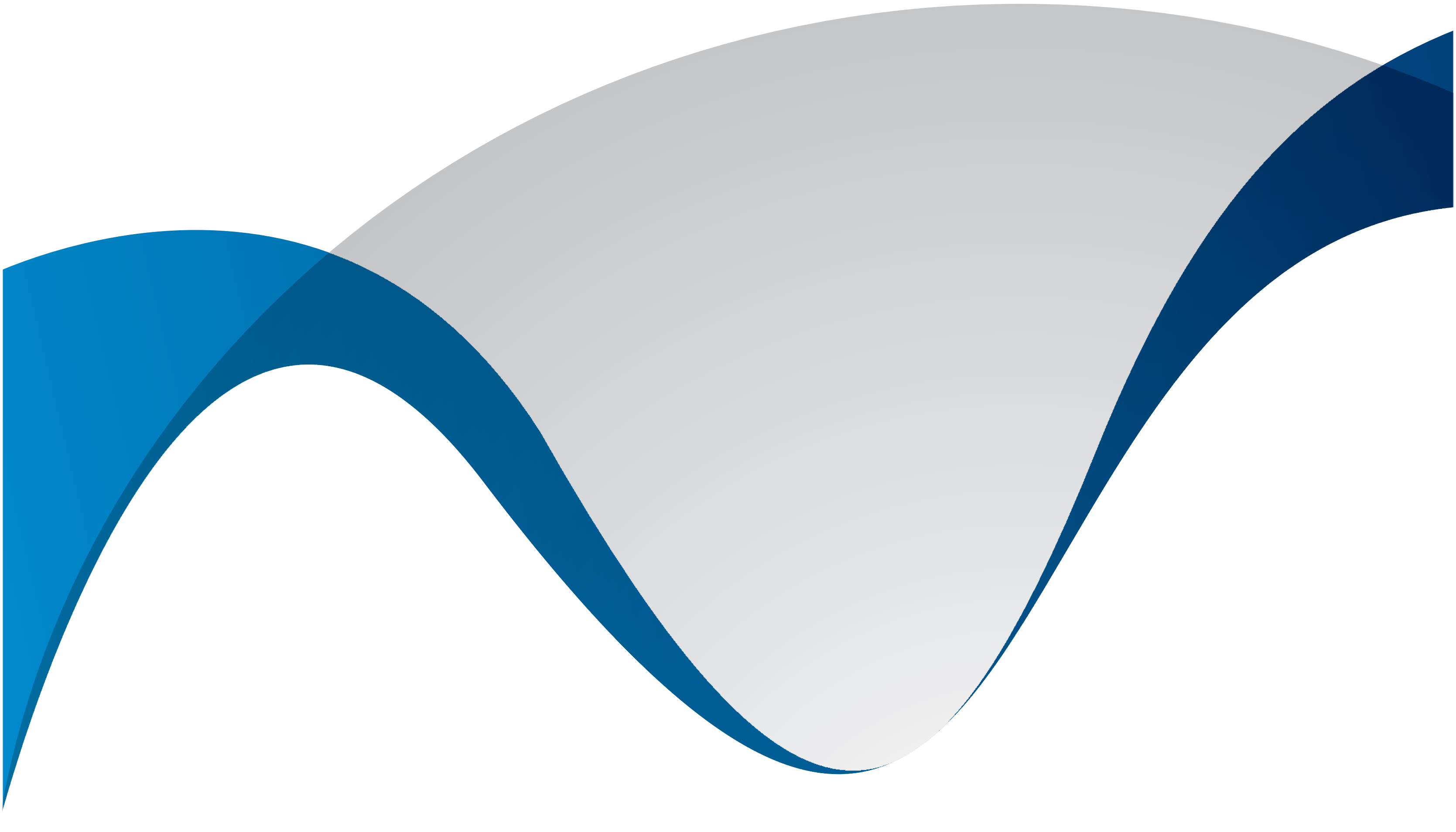
Business Intelligence and Analytics



Drew Bentley

Business Intelligence and Analytics

Business Intelligence and Analytics

Edited by

Drew Bentley

Business Intelligence and Analytics
Edited by Drew Bentley
ISBN: 978-1-9789-2136-8

© 2017 Library Press

Published by Library Press,
5 Penn Plaza,
19th Floor,
New York, NY 10001, USA

Cataloging-in-Publication Data

Business intelligence and analytics / edited by Drew Bentley.
p. cm.
Includes bibliographical references and index.
ISBN 978-1-9789-2136-8
1. Business intelligence. 2. Industrial management--Statistical methods--Data processing.
3. Business planning--Data processing.
4. Management. I. Bentley, Drew.

HD38.7 .B87 2017
658.472--dc23

This book contains information obtained from authentic and highly regarded sources. All chapters are published with permission under the Creative Commons Attribution Share Alike License or equivalent. A wide variety of references are listed. Permissions and sources are indicated; for detailed attributions, please refer to the permissions page. Reasonable efforts have been made to publish reliable data and information, but the authors, editors and publisher cannot assume any responsibility for the validity of all materials or the consequences of their use.

Copyright of this ebook is with Library Press, rights acquired from the original print publisher, Larsen and Keller Education.

Trademark Notice: All trademarks used herein are the property of their respective owners. The use of any trademark in this text does not vest in the author or publisher any trademark ownership rights in such trademarks, nor does the use of such trademarks imply any affiliation with or endorsement of this book by such owners.

The publisher's policy is to use permanent paper from mills that operate a sustainable forestry policy. Furthermore, the publisher ensures that the text paper and cover boards used have met acceptable environmental accreditation standards.

Table of Contents

Preface VII

Chapter 1 **Introduction to Business Intelligence** **1**
- Business Intelligence 1
- Mobile Business Intelligence 11
- Real-time Business Intelligence 18

Chapter 2 **Analytics: A Comprehensive Study** **22**
- Business Analytics 22
- Analytics 24
- Software Analytics 28
- Embedded Analytics 30
- Learning Analytics 31
- Predictive Analytics 37
- Prescriptive Analytics 51
- Social Media Analytics 56
- Behavioral Analytics 57

Chapter 3 **Data Mining: An Overview** **64**
- Data Mining 64
- Anomaly Detection 73
- Association Rule Learning 75
- Cluster Analysis 82
- Statistical Classification 97
- Regression Analysis 101
- Automatic Summarization 110
- Examples of Data Mining 120

Chapter 4 **Understanding Data Warehousing** **129**
- Data Warehouse 129
- Data Mart 137
- Master Data Management 139
- Dimension (Data Warehouse) 142
- Slowly Changing Dimension 146
- Data Vault Modeling 154
- Extract, Transform, Load 161
- Star Schema 169

Chapter 5 **Market Research: An Integrated Study** **173**
- Market Research 173
- Market Segmentation 178
- Market Trend 195
- SWOT Analysis 200
- Marketing Research 207

Chapter 6 **Essential Aspects of Business Intelligence** **220**
- Context Analysis 220
- Business Performance Management 225
- Business Process Discovery 230
- Information System 234
- Organizational Intelligence 241
- Data Visualization 245
- Data Profiling 256
- Data Cleansing 258
- Process Mining 264
- Competitive Intelligence 266

Chapter 7 **Operational Intelligence: Technological Components** **272**
- Operational Intelligence 272
- Business Activity Monitoring 275
- Complex Event Processing 277
- Business Process Management 281
- Metadata 288
- Root Cause Analysis 302

Permissions

Index

Preface

Data is raw facts and figures and information is meaningful data that would be helpful for a person or company. Business intelligence extracts information from raw data through tools like data mining, perspective analysis, online analytical processing etc. The textbook will provide comprehensive information to readers about business intelligence and analytics. This book explores all the important aspects of business intelligence and analytics in the present day scenario. The topics covered in this extensive book deal with the core subjects of business intelligence. It aims to serve as a resource guide for students and facilitate the study of the discipline.

Given below is the chapter wise description of the book:

Chapter 1- The strategy and the planning that is incorporated in any business is known as business intelligence. It may also include products, technologies and analysis and presentation of business information. This chapter will provide an integrated understanding of business intelligence.

Chapter 2- Analytics is the understanding and communication of significant patterns of data. Analytics is applied in businesses to improve their performances. Some of the aspects explained in this text are software analytics, embedded analytics, learning analytics and social media analytics. The section on analytics offers an insightful focus, keeping in mind the complex subject matter.

Chapter 3- The process of understanding the patterns found in large data sets is known as data mining. Some of the aspects of data mining that have been elucidated in the following section are association rule learning, cluster analysis, regression analysis, automatic summarization and examples of data mining. The chapter on data mining offers an insightful focus, keeping in mind the complex subject matter.

Chapter 4- Data warehouse is the core of business intelligence. It is majorly used for reporting and analyzing data. Data mart, master data management, dimension, slowly changing dimension and star schema. This text elucidates the crucial theories and principles of data warehousing.

Chapter 5- The effort put in to gather information related to customers or markets is known as market research. Market research is an important part of business strategy. Market segmentation, market trend, SWOT analysis and market research are some of the topics elucidated in this chapter.

Chapter 6- The essential aspects of business intelligence are context analysis, business performance management, business process discovery, information system, organization intelligence and process mining. The method to analyze the environment of any business is known as context analysis. The topics discussed in this section are of great importance to broaden the existing knowledge on business intelligence.

Chapter 7- Operational intelligence has a number of aspects that have been elucidated in this chapter. Some of these features are complex event processing, business process management, metadata and root cause analysis. The components discussed in this text are of great importance to broaden the existing knowledge on operational intelligence.

Indeed, my job was extremely crucial and challenging as I had to ensure that every chapter is informative and structured in a student-friendly manner. I am thankful for the support provided by my family and colleagues during the completion of this book.

Editor

1

Introduction to Business Intelligence

The strategy and the planning that is incorporated in any business is known as business intelligence. It may also include products, technologies and analysis and presentation of business information. This chapter will provide an integrated understanding of business intelligence.

Business Intelligence

Business intelligence (BI) can be described as "a set of techniques and tools for the acquisition and transformation of raw data into meaningful and useful information for business analysis purposes". The term "data surfacing" is also more often associated with BI functionality. BI technologies are capable of handling large amounts of structured and sometimes unstructured data to help identify, develop and otherwise create new strategic business opportunities. The goal of BI is to allow for the easy interpretation of these large volumes of data. Identifying new opportunities and implementing an effective strategy based on insights can provide businesses with a competitive market advantage and long-term stability.

BI technologies provide historical, current and predictive views of business operations. Common functions of business intelligence technologies are reporting, online analytical processing, analytics, data mining, process mining, complex event processing, business performance management, benchmarking, text mining, predictive analytics and prescriptive analytics.

BI can be used to support a wide range of business decisions ranging from operational to strategic. Basic operating decisions include product positioning or pricing. Strategic business decisions include priorities, goals and directions at the broadest level. In all cases, BI is most effective when it combines data derived from the market in which a company operates (external data) with data from company sources internal to the business such as financial and operations data (internal data). When combined, external and internal data can provide a more complete picture which, in effect, creates an "intelligence" that cannot be derived by any singular set of data. Amongst myriad uses, BI tools empower organisations to gain insight into new markets, assess demand and suitability of products and services for different market segments and gauge the impact of marketing efforts.

Components

Business intelligence is made up of an increasing number of components including:

- Multidimensional aggregation and allocation
- Denormalization, tagging and standardization

- Realtime reporting with analytical alert
- A method of interfacing with unstructured data sources
- Group consolidation, budgeting and rolling forecasts
- Statistical inference and probabilistic simulation
- Key performance indicators optimization
- Version control and process management
- Open item management

History

The earliest known use of the term "Business Intelligence" is in Richard Millar Devens' in the 'Cyclopædia of Commercial and Business Anecdotes' from 1865. Devens used the term to describe how the banker, Sir Henry Furnese, gained profit by receiving and acting upon information about his environment, prior to his competitors. "*Throughout Holland, Flanders, France, and Germany, he maintained a complete and perfect train of business intelligence. The news of the many battles fought was thus received first by him, and the fall of Namur added to his profits, owing to his early receipt of the news.*" (Devens, (1865), p. 210). The ability to collect and react accordingly based on the information retrieved, an ability that Furnese excelled in, is today still at the very heart of BI.

In a 1958 article, IBM researcher Hans Peter Luhn used the term business intelligence. He employed the Webster's dictionary definition of intelligence: "the ability to apprehend the interrelationships of presented facts in such a way as to guide action towards a desired goal."

Business intelligence as it is understood today is said to have evolved from the decision support systems (DSS) that began in the 1960s and developed throughout the mid-1980s. DSS originated in the computer-aided models created to assist with decision making and planning. From DSS, data warehouses, Executive Information Systems, OLAP and business intelligence came into focus beginning in the late 80s.

In 1989, Howard Dresner (later a Gartner analyst) proposed "business intelligence" as an umbrella term to describe "concepts and methods to improve business decision making by using fact-based support systems." It was not until the late 1990s that this usage was widespread.

Data Warehousing

Often BI applications use data gathered from a data warehouse (DW) or from a data mart, and the concepts of BI and DW sometimes combine as "BI/DW" or as "BIDW". A data warehouse contains a copy of analytical data that facilitates decision support. However, not all data warehouses serve for business intelligence, nor do all business intelligence applications require a data warehouse.

To distinguish between the concepts of business intelligence and data warehouses, Forrester Research defines business intelligence in one of two ways:

1. Using a broad definition: "Business Intelligence is a set of methodologies, processes, ar-

chitectures, and technologies that transform raw data into meaningful and useful information used to enable more effective strategic, tactical, and operational insights and decision-making." Under this definition, business intelligence also includes technologies such as data integration, data quality, data warehousing, master-data management, text- and content-analytics, and many others that the market sometimes lumps into the "Information Management" segment. Therefore, Forrester refers to *data preparation* and *data usage* as two separate but closely linked segments of the business-intelligence architectural stack.

2. Forrester defines the narrower business-intelligence market as, "...referring to just the top layers of the BI architectural stack such as reporting, analytics and dashboards."

Comparison with Competitive Intelligence

Though the term business intelligence is sometimes a synonym for competitive intelligence (because they both support decision making), BI uses technologies, processes, and applications to analyze mostly internal, structured data and business processes while competitive intelligence gathers, analyzes and disseminates information with a topical focus on company competitors. If understood broadly, business intelligence can include the subset of competitive intelligence.

Comparison with Business Analytics

Business intelligence and business analytics are sometimes used interchangeably, but there are alternate definitions. One definition contrasts the two, stating that the term business intelligence refers to collecting business data to find information primarily through asking questions, reporting, and online analytical processes. Business analytics, on the other hand, uses statistical and quantitative tools for explanatory and predictive modeling.

In an alternate definition, Thomas Davenport, professor of information technology and management at Babson College argues that business intelligence should be divided into querying, reporting, Online analytical processing (OLAP), an "alerts" tool, and business analytics. In this definition, business analytics is the subset of BI focusing on statistics, prediction, and optimization, rather than the reporting functionality.

Applications in an Enterprise

Business intelligence can be applied to the following business purposes, in order to drive business value.

1. Measurement – program that creates a hierarchy of performance metrics and benchmarking that informs business leaders about progress towards business goals (business process management).
2. Analytics – program that builds quantitative processes for a business to arrive at optimal decisions and to perform business knowledge discovery. Frequently involves: data mining, process mining, statistical analysis, predictive analytics, predictive modeling, business process modeling, data lineage, complex event processing and prescriptive analytics.

3. Reporting/enterprise reporting – program that builds infrastructure for strategic reporting to serve the strategic management of a business, not operational reporting. Frequently involves data visualization, executive information system and OLAP.
4. Collaboration/collaboration platform – program that gets different areas (both inside and outside the business) to work together through data sharing and electronic data interchange.
5. Knowledge management – program to make the company data-driven through strategies and practices to identify, create, represent, distribute, and enable adoption of insights and experiences that are true business knowledge. Knowledge management leads to learning management and regulatory compliance.

In addition to the above, business intelligence can provide a pro-active approach, such as alert functionality that immediately notifies the end-user if certain conditions are met. For example, if some business metric exceeds a pre-defined threshold, the metric will be highlighted in standard reports, and the business analyst may be alerted via e-mail or another monitoring service. This end-to-end process requires data governance, which should be handled by the expert.

Prioritization of Projects

It can be difficult to provide a positive business case for business intelligence initiatives, and often the projects must be prioritized through strategic initiatives. BI projects can attain higher prioritization within the organization if managers consider the following:

- As described by Kimball the BI manager must determine the tangible benefits such as eliminated cost of producing legacy reports.
- Data access for the entire organization must be enforced. In this way even a small benefit, such as a few minutes saved, makes a difference when multiplied by the number of employees in the entire organization.
- As described by Ross, Weil & Roberson for Enterprise Architecture, managers should also consider letting the BI project be driven by other business initiatives with excellent business cases. To support this approach, the organization must have enterprise architects who can identify suitable business projects.
- Using a structured and quantitative methodology to create defensible prioritization in line with the actual needs of the organization, such as a weighted decision matrix.

Success Factors of Implementation

According to Kimball et al., there are three critical areas that organizations should assess before getting ready to do a BI project:

1. The level of commitment and sponsorship of the project from senior management.
2. The level of business need for creating a BI implementation.
3. The amount and quality of business data available.

Business Sponsorship

The commitment and sponsorship of senior management is according to Kimball *et al.*, the most important criteria for assessment. This is because having strong management backing helps overcome shortcomings elsewhere in the project. However, as Kimball *et al.* state: “even the most elegantly designed DW/BI system cannot overcome a lack of business [management] sponsorship”.

It is important that personnel who participate in the project have a vision and an idea of the benefits and drawbacks of implementing a BI system. The best business sponsor should have organizational clout and should be well connected within the organization. It is ideal that the business sponsor is demanding but also able to be realistic and supportive if the implementation runs into delays or drawbacks. The management sponsor also needs to be able to assume accountability and to take responsibility for failures and setbacks on the project. Support from multiple members of the management ensures the project does not fail if one person leaves the steering group. However, having many managers work together on the project can also mean that there are several different interests that attempt to pull the project in different directions, such as if different departments want to put more emphasis on their usage. This issue can be countered by an early and specific analysis of the business areas that benefit the most from the implementation. All stakeholders in the project should participate in this analysis in order for them to feel invested in the project and to find common ground.

Another management problem that may be encountered before the start of an implementation is an overly aggressive business sponsor. Problems of scope creep occur when the sponsor requests data sets that were not specified in the original planning phase.

Business Needs

Because of the close relationship with senior management, another critical thing that must be assessed before the project begins is whether or not there is a business need and whether there is a clear business benefit by doing the implementation. The needs and benefits of the implementation are sometimes driven by competition and the need to gain an advantage in the market. Another reason for a business-driven approach to implementation of BI is the acquisition of other organizations that enlarge the original organization it can sometimes be beneficial to implement DW or BI in order to create more oversight.

Companies that implement BI are often large, multinational organizations with diverse subsidiaries. A well-designed BI solution provides a consolidated view of key business data not available anywhere else in the organization, giving management visibility and control over measures that otherwise would not exist.

Amount and Quality of Available Data

Without proper data, or with too little quality data, any BI implementation fails; it does not matter how good the management sponsorship or business-driven motivation is. Before implementation it is a good idea to do data profiling. This analysis identifies the “content, consistency and structure [..]” of the data. This should be done as early as possible in the process and if the analysis shows that data is lacking, put the project on hold temporarily while the IT department figures out how to properly collect data.

When planning for business data and business intelligence requirements, it is always advisable to consider specific scenarios that apply to a particular organization, and then select the business intelligence features best suited for the scenario.

Often, scenarios revolve around distinct business processes, each built on one or more data sources. These sources are used by features that present that data as information to knowledge workers, who subsequently act on that information. The business needs of the organization for each business process adopted correspond to the essential steps of business intelligence. These essential steps of business intelligence include but are not limited to:

1. Go through business data sources in order to collect needed data
2. Convert business data to information and present appropriately
3. Query and analyze data
4. Act on the collected data

The quality aspect in business intelligence should cover all the process from the source data to the final reporting. At each step, the quality gates are different:

1. Source Data:
    - Data Standardization: make data comparable (same unit, same pattern...)
    - Master Data Management: unique referential
2. Operational Data Store (ODS):
    - Data Cleansing: detect & correct inaccurate data
    - Data Profiling: check inappropriate value, null/empty
3. Data Warehouse:
    - Completeness: check that all expected data are loaded
    - Referential integrity: unique and existing referential over all sources
    - Consistency between sources: check consolidated data vs sources
4. Reporting:
    - Uniqueness of indicators: only one share dictionary of indicators
    - Formula accuracy: local reporting formula should be avoided or checked

User Aspect

Some considerations must be made in order to successfully integrate the usage of business intelligence systems in a company. Ultimately the BI system must be accepted and utilized by the users in order for it to add value to the organization. If the usability of the system is poor, the users may become frustrated and spend a considerable amount of time figuring out how to use the system or may not be able to really use the system. If the system does not add value to the users´ mission, they simply don’t use it.

To increase user acceptance of a BI system, it can be advisable to consult business users at an early stage of the DW/BI lifecycle, for example at the requirements gathering phase. This can provide an insight into the business process and what the users need from the BI system. There are several methods for gathering this information, such as questionnaires and interview sessions.

When gathering the requirements from the business users, the local IT department should also be consulted in order to determine to which degree it is possible to fulfill the business's needs based on the available data.

Taking a user-centered approach throughout the design and development stage may further increase the chance of rapid user adoption of the BI system.

Besides focusing on the user experience offered by the BI applications, it may also possibly motivate the users to utilize the system by adding an element of competition. Kimball suggests implementing a function on the Business Intelligence portal website where reports on system usage can be found. By doing so, managers can see how well their departments are doing and compare themselves to others and this may spur them to encourage their staff to utilize the BI system even more.

In a 2007 article, H. J. Watson gives an example of how the competitive element can act as an incentive. Watson describes how a large call centre implemented performance dashboards for all call agents, with monthly incentive bonuses tied to performance metrics. Also, agents could compare their performance to other team members. The implementation of this type of performance measurement and competition significantly improved agent performance.

BI chances of success can be improved by involving senior management to help make BI a part of the organizational culture, and by providing the users with necessary tools, training, and support. Training encourages more people to use the BI application.

Providing user support is necessary to maintain the BI system and resolve user problems. User support can be incorporated in many ways, for example by creating a website. The website should contain great content and tools for finding the necessary information. Furthermore, helpdesk support can be used. The help desk can be manned by power users or the DW/BI project team.

BI Portals

A Business Intelligence portal (BI portal) is the primary access interface for Data Warehouse (DW) and Business Intelligence (BI) applications. The BI portal is the user's first impression of the DW/BI system. It is typically a browser application, from which the user has access to all the individual services of the DW/BI system, reports and other analytical functionality. The BI portal must be implemented in such a way that it is easy for the users of the DW/BI application to call on the functionality of the application.

The BI portal's main functionality is to provide a navigation system of the DW/BI application. This means that the portal has to be implemented in a way that the user has access to all the functions of the DW/BI application.

The most common way to design the portal is to custom fit it to the business processes of the organization for which the DW/BI application is designed, in that way the portal can best fit the needs and requirements of its users.

The BI portal needs to be easy to use and understand, and if possible have a look and feel similar to other applications or web content of the organization the DW/BI application is designed for (consistency).

The following is a list of desirable features for web portals in general and BI portals in particular:

Usable

User should easily find what they need in the BI tool.

Content Rich

The portal is not just a report printing tool, it should contain more functionality such as advice, help, support information and documentation.

Clean

The portal should be designed so it is easily understandable and not over-complex as to confuse the users

Current

The portal should be updated regularly.

Interactive

The portal should be implemented in a way that makes it easy for the user to use its functionality and encourage them to use the portal. Scalability and customization give the user the means to fit the portal to each user.

Value Oriented

It is important that the user has the feeling that the DW/BI application is a valuable resource that is worth working on.

Marketplace

There are a number of business intelligence vendors, often categorized into the remaining independent “pure-play” vendors and consolidated “megavendors” that have entered the market through a recent trend of acquisitions in the BI industry. The business intelligence market is gradually growing. In 2012 business intelligence services brought in $13.1 billion in revenue.

Some companies adopting BI software decide to pick and choose from different product offerings (best-of-breed) rather than purchase one comprehensive integrated solution (full-service).

Industry-specific

Specific considerations for business intelligence systems have to be taken in some sectors such as governmental banking regulations or healthcare. The information collected by banking institutions and analyzed with BI software must be protected from some groups or individuals, while being fully available to other groups or individuals. Therefore, BI solutions must be sensitive to those needs and be flexible enough to adapt to new regulations and changes to existing law.

Semi-structured or Unstructured Data

Businesses create a huge amount of valuable information in the form of e-mails, memos, notes from call-centers, news, user groups, chats, reports, web-pages, presentations, image-files, video-files, and marketing material and news. According to Merrill Lynch, more than 85% of all business information exists in these forms. These information types are called either *semi-structured* or *unstructured* data. However, organizations often only use these documents once.

The managements of semi-structured data is recognized as a major unsolved problem in the information technology industry. According to projections from Gartner (2003), white collar workers spend anywhere from 30 to 40 percent of their time searching, finding and assessing unstructured data. BI uses both structured and unstructured data, but the former is easy to search, and the latter contains a large quantity of the information needed for analysis and decision making. Because of the difficulty of properly searching, finding and assessing unstructured or semi-structured data, organizations may not draw upon these vast reservoirs of information, which could influence a particular decision, task or project. This can ultimately lead to poorly informed decision making.

Therefore, when designing a business intelligence/DW-solution, the specific problems associated with semi-structured and unstructured data must be accommodated for as well as those for the structured data.

Unstructured Data vs. Semi-structured

Unstructured and semi-structured data have different meanings depending on their context. In the context of relational database systems, unstructured data cannot be stored in predictably ordered columns and rows. One type of unstructured data is typically stored in a BLOB (binary large object), a catch-all data type available in most relational database management systems. Unstructured data may also refer to irregularly or randomly repeated column patterns that vary from row to row within each file or document.

Many of these data types, however, like e-mails, word processing text files, PPTs, image-files, and video-files conform to a standard that offers the possibility of metadata. Metadata can include information such as author and time of creation, and this can be stored in a relational database. Therefore, it may be more accurate to talk about this as semi-structured documents or data, but no specific consensus seems to have been reached.

Unstructured data can also simply be the knowledge that business users have about future business trends. Business forecasting naturally aligns with the BI system because business users think of their business in aggregate terms. Capturing the business knowledge that may only exist in the minds of business users provides some of the most important data points for a complete BI solution.

Problems with Semi-structured or Unstructured Data

There are several challenges to developing BI with semi-structured data. According to Inmon & Nesavich, some of those are:

1. Physically accessing unstructured textual data – unstructured data is stored in a huge variety of formats.

2. Terminology – Among researchers and analysts, there is a need to develop a standardized terminology.
3. Volume of data – As stated earlier, up to 85% of all data exists as semi-structured data. Couple that with the need for word-to-word and semantic analysis.
4. Searchability of unstructured textual data – A simple search on some data, e.g. apple, results in links where there is a reference to that precise search term. (Inmon & Nesavich, 2008) gives an example: “a search is made on the term felony. In a simple search, the term felony is used, and everywhere there is a reference to felony, a hit to an unstructured document is made. But a simple search is crude. It does not find references to crime, arson, murder, embezzlement, vehicular homicide, and such, even though these crimes are types of felonies.”

The Use of Metadata

To solve problems with searchability and assessment of data, it is necessary to know something about the content. This can be done by adding context through the use of metadata. Many systems already capture some metadata (e.g. filename, author, size, etc.), but more useful would be metadata about the actual content – e.g. summaries, topics, people or companies mentioned. Two technologies designed for generating metadata about content are automatic categorization and information extraction.

Future

A 2009 paper predicted these developments in the business intelligence market:

- Because of lack of information, processes, and tools, through 2012, more than 35 percent of the top 5,000 global companies regularly fail to make insightful decisions about significant changes in their business and markets.
- By 2012, business units will control at least 40 percent of the total budget for business intelligence.
- By 2012, one-third of analytic applications applied to business processes will be delivered through coarse-grained application mashups.
- BI has a huge scope in Entrepreneurship however majority of new entrepreneurs ignore its potential.

A 2009 *Information Management* special report predicted the top BI trends: “green computing, social networking services, data visualization, mobile BI, predictive analytics, composite applications, cloud computing and multitouch”. Research undertaken in 2014 indicated that employees are more likely to have access to, and more likely to engage with, cloud-based BI tools than traditional tools.

Other business intelligence trends include the following:

- Third party SOA-BI products increasingly address ETL issues of volume and throughput.

- Companies embrace in-memory processing, 64-bit processing, and pre-packaged analytic BI applications.
- Operational applications have callable BI components, with improvements in response time, scaling, and concurrency.
- Near or real time BI analytics is a baseline expectation.
- Open source BI software replaces vendor offerings.

Other lines of research include the combined study of business intelligence and uncertain data. In this context, the data used is not assumed to be precise, accurate and complete. Instead, data is considered uncertain and therefore this uncertainty is propagated to the results produced by BI.

According to a study by the Aberdeen Group, there has been increasing interest in Software-as-a-Service (SaaS) business intelligence over the past years, with twice as many organizations using this deployment approach as one year ago – 15% in 2009 compared to 7% in 2008.

An article by InfoWorld's Chris Kanaracus points out similar growth data from research firm IDC, which predicts the SaaS BI market will grow 22 percent each year through 2013 thanks to increased product sophistication, strained IT budgets, and other factors.

An analysis of top 100 Business Intelligence and Analytics scores and ranks the firms based on several open variables

Mobile Business Intelligence

Mobile Business Intelligence (Mobile BI or Mobile Intelligence) is defined as "The capability that enables the mobile workforce to gain business insights through information analysis using applications optimized for mobile devices" Verkooij(2012) Business intelligence (BI) refers to computer-based techniques used in spotting, digging-out, and analyzing business data, such as sales revenue by products and/or departments or associated costs and incomes.

Although the concept of mobile computing has been prevalent for over a decade, Mobile BI has shown a momentum/growth only very recently. This change has been partly encouraged by a change from the 'wired world' to a wireless world with the advantage of smartphones which has led to a new era of mobile computing, especially in the field of BI.

According to the Aberdeen Group, a large number of companies are rapidly undertaking mobile BI owing to a large number of market pressures such as the need for higher efficiency in business processes, improvement in employee productivity (e.g., time spent looking for information), better and faster decision making, better customer service, and delivery of real-time bi-directional data access to make decisions anytime and anywhere. But despite the apparent advantages of mobile information delivery, mobile BI is still in the 'early adopter' phase. Some CFOs remain sceptical of the business benefits and with the perceived lack of specific business use cases and tangible ROI, mobile BI adoption is still behind the curve compared with other enterprise mobile applications.

History

Information Delivery to Mobile Devices

The predominant method for accessing BI information is using proprietary software or a Web browser on a personal computer to connect to BI applications. These BI applications request data from databases. Starting in the late 1990s, BI systems offered alternatives for receiving data, including email and mobile devices.

Static Data Push

Initially, mobile devices such as pagers and mobile phones received pushed data using a short message service (SMS) or text messages. These applications were designed for specific mobile devices, contained minimal amounts of information, and provided no data interactivity. As a result, the early mobile BI applications were expensive to design and maintain while providing limited informational value, and garnered little interest.

Data Access Via a Mobile Browser

The mobile browser on a smartphone, a handheld computer integrated with a mobile phone, provided a means to read simple tables of data. The small screen space, immature mobile browsers, and slow data transmission could not provide a satisfactory BI experience. Accessibility and bandwidth may be perceived as issues when it comes to mobile technology, but BI solutions provide advanced functionality to predict and outperform such potential challenges. While Web-based mobile BI solutions provide little to no control over the processing of data in a network, managed BI solutions for mobile devices only utilize the server for specific operations. In addition, local reports are compressed both during transmission and on the device, permitting greater flexibility for storage and receipt of these reports. Within a mobile environment, users capitalize on easy access to information because the mobile application operates within a single authoring environment that permits access to all BI content (respecting existing security) regardless of language or locale. Furthermore, the user will not need to build and maintain a separate mobile BI deployment. In addition, mobile BI requires much less bandwidth for functionality. Mobile BI promises a small report footprint on memory, encryption during transmission as well as on the device, and compressed data storage for offline viewing and use.

Mobile Client Application

In 2002, Research in Motion released the first BlackBerry smartphone optimized for wireless email use. Wireless e-mail proved to be the “killer app” that accelerated the popularity of the smartphone market. By the mid-2000s, Research in Motion’s BlackBerry had solidified its hold on the smartphone market with both corporate and governmental organizations. The BlackBerry smartphones eliminated the obstacles to mobile business intelligence. The BlackBerry offered a consistent treatment of data across its many models, provided a much larger screen for viewing data, and allowed user interactivity via the thumbwheel and keyboard. BI vendors re-entered the market with offerings spanning different mobile operating systems (BlackBerry, Windows, Symbian) and data access methods. The two most popular data access options were:

- to use the mobile browser to access data, similar to desktop computer, and

- to create a native application designed specifically for the mobile device.

Research in Motion is continuing to lose market share to Apple and Android smartphones. In the first three months of 2011 Google's Android OS gained 7 points of market share. During the same time period RIM's market share collapsed and dropped almost 5 points.

Purpose-built Mobile BI Apps

Apple quickly set the standard for mobile devices with the introduction of the iPhone. In the first three years, Apple sold over 33.75 million units. Similarly, in 2010, Apple sold over 1 million iPads in just under three months. Both devices feature an interactive touchscreen display that is the de facto standard on many mobile phones and tablet computers.

In 2008, Apple published the SDK for which developers can build applications that run natively on the iPhone and iPad instead of Safari-based applications. These native applications can give the user a robust, easier-to-read and easier-to-navigate experience.

Others were quick to join in the success of mobile devices and app downloads. The Google Play Store now has over 700,000 apps available for the mobile devices running the Android operating system.

More importantly, the advent of the mobile device has radically changed the way people use data on their mobile devices. This includes mobile BI. Business intelligence applications can be used to transform reports and data into mobile dashboards, and have them instantly delivered to any mobile device.

Google Inc.'s Android has overtaken Apple Inc.'s iOS in the wildly growing arena of app downloads. In the second quarter of 2011, 44% of all apps downloaded from app marketplaces across the web were for Android devices while 31% were for Apple devices, according to new data from ABI Research. The remaining apps were for various other mobile operating systems, including BlackBerry and Windows Phone 7.

Mobile BI applications have evolved from being a client application for viewing data to a purpose-built application designed to provide information and workflows necessary to quickly make business decisions and take action.

Web Applications vs. Device-specific Applications for Mobile BI

In early 2011, as the mobile BI software market started to mature and adoption started to grow at a significant pace in both small and large enterprises, most vendors adopted either a purpose-built, device-specific application strategy (e.g. iPhone or Android apps, downloaded from iTunes or the Google Play Store) or a web application strategy (browser-based, works on most devices without an application being installed on the device). This debate continues and there are benefits and drawbacks to both methods. One potential solution will be the wider adoption of HTML5 on mobile devices which will give web applications many of the characteristics of dedicated applications while still allowing them to work on many devices without an installed application.

Microsoft has announced their mobile BI strategy. Microsoft plans to support browser-based applications such as Reporting Services and PerformancePoint on iOS in the first half of 2012 and touch-based applications on iOS and Android by the second half of 2012. Despite popular perception that Microsoft only acknowledges its own existence, recent moves suggest the company is aware that it is not the only player in the technology ecosystem. Instead of attempting to squelch competition or suggesting new technology developments were ridiculous, the company has instead decided to make its technology accessible to a wider audience.

There are many mobile devices and platforms available today. The list is constantly growing and so is the platform support. There are hundreds of models available today, with multiple hardware and software combinations. The enterprise must select a device very carefully. The target devices will impact the mobile BI design itself because the design for a smartphone will be different than for a tablet. The screen size, processor, memory, etc. all vary. The mobile BI program must account for lack of device standardization from the providers by constantly testing devices for the mobile BI apps. Some best practices can always be followed. For example, a smartphone is a good candidate for operational mobile BI. However, for analytics and what-if analysis, tablets are the best option. Hence, the selection or availability of the device plays a big role in the implementation.

Demand

Gartner analyst Ted Friedman believes that mobile delivery of BI is all about practical, tactical information needed to make immediate decisions – “The biggest value is in operational BI — information in the context of applications — not in pushing lots of data to somebody’s phone.”

Accessing the Internet through a mobile device such as a smartphone is also known as the mobile Internet or mobile Web. IDC expects the US mobile workforce to increase by 73% in 2011. Morgan Stanley reports the mobile Internet is ramping up faster than its predecessor, the desktop Internet, enabling companies to deliver knowledge to their mobile workforce to help them make more profitable decisions.

Michael Cooney from Gartner has identified bring-your-own-technology at work as becoming the norm, not the exception. By 2015 media tablet shipments will reach around 50% of laptop shipments and Windows 8 will likely be in third place behind Android and Apple. The net result is that Microsoft’s share of the client platform, be it PC, tablet or smartphone, will likely be reduced to 60% and it could fall below 50%.

Business Benefits

In its latest Magic Quadrant for Business Intelligence Platforms, Gartner examines whether the platform enables users to “fully interact with BI content delivered to mobile devices.” The phrase “fully interact” is the key. The ability to send alerts embedded in email or text messages, or links to static content in email messages hardly represents sophistication in mobile analytics. For users to benefit from mobile BI, they must be able to navigate dashboard and guided analytics comfortably—or as comfortably as the mobile device will allow, which is where devices with high-resolution screens and touch interfaces (like the iPhone and Android-based phones) have a clear edge over, say, earlier editions of BlackBerry. It is equally important to take a step back to define your purpose and adoption patterns. Which business users can benefit the most from mobile analytics—and what, exactly, is their requirement? You don’t need mobile analytics to send a few alerts or summary reports to their handhelds—without interactivity, mobile BI is indistinguishable from merely informative email or text messages.

Applications

Similar to consumer applications, which have shown an ever increasing growth over the past few years, a constant demand for anytime, anywhere access to BI is leading to a number of custom mobile application development. Businesses have also started adopting mobile solutions for their workforce and are soon becoming key components of core business processes. In an Aberdeen survey conducted in May 2010, 23% of companies participating indicated that they now have a mobile BI app or dashboard in place, while another 31% indicated that they plan to implement some form of mobile BI in the next year.

Definitions

Mobile BI applications can be defined/segregated as follows:

- Mobile Browser Rendered App: Almost any mobile device enables Web-based, thin client, HTML-only BI applications. However, these apps are static and provide little data interactivity. Data is viewed just as it would be over a browser from a personal computer. Little additional effort is required to display data but mobile browsers can typically only support a small subset of the interactivity of a web browser.
- Customized App: A step up from this approach is to render each (or all) reports and dashboards in device-specific format. In other words, provide information specific to the screen size, optimize usage of screen real estate, and enable device-specific navigation controls. Examples of these include thumb wheel or thumb button for BlackBerry, up/down/left/right arrows for Palm, gestural manipulation for iPhone. This approach requires more effort than the previous but no additional software.
- Mobile Client App: The most advanced, the client app provides full interactivity with the BI content viewed on the device. In addition, this approach provides periodic caching of data which can be viewed and analyzed even offline.

Companies across all verticals, from retail to even non-profit organizations are realizing the value of purpose-specific mobile applications suited for their mobile workforce.

Development

Developing a native mobile BI app poses challenges, especially concerning data display rendering and user interactivity. Mobile BI App development has traditionally been a time-consuming and expensive effort requiring businesses to justify the investment for the mobile workforce. They do not only require texting and alerts, they need information customized for their line of work which they can interact with and analyze to gain deeper information.

Custom-coded Mobile BI Apps

Mobile BI applications are often custom-coded apps specific to the underlying mobile operating system. For example, the iPhone apps require coding in Objective-C while Android apps require coding in Java. In addition to the user functionality of the app, the app must be coded to work with the supporting server infrastructure required to serve data to the mobile BI app. While custom-coded apps offer near limitless options, the specialized software coding expertise and infrastructure can be expensive to develop, modify, and maintain.

Fixed-form Mobile BI Applications

Business data can be displayed in a mobile BI client (or web browser) that serves as a user interface to existing BI platforms or other data sources, eliminating the need for new master sources of data and specialized server infrastructure. This option offers fixed and configurable data visualizations such as charts, tables, trends, KPIs, and links, and can usually be deployed quickly using existing data sources. However, the data visualizations are not limitless and cannot always be extended to beyond what is available from the vendor.

Graphical Tool-developed Mobile BI Apps

Mobile BI apps can also be developed using the graphical, drag-and-drop development environments of BI platforms. The advantages including the following:

1. Apps can be developed without coding,
2. Apps can be easily modified and maintained using the BI platform change management tools,
3. Apps can use any range of data visualizations and not be limited to just a few,
4. Apps can incorporate specific business workflows, and
5. The BI platform provides the server infrastructure.

Using graphical BI development tools can allow faster mobile BI app development when a custom application is required.

Security Considerations for Mobile BI Apps

High adoption rates and reliance on mobile devices makes safe mobile computing a critical concern. The Mobile Business Intelligence Market Study discovered that security is the number one issue (63%) for organizations.

A comprehensive mobile security solution must provide security at these levels:

- Device
- Transmission
- Authorization, Authentication, and Network Security

Device Security

A senior analyst at the Burton Group research firm recommends that the best way to ensure data will not be tampered with is to not store it on the client device (mobile device). As such, there is no local copy to lose if the mobile device is stolen and the data can reside on servers within the data center with access permitted only over the network. Most smartphone manufacturers provide a complete set of security features including full-disk encryption, email encryption, as well as remote management which includes the ability to wipe contents if device is lost or stolen. Also, some devices have embedded third-party antivirus and firewall software such as RIM's BlackBerry.

Transmission Security

Transmission security refers to measures that are designed to protect data from unauthorized interception, traffic analysis, and imitative deception. These measures include Secure Sockets Layer (SSL), iSeries Access for Windows, and virtual private network (VPN) connections. A secure data transmission should enable the identity of the sender and receiver to be verified by using a cryptographic shared key system as well as protect the data to be modified by a third party when it crosses the network. This can be done using AES or Triple DES with an encrypted SSL tunnel.

Authorization, Authentication, and Network Security

Authorization refers to the act of specifying access rights to control access of information to users. Authentication refers to the act of establishing or confirming the user as true or authentic. Network security refers to all the provisions and policies adopted by the network administrator to prevent and monitor unauthorized access, misuse, modification, or denial of the computer network and network-accessible resources. The mobility adds to unique security challenges. As data is trafficked beyond the enterprise firewall towards unknown territories, ensuring that it is handled safely is of paramount importance. Towards this, proper authentication of user connections, centralized access control (like LDAP Directory), encrypted data transfer mechanisms can be implemented.

Role of BI for Securing Mobile Apps

To ensure high security standards, BI software platforms must extend the authentication options and policy controls to the mobile platform. Business intelligence software platforms need to ensure a secure encrypted keychain for storage of credentials. Administrative control of password policies should allow creation of security profiles for each user and seamless integration with centralized security directories to reduce administration and maintenance of users.

Products

A number of BI vendors and niche software vendors offer mobile BI solutions. Some notable examples include:

- CollabMobile
- Cognos
- Cherrywork
- Dimensional Insight
- InetSoft
- Infor
- Information Builders

- MicroStrategy
- QlikView
- Roambi
- SAP
- Tableau Software
- Sisense
- TARGIT Business Intelligence

Real-time Business Intelligence

Real-time business intelligence (RTBI) is the process of delivering business intelligence (BI) or information about [business operations] as they occur. Real time means near to zero latency and access to information whenever it is required.

The speed of today's processing systems has moved classical data warehousing into the realm of real-time. The result is real-time business intelligence. Business transactions as they occur are fed to a real-time BI system that maintains the current state of the enterprise. The RTBI system not only supports the classic strategic functions of data warehousing for deriving information and knowledge from past enterprise activity, but it also provides real-time tactical support to drive enterprise actions that react immediately to events as they occur. As such, it replaces both the classic data warehouse and the enterprise application integration (EAI) functions. Such event-driven processing is a basic tenet of real-time business intelligence.

In this context, "real-time" means a range from milliseconds to a few seconds (5s) after the business event has occurred. While traditional BI presents historical data for manual analysis, RTBI compares current business events with historical patterns to detect problems or opportunities automatically. This automated analysis capability enables corrective actions to be initiated and/or business rules to be adjusted to optimize business processes.

RTBI is an approach in which up-to-a-minute data is analyzed, either directly from Operational sources or feeding business transactions into a real time data warehouse and Business Intelligence system. RTBI analyzes real time data.

Real-time business intelligence makes sense for some applications but not for others – a fact that organizations need to take into account as they consider investments in real-time BI tools. Key to deciding whether a real-time BI strategy would pay dividends is understanding the needs of the business and determining whether end users require immediate access to data for analytical purposes, or if something less than real time is fast enough.

Evolution of RTBI

In today's competitive environment with high consumer expectation, decisions that are based on the most current data available to improve customer relationships, increase revenue, maximize operational efficiencies, and yes – even save lives. This technology is real-time business intelligence. Real-time business intelligence systems provide the information necessary to strategically improve an enterprise's processes as well as to take tactical advantage of events as they occur.

Latency

All real-time business intelligence systems have some latency, but the goal is to minimize the time from the business event happening to a corrective action or notification being initiated. Analyst Richard Hackathorn describes three types of latency:

- Data latency; the time taken to collect and store the data
- Analysis latency; the time taken to analyze the data and turn it into actionable information
- Action latency; the time taken to react to the information and take action

Real-time business intelligence technologies are designed to reduce all three latencies to as close to zero as possible, whereas traditional business intelligence only seeks to reduce data latency and does not address analysis latency or action latency since both are governed by manual processes.

Some commentators have introduced the concept of *right time business intelligence* which proposes that information should be delivered just before it is required, and not necessarily in real-time.

Architectures

Event-based

Real-time Business Intelligence systems are event driven, and may use Complex Event Processing, Event Stream Processing and Mashup (web application hybrid) techniques to enable events to be analysed without being first transformed and stored in a database. These in- memory techniques have the advantage that high rates of events can be monitored, and since data does not have to be written into databases data latency can be reduced to milliseconds.

Data Warehouse

An alternative approach to event driven architectures is to increase the refresh cycle of an existing data warehouse to update the data more frequently. These real-time data warehouse systems can achieve near real-time update of data, where the data latency typically is in the range from minutes to hours. The analysis of the data is still usually manual, so the total latency is significantly different from event driven architectural approaches.

Server-less Technology

The latest alternative innovation to "real-time" event driven and/or "real-time" data warehouse architectures is MSSO Technology (Multiple Source Simple Output) which removes the need for the data warehouse and intermediary servers altogether since it is able to access live data directly from the source (even from multiple, disparate sources). Because live data is accessed directly by server-less means, it provides the potential for zero-latency, real-time data in the truest sense.

Process-aware

This is sometimes considered a subset of Operational intelligence and is also identified with Business Activity Monitoring. It allows entire processes (transactions, steps) to be monitored, metrics (latency, completion/failed ratios, etc.) to be viewed, compared with warehoused historic data, and trended in real-time. Advanced implementations allow threshold detection, alerting and providing feedback to the process execution systems themselves, thereby ‘closing the loop’.

Technologies that Support Real-time Analytics

Technologies that can be supported to enable real-time business intelligence are data visualization, data federation, enterprise information integration, enterprise application integration and service oriented architecture. Complex event processing tools can be used to analyze data streams in real time and either trigger automated actions or alert workers to patterns and trends.

Data Warehouse Appliance

> Data warehouse appliance is a combination of hardware and software product which was designed exclusively for analytical processing. In data warehouse implementation, tasks that involve tuning, adding or editing structure around the data, data migration from other databases, reconciliation of data are done by DBA. Another task for DBA was to make the database to perform well for large sets of users. Whereas with data warehouse appliances, it is the vendor responsibility of the physical design and tuning the software as per hardware requirements. Data warehouse appliance package comes with its own operating system, storage, DBMS, software, and required hardware. If required data warehouse appliances can be easily integrated with other tools.

Mobile Technology

> There are very limited vendors for providing Mobile business intelligence; MBI is integrated with existing BI architecture. MBI is a package that uses existing BI applications so people can use on their mobile phone and make informed decision in real time.

Application Areas

- Algorithmic trading
- Fraud detection
- Systems monitoring
- Application performance monitoring
- Customer Relationship Management
- Demand sensing
- Dynamic pricing and yield management
- Data validation
- Operational intelligence and risk management

- Payments & cash monitoring
- Data security monitoring
- Supply chain optimization
- RFID/sensor network data analysis
- Workstreaming
- Call center optimization
- Enterprise Mashups and Mashup Dashboards
- Transportation industry

Transportation industry can be benefited by using real-time analytics. For an example railroad network. Depending on the results provided by the real-time analytics, dispatcher can make a decision on what kind of train he can dispatch on the track depending on the train traffic and commodities shipped.

References

- (Rud, Olivia (2009). Business Intelligence Success Factors: Tools for Aligning Your Business in the Global Economy. Hoboken, N.J: Wiley & Sons. ISBN 978-0-470-39240-9.)
- Coker, Frank (2014). Pulse: Understanding the Vital Signs of Your Business. Ambient Light Publishing. pp. 41–42. ISBN 978-0-9893086-0-1.
- Jeanne W. Ross, Peter Weill, David C. Robertson (2006) Enterprise Architecture As Strategy, p. 117 ISBN Julian, Taylor (10 January 2010). "Business intelligence implementation according to customer's needs". APRO Software. Retrieved 16 May 2016.
- "How Companies Are Implementing Business Intelligence Competency Centers" (PDF). Computer World. Archived from the original (PDF) on 28 May 2013. Retrieved 1 April 2014.

2

Analytics: A Comprehensive Study

Analytics is the understanding and communication of significant patterns of data. Analytics is applied in businesses to improve their performances. Some of the aspects explained in this text are software analytics, embedded analytics, learning analytics and social media analytics. The section on analytics offers an insightful focus, keeping in mind the complex subject matter.

Business Analytics

Business analytics (BA) refers to the skills, technologies, practices for continuous iterative exploration and investigation of past business performance to gain insight and drive business planning. Business analytics focuses on developing new insights and understanding of business performance based on data and statistical methods. In contrast, business intelligence traditionally focuses on using a consistent set of metrics to both measure past performance and guide business planning, which is also based on data and statistical methods.

Business analytics makes extensive use of statistical analysis, including explanatory and predictive modeling, and fact-based management to drive decision making. It is therefore closely related to management science. Analytics may be used as input for human decisions or may drive fully automated decisions. Business intelligence is querying, reporting, online analytical processing (OLAP), and "alerts."

In other words, querying, reporting, OLAP, and alert tools can answer questions such as what happened, how many, how often, where the problem is, and what actions are needed. Business analytics can answer questions like why is this happening, what if these trends continue, what will happen next (that is, predict), what is the best that can happen (that is, optimize).

Examples of Application

Banks, such as Capital One, use data analysis (or *analytics*, as it is also called in the business setting), to differentiate among customers based on credit risk, usage and other characteristics and then to match customer characteristics with appropriate product offerings. Harrah's, the gaming firm, uses analytics in its customer loyalty programs. E & J Gallo Winery quantitatively analyzes and predicts the appeal of its wines. Between 2002 and 2005, Deere & Company saved more than $1 billion by employing a new analytical tool to better optimize inventory. A telecoms company that pursues efficient call centre usage over customer service may save money.

Types of Analytics

- Decision analytics: supports human decisions with visual analytics the user models to reflect reasoning.

- Descriptive analytics: gains insight from historical data with reporting, scorecards, clustering etc.
- Predictive analytics: employs predictive modeling using statistical and machine learning techniques
- Prescriptive analytics: recommends decisions using optimization, simulation, etc.

Basic Domains within Analytics

- Behavioral analytics
- Cohort Analysis
- Collections analytics
- Contextual data modeling - supports the human reasoning that occurs after viewing "executive dashboards" or any other visual analytics
- Cyber analytics
- Enterprise Optimization
- Financial services analytics
- Fraud analytics
- Marketing analytics
- Pricing analytics
- Retail sales analytics
- Risk & Credit analytics
- Supply Chain analytics
- Talent analytics
- Telecommunications
- Transportation analytics

History

Analytics have been used in business since the management exercises were put into place by Frederick Winslow Taylor in the late 19th century. Henry Ford measured the time of each component in his newly established assembly line. But analytics began to command more attention in the late 1960s when computers were used in decision support systems. Since then, analytics have changed and formed with the development of enterprise resource planning (ERP) systems, data warehouses, and a large number of other software tools and processes.

In later years the business analytics have exploded with the introduction to computers. This change has brought analytics to a whole new level and has made the possibilities endless. As far as analytics has come in history, and what the current field of analytics is today many people would never think that analytics started in the early 1900s with Mr. Ford himself.

Challenges

Business analytics depends on sufficient volumes of high quality data. The difficulty in ensuring data quality is integrating and reconciling data across different systems, and then deciding what subsets of data to make available.

Previously, analytics was considered a type of after-the-fact method of forecasting consumer behavior by examining the number of units sold in the last quarter or the last year. This type of data warehousing required a lot more storage space than it did speed. Now business analytics is becoming a tool that can influence the outcome of customer interactions. When a specific customer type is considering a purchase, an analytics-enabled enterprise can modify the sales pitch to appeal to that consumer. This means the storage space for all that data must react extremely fast to provide the necessary data in real-time.

Competing on Analytics

Thomas Davenport, professor of information technology and management at Babson College argues that businesses can optimize a distinct business capability via analytics and thus better compete. He identifies these characteristics of an organization that are apt to compete on analytics:

- One or more senior executives who strongly advocate fact-based decision making and, specifically, analytics
- Widespread use of not only descriptive statistics, but also predictive modeling and complex optimization techniques
- Substantial use of analytics across multiple business functions or processes
- Movement toward an enterprise level approach to managing analytical tools, data, and organizational skills and capabilities

Analytics

Analytics is the discovery, interpretation, and communication of meaningful patterns in data. Especially valuable in areas rich with recorded information, analytics relies on the simultaneous application of statistics, computer programming and operations research to quantify performance. Analytics often favors data visualization to communicate insight.

Organizations may apply analytics to business data to describe, predict, and improve business performance. Specifically, areas within analytics include predictive analytics, prescriptive analytics, enterprise decision management, retail analytics, store assortment and stock-keeping unit optimization, marketing optimization and marketing mix modeling, web analytics, sales force sizing and optimization, price and promotion modeling, predictive science, credit risk analysis, and fraud analytics. Since analytics can require extensive computation, the algorithms and software used for analytics harness the most current methods in computer science, statistics, and mathematics.

Analytics vs. Analysis

Analytics is multidisciplinary. There is extensive use of mathematics and statistics, the use of descriptive techniques and predictive models to gain valuable knowledge from data—data analysis. The insights from data are used to recommend action or to guide decision making rooted in business context. Thus, analytics is not so much concerned with individual analyses or analysis steps, but with the entire methodology. There is a pronounced tendency to use the term *analytics* in business settings e.g. text analytics vs. the more generic text mining to emphasize this broader perspective.. There is an increasing use of the term *advanced analytics*, typically used to describe the technical aspects of analytics, especially in the emerging fields such as the use of machine learning techniques like neural networks to do predictive modeling.

Examples

Marketing Optimization

Marketing has evolved from a creative process into a highly data-driven process. Marketing organizations use analytics to determine the outcomes of campaigns or efforts and to guide decisions for investment and consumer targeting. Demographic studies, customer segmentation, conjoint analysis and other techniques allow marketers to use large amounts of consumer purchase, survey and panel data to understand and communicate marketing strategy.

Web analytics allows marketers to collect session-level information about interactions on a website using an operation called sessionization. Google Analytics is an example of a popular free analytics tool that marketers use for this purpose. Those interactions provide web analytics information systems with the information necessary to track the referrer, search keywords, identify IP address, and track activities of the visitor. With this information, a marketer can improve marketing campaigns, website creative content, and information architecture.

Analysis techniques frequently used in marketing include marketing mix modeling, pricing and promotion analyses, sales force optimization and customer analytics e.g.: segmentation. Web analytics and optimization of web sites and online campaigns now frequently work hand in hand with the more traditional marketing analysis techniques. A focus on digital media has slightly changed the vocabulary so that *marketing mix modeling* is commonly referred to as *attribution modeling* in the digital or marketing mix modeling context.

These tools and techniques support both strategic marketing decisions (such as how much overall to spend on marketing, how to allocate budgets across a portfolio of brands and the marketing mix) and more tactical campaign support, in terms of targeting the best potential customer with the optimal message in the most cost effective medium at the ideal time.

Portfolio Analytics

A common application of business analytics is portfolio analysis. In this, a bank or lending agency has a collection of accounts of varying value and risk. The accounts may differ by the social status (wealthy, middle-class, poor, etc.) of the holder, the geographical location, its net value, and many other factors. The lender must balance the return on the loan with the risk of default for each loan. The question is then how to evaluate the portfolio as a whole.

The least risk loan may be to the very wealthy, but there are a very limited number of wealthy people. On the other hand, there are many poor that can be lent to, but at greater risk. Some balance must be struck that maximizes return and minimizes risk. The analytics solution may combine time series analysis with many other issues in order to make decisions on when to lend money to these different borrower segments, or decisions on the interest rate charged to members of a portfolio segment to cover any losses among members in that segment.

Risk Analytics

Predictive models in the banking industry are developed to bring certainty across the risk scores for individual customers. Credit scores are built to predict individual's delinquency behavior and widely used to evaluate the credit worthiness of each applicant. Furthermore, risk analyses are carried out in the scientific world and the insurance industry. It is also extensively used in financial institutions like Online Payment Gateway companies to analyse if a transaction was genuine or fraud. For this purpose they use the transaction history of the customer. This is more commonly used in Credit Card purchase, when there is a sudden spike in the customer transaction volume the customer gets a call of confirmation if the transaction was initiated by him/her. This helps in reducing loss due to such circumstances.

Digital Analytics

Digital analytics is a set of business and technical activities that define, create, collect, verify or transform digital data into reporting, research, analyses, recommendations, optimizations, predictions, and automations. This also includes the SEO (Search Engine Optimization) where the keyword search is tracked and that data is used for marketing purposes. Even banner ads and clicks come under digital analytics. All marketing firms rely on digital analytics for their digital marketing assignments, where MROI (Marketing Return on Investment) is important.

Security Analytics

Security analytics refers to information technology (IT) solutions that gather and analyze security events to bring situational awareness and enable IT staff to understand and analyze events that pose the greatest risk. Solutions in this area include security information and event management solutions and user behavior analytics solutions.

Software Analytics

Software analytics is the process of collecting information about the way a piece of software is used and produced.

Challenges

In the industry of commercial analytics software, an emphasis has emerged on solving the challenges of analyzing massive, complex data sets, often when such data is in a constant state of change. Such data sets are commonly referred to as big data. Whereas once the problems posed by big data were only found in the scientific community, today big data is a problem for many businesses that operate transactional systems online and, as a result, amass large volumes of data quickly.

The analysis of unstructured data types is another challenge getting attention in the industry. Unstructured data differs from structured data in that its format varies widely and cannot be stored in traditional relational databases without significant effort at data transformation. Sources of unstructured data, such as email, the contents of word processor documents, PDFs, geospatial data, etc., are rapidly becoming a relevant source of business intelligence for businesses, governments and universities. For example, in Britain the discovery that one company was illegally selling fraudulent doctor's notes in order to assist people in defrauding employers and insurance companies, is an opportunity for insurance firms to increase the vigilance of their unstructured data analysis. The McKinsey Global Institute estimates that big data analysis could save the American health care system $300 billion per year and the European public sector €250 billion.

These challenges are the current inspiration for much of the innovation in modern analytics information systems, giving birth to relatively new machine analysis concepts such as complex event processing, full text search and analysis, and even new ideas in presentation. One such innovation is the introduction of grid-like architecture in machine analysis, allowing increases in the speed of massively parallel processing by distributing the workload to many computers all with equal access to the complete data set.

Analytics is increasingly used in education, particularly at the district and government office levels. However, the complexity of student performance measures presents challenges when educators try to understand and use analytics to discern patterns in student performance, predict graduation likelihood, improve chances of student success, etc. For example, in a study involving districts known for strong data use, 48% of teachers had difficulty posing questions prompted by data, 36% did not comprehend given data, and 52% incorrectly interpreted data. To combat this, some analytics tools for educators adhere to an over-the-counter data format (embedding labels, supplemental documentation, and a help system, and making key package/display and content decisions) to improve educators' understanding and use of the analytics being displayed.

One more emerging challenge is dynamic regulatory needs. For example, in the banking industry, Basel III and future capital adequacy needs are likely to make even smaller banks adopt internal risk models. In such cases, cloud computing and open source R (programming language) can help smaller banks to adopt risk analytics and support branch level monitoring by applying predictive analytics.

Risks

The main risk for the people is discrimination like price discrimination or statistical discrimination. Analytical processes can also result in discriminatory outcomes that may violate anti-discrimination and civil rights laws. There is also the risk that a developer could profit from the ideas or work done by users, like this example: Users could write new ideas in a note taking app, which could then be sent as a custom event, and the developers could profit from those ideas. This can happen because the ownership of content is usually unclear in the law.

If a user's identity is not protected, there are more risks; for example, the risk that private information about users is made public on the internet.

In the extreme, there is the risk that governments could gather too much private information, now that the governments are giving themselves more powers to access citizens' information.

Software Analytics

Software Analytics refers to analytics specific to software systems and related software development processes. It aims at describing, predicting, and improving development, maintenance, and management of complex software systems. Methods and techniques of software analytics typically rely on gathering, analyzing, and visualizing information found in the manifold data sources in the scope of software systems and their software development processes---software analytics "turns it into actionable insight to inform better decisions related to software".

Software analytics represents a base component of software diagnosis that generally aims at generating findings, conclusions, and evaluations about software systems and their implementation, composition, behavior, and evolution. Software analytics frequently uses and combines approaches and techniques from statistics, prediction analysis, data mining, and scientific visualization. For example, software analytics can map data by means of software maps that allow for interactive exploration.

Data under exploration and analysis by Software Analytics exists in software lifecycle, including source code, software requirement specifications, bug reports, test cases, execution traces/logs, and real-world user feedback, etc. Data plays a critical role in modern software development, because hidden in the data is the information and insight about the quality of software and services, the experience that software users receive, as well as the dynamics of software development.

Insightful information obtained by Software Analytics is information that conveys meaningful and useful understanding or knowledge towards performing the target task. Typically insightful information cannot be easily obtained by direct investigation on the raw data without the aid of analytic technologies.

Actionable information obtained by Software Analytics is information upon which software practitioners can come up with concrete solutions (better than existing solutions if any) towards completing the target task.

Software Analytics focuses on trinity of software systems, software users, and software development process:

Software Systems. Depending on scale and complexity, the spectrum of software systems can span from operating systems for devices to large networked systems that consist of thousands of servers. System quality such as reliability, performance and security, etc., is the key to success of modern software systems. As the system scale and complexity greatly increase, larger amount of data, e.g., run-time traces and logs, is generated; and data becomes a critical means to monitor, analyze, understand and improve system quality.

Software Users. Users are (almost) always right because ultimately they will use the software and services in various ways. Therefore, it is important to continuously provide the best experience to users. Usage data collected from the real world reveals how users interact with software and services. The data is incredibly valuable for software practitioners to better understand their customers and gain insights on how to improve user experience accordingly.

Software Development Process. Software development has evolved from its traditional form to exhibiting different characteristics. The process is more agile and engineers are more collaborative than that in the past. Analytics on software development data provides a powerful mechanism that software practitioners can leverage to achieve higher development productivity.

In general, the primary technologies employed by Software Analytics include analytical technologies such as machine learning, data mining and pattern recognition, information visualization, as well as large-scale data computing & processing.

History

In May 2009, Software Analytics was first coined and proposed when Dr. Dongmei Zhang founded the Software Analytics Group (SA) at Microsoft Research Asia (MSRA). The term has become well known in the software engineering research community after a series of tutorials and talks on software analytics were given by Dr. Dongmei Zhang and her colleagues, in collaboration with Professor Tao Xie from North Carolina State University, at software engineering conferences including a tutorial at the IEEE/ACM International Conference on Automated Software Engineering (ASE 2011), a talk at the International Workshop on Machine Learning Technologies in Software Engineering (MALETS 2011), a tutorial and a keynote talk given by Dr. Dongmei Zhang at the IEEE-CS Conference on Software Engineering Education and Training (CSEE&T 2012), a tutorial at the International Conference on Software Engineering (ICSE 2012) - Software Engineering in Practice Track, and a keynote talk given by Dr. Dongmei Zhang at the Working Conference on Mining Software Repositories (MSR 2012).

In November 2010, Software Development Analytics (Software Analytics with focus on Software Development) was proposed by Thomas Zimmermann and his colleagues at the Empirical Software Engineering Group (ESE) at Microsoft Research Redmond in their FoSER 2010 paper. A goldfish bowl panel on software development analytics was organized by Thomas Zimmermann and Professor Tim Menzies from West Virginia University at the International Conference on Software Engineering (ICSE 2012), Software Engineering in Practice track.

Software Analytics Providers

- CAST Software
- IBM Cognos Business Intelligence
- Kiuwan
- Microsoft Azure Application Insights
- Nalpeiron Software Analytics
- New Relic

- Squore
- Tableau Software
- Trackerbird Software Analytics

Embedded Analytics

Embedded analytics is the technology designed to make data analysis and business intelligence more accessible by all kind of application or user.

Definition

According to Gartner analysts Kurt Schlegel, traditional business intelligence were suffering in 2008 a lack of integration between the data and the business users. This technology intention is to be more pervasive by real-time autonomy and self-service of data visualization or customization, meanwhile decision makers, business users or even customers are doing their own daily workflow and tasks.

History

First mentions of the concept were made by Howard Dresner, consultant, author, former Gartner analyst and inventor of the term “business intelligence”. Consolidation of business intelligence “doesn’t mean the BI market has reached maturity” said Howard Dresner while he was working for Hyperion Solutions, a company that Oracle bought in 2007. Oracle started then to use the term “embedded analytics” at their press release for Oracle® Rapid Planning on 2009. Gartner Group, a company for which Howard Dresner has been working, finally added the term to their IT Glossary on November 5, 2012. . It was clear this was a mainstream technology when Dresner Advisory Services published the 2014 Embedded Business Intelligence Market Study as part of the Wisdom of Crowds® Series of Research, including 24 vendors.

Tools

- Actuate
- Dundas Data Visualization
- GoodData
- IBM
- icCube
- Logi Analytics
- Pentaho
- Qlik
- SAP

- SAS
- Tableau
- TIBCO
- Sisense

Learning Analytics

Learning analytics is the measurement, collection, analysis and reporting of data about learners and their contexts, for purposes of understanding and optimizing learning and the environments in which it occurs. A related field is educational data mining. For general audience introductions, see:

- The Educause Learning Initiative Briefing
- The Educause Review on Learning analytics
- And the UNESCO “Learning Analytics Policy Brief” (2012)

What is Learning Analytics?

The definition and aims of Learning Analytics are contested. One earlier definition discussed by the community suggested that “Learning analytics is the use of intelligent data, learner-produced data, and analysis models to discover information and social connections for predicting and advising people’s learning.”

But this definition has been criticised:

1. *“I somewhat disagree with this definition - it serves well as an introductory concept if we use analytics as a support structure for existing education models. I think learning analytics - at an advanced and integrated implementation - can do away with pre-fab curriculum models”*. George Siemens, 2010.
2. *“In the descriptions of learning analytics we talk about using data to “predict success”. I’ve struggled with that as I pore over our databases. I’ve come to realize there are different views/levels of success.”* Mike Sharkey 2010.

A more holistic view than a mere definition is provided by the framework of learning analytics by Greller and Drachsler (2012). It uses a general morphological analysis (GMA) to divide the domain into six “critical dimensions”.

A systematic overview on learning analytics and its key concepts is provided by Chatti et al. (2012) and Chatti et al. (2014) through a reference model for learning analytics based on four dimensions, namely data, environments, context (what?), stakeholders (who?), objectives (why?), and methods (how?).

It has been pointed out that there is a broad awareness of analytics across educational institutions for various stakeholders, but that the way ‘learning analytics’ is defined and implemented may vary, including:

1. for individual learners to reflect on their achievements and patterns of behaviour in relation to others;
2. as predictors of students requiring extra support and attention;
3. to help teachers and support staff plan supporting interventions with individuals and groups;
4. for functional groups such as course team seeking to improve current courses or develop new curriculum offerings; and
5. for institutional administrators taking decisions on matters such as marketing and recruitment or efficiency and effectiveness measures."

In that briefing paper, Powell and MacNeill go on to point out that some motivations and implementations of analytics may come into conflict with others, for example highlighting potential conflict between analytics for individual learners and organisational stakeholders.

Gašević, Dawson, and Siemens argue that computational aspects of learning analytics need to be linked with the existing educational research if the field of learning analytics is to deliver to its promise to understand and optimize learning.

Differentiating Learning Analytics and Educational Data Mining

Differentiating the fields of educational data mining (EDM) and learning analytics (LA) has been a concern of several researchers. George Siemens takes the position that educational data mining encompasses both learning analytics and academic analytics, the former of which is aimed at governments, funding agencies, and administrators instead of learners and faculty. Baepler and Murdoch define academic analytics as an area that *"...combines select institutional data, statistical analysis, and predictive modeling to create intelligence upon which learners, instructors, or administrators can change academic behavior"*. They go on to attempt to disambiguate educational data mining from academic analytics based on whether the process is hypothesis driven or not, though Brooks questions whether this distinction exists in the literature. Brooks instead proposes that a better distinction between the EDM and LA communities is in the roots of where each community originated, with authorship at the EDM community being dominated by researchers coming from intelligent tutoring paradigms, and learning anaytics researchers being more focused on enterprise learning systems (e.g. learning content management systems).

Regardless of the differences between the LA and EDM communities, the two areas have significant overlap both in the objectives of investigators as well as in the methods and techniques that are used in the investigation. In the MS program offering in Learning Analytics at Teachers College, Columbia University, students are taught both EDM and LA methods.

History

The Context of Learning Analytics

In "The State of Learning Analytics in 2012: A Review and Future Challenges" Rebecca Ferguson tracks the progress of analytics for learning as a development through:

1. The increasing interest in 'big data' for business intelligence

2. The rise of online education focussed around Virtual Learning Environments (VLEs), Content Management Systems (CMSs), and Management Information Systems (MIS) for education, which saw an increase in digital data regarding student background (often held in the MIS) and learning log data (from VLEs). This development afforded the opportunity to apply 'business intelligence' techniques to educational data
3. Questions regarding the optimisation of systems to support learning particularly given the question regarding how we can know whether a student is engaged/understanding if we can't see them?
4. Increasing focus on evidencing progress and professional standards for accountability systems
5. This focus led to a teacher stakehold in the analytics - given that they are associated with accountability systems
6. Thus an increasing emphasis was placed on the pedagogic affordances of learning analytics
7. This pressure is increased by the economic desire to improve engagement in online education for the deliverance of high quality - affordable - education

History of The Techniques and Methods of Learning Analytics

In a discussion of the history of analytics, Cooper highlights a number of communities from which learning analytics draws techniques, including:

1. Statistics - which are a well established means to address hypothesis testing
2. Business Intelligence - which has similarities with learning analytics, although it has historically been targeted at making the production of reports more efficient through enabling data access and summarising performance indicators.
3. Web analytics - tools such as Google analytics report on web page visits and references to websites, brands and other keyterms across the internet. The more 'fine grain' of these techniques can be adopted in learning analytics for the exploration of student trajectories through learning resources (courses, materials, etc.).
4. Operational research - aims at highlighting design optimisation for maximising objectives through the use of mathematical models and statistical methods. Such techniques are implicated in learning analytics which seek to create models of real world behaviour for practical application.
5. Artificial intelligence and Data mining - Machine learning techniques built on data mining and AI methods are capable of detecting patterns in data. In learning analytics such techniques can be used for intelligent tutoring systems, classification of students in more dynamic ways than simple demographic factors, and resources such as 'suggested course' systems modelled on collaborative filtering techniques.
6. Social Network Analysis - SNA analyses relationships between people by exploring implicit (e.g. interactions on forums) and explicit (e.g. 'friends' or 'followers') ties online and offline. SNA developed from the work of sociologists like Wellman and Watts, and mathe-

maticians like Barabasi and Strogatz. The work of these individuals has provided us with a good sense of the patterns that networks exhibit (small world, power laws), the attributes of connections (in early 70's, Granovetter explored connections from a perspective of tie strength and impact on new information), and the social dimensions of networks (for example, geography still matters in a digital networked world). It is particularly used to explore clusters of networks, influence networks, engagement and disengagement, and has been deployed for these purposes in learning analytic contexts.

7. Information visualization - visualisation is an important step in many analytics for sensemaking around the data provided - it is thus used across most techniques (including those above).

History of Learning Analytics in Higher Education

The first graduate program focused specifically on learning analytics was created by Dr. Ryan Baker and launched in the Fall 2015 semester at Teachers College - Columbia University. The program description states that "data about learning and learners are being generated today on an unprecedented scale. The fields of learning analytics (LA) and educational data mining (EDM) have emerged with the aim of transforming this data into new insights that can benefit students, teachers, and administrators. As one of world's leading teaching and research institutions in education, psychology, and health, we are proud to offer an innovative graduate curriculum dedicated to improving education through technology and data analysis."

Analytic Methods

Methods for learning analytics include:

- Content analysis - particularly of resources which students create (such as essays)
- Discourse Analytics Discourse analytics aims to capture meaningful data on student interactions which (unlike 'social network analytics') aims to explore the properties of the language used, as opposed to just the network of interactions, or forum-post counts, etc.
- Social Learning Analytics which is aimed at exploring the role of social interaction in learning, the importance of learning networks, discourse used to sensemake, etc.
- Disposition Analytics which seeks to capture data regarding student's dispositions to their own learning, and the relationship of these to their learning. For example, "curious" learners may be more inclined to ask questions - and this data can be captured and analysed for learning analytics.

Analytic Outcomes

Analytics have been used for:

- Prediction purposes, for example to identify 'at risk' students in terms of drop out or course failure
- Personalization & Adaptation, to provide students with tailored learning pathways, or assessment materials

- Intervention purposes, providing educators with information to intervene to support students
- Information visualization, typically in the form of so-called learning dashboards which provide overview learning data through data visualisation tools

Software

Much of the software that is currently used for learning analytics duplicates functionality of web analytics software, but applies it to learner interactions with content. Social network analysis tools are commonly used to map social connections and discussions. Some examples of learning analytics software tools:

- Student Success System - a predictive learning analytics tool that predicts student performance and plots learners into risk quadrants based upon engagement and performance predictions, and provides indicators to develop understanding as to why a learner is not on track through visualizations such as the network of interactions resulting from social engagement (e.g. discussion posts and replies), performance on assessments, engagement with content, and other indicators
- SNAPP - a learning analytics tool that visualizes the network of interactions resulting from discussion forum posts and replies.
- LOCO-Analyst - a context-aware learning tool for analytics of learning processes taking place in a web-based learning environment
- SAM - a Student Activity Monitor intended for Personal Learning Environments
- BEESTAR INSIGHT - a real-time system that automatically collects student engagement and attendance & provides analytics tools and dashboards for students, teachers & management
- Solutionpath StREAM- A leading UK based real-time system that leverage predictive models to determine all facets of student engagement using structured and unstructured sources for all institutional roles

Ethics & Privacy

The ethics of data collection, analytics, reporting and accountability has been raised as a potential concern for Learning Analytics (e.g.,), with concerns raised regarding:

- Data ownership
- Communications around the scope and role of Learning Analytics
- The necessary role of human feedback and error-correction in Learning Analytics systems
- Data sharing between systems, organisations, and stakeholders
- Trust in data clients

As Kay, Kom and Oppenheim point out, the range of data is wide, potentially derived from: "*Recorded activity; student records, attendance, assignments, researcher information (CRIS).

- Systems interactions; VLE, library / repository search, card transactions.
- Feedback mechanisms; surveys, customer care.
- External systems that offer reliable identification such as sector and shared services and social networks."

Thus the legal and ethical situation is challenging and different from country to country, raising implications for: "*Variety of data - principles for collection, retention and exploitation.

- Education mission - underlying issues of learning management, including social and performance engineering.
- Motivation for development of analytics – mutuality, a combination of corporate, individual and general good.
- Customer expectation – effective business practice, social data expectations, cultural considerations of a global customer base. *Obligation to act – duty of care arising from knowledge and the consequent challenges of student and employee performance management."

In some prominent cases like the inBloom disaster even full functional systems have been shut down due to lack of trust in the data collection by governments, stakeholders and civil rights groups. Since then, the Learning Analytics community has extensively studied legal conditions in a series of experts workshops on 'Ethics & Privacy 4 Learning Analytics' that constitute the use of trusted Learning Analytics. Drachsler & Greller released a 8-point checklist named DELICATE that is based on the intensive studies in this area to demystify the ethics and privacy discussions around Learning Analytics.

1. D-etermination: Decide on the purpose of learning analytics for your institution.
2. E-xplain: Define the scope of data collection and usage.
3. L-egitimate: Explain how you operate within the legal frameworks, refer to the essential legislation.
4. I-nvolve: Talk to stakeholders and give assurances about the data distribution and use.
5. C-onsent: Seek consent through clear consent questions.
6. A-nonymise: De-identify individuals as much as possible
7. T-echnical aspects: Monitor who has access to data, especially in areas with high staff turnover.
8. E-xternal partners: Make sure externals provide highest data security standards

It shows ways to design and provide privacy conform Learning Analytics that can benefit all stakeholders. The full DELICATE checklist is publicly available here.

Open Learning Analytics

Chatti, Muslim and Schroeder note that the aim of Open Learning Analytics (OLA) is to improve learning effectiveness in lifelong learning environments. The authors refer to OLA as an ongoing analytics process that encompasses diversity at all four dimensions of the learning analytics reference model.

Predictive Analytics

Predictive analytics encompasses a variety of statistical techniques from predictive modeling, machine learning, and data mining that analyze current and historical facts to make predictions about future or otherwise unknown events.

In business, predictive models exploit patterns found in historical and transactional data to identify risks and opportunities. Models capture relationships among many factors to allow assessment of risk or potential associated with a particular set of conditions, guiding decision making for candidate transactions.

The defining functional effect of these technical approaches is that predictive analytics provides a predictive score (probability) for each individual (customer, employee, healthcare patient, product SKU, vehicle, component, machine, or other organizational unit) in order to determine, inform, or influence organizational processes that pertain across large numbers of individuals, such as in marketing, credit risk assessment, fraud detection, manufacturing, healthcare, and government operations including law enforcement.

Predictive analytics is used in actuarial science, marketing, financial services, insurance, telecommunications, retail, travel, healthcare, child protection, pharmaceuticals, capacity planning and other fields.

One of the most well known applications is credit scoring, which is used throughout financial services. Scoring models process a customer's credit history, loan application, customer data, etc., in order to rank-order individuals by their likelihood of making future credit payments on time.

Definition

Predictive analytics is an area of data mining that deals with extracting information from data and using it to predict trends and behavior patterns. Often the unknown event of interest is in the future, but predictive analytics can be applied to any type of unknown whether it be in the past, present or future. For example, identifying suspects after a crime has been committed, or credit card fraud as it occurs. The core of predictive analytics relies on capturing relationships between explanatory variables and the predicted variables from past occurrences, and exploiting them to predict the unknown outcome. It is important to note, however, that the accuracy and usability of results will depend greatly on the level of data analysis and the quality of assumptions.

Predictive analytics is often defined as predicting at a more detailed level of granularity, i.e., generating predictive scores (probabilities) for each individual organizational element. This distinguishes it from forecasting. For example, "Predictive analytics—Technology that learns from experience (data) to predict the future behavior of individuals in order to drive better decisions." In future industrial systems, the value of predictive analytics will be to predict and prevent potential issues to achieve near-zero break-down and further be integrated into prescriptive analytics for decision optimization. Furthermore, the converted data can be used for closed-loop product life cycle improvement which is the vision of Industrial Internet Consortium.

Types

Generally, the term predictive analytics is used to mean predictive modeling, "scoring" data with predictive models, and forecasting. However, people are increasingly using the term to refer to related analytical disciplines, such as descriptive modeling and decision modeling or optimization. These disciplines also involve rigorous data analysis, and are widely used in business for segmentation and decision making, but have different purposes and the statistical techniques underlying them vary.

Predictive Models

Predictive models are models of the relation between the specific performance of a unit in a sample and one or more known attributes or features of the unit. The objective of the model is to assess the likelihood that a similar unit in a different sample will exhibit the specific performance. This category encompasses models in many areas, such as marketing, where they seek out subtle data patterns to answer questions about customer performance, or fraud detection models. Predictive models often perform calculations during live transactions, for example, to evaluate the risk or opportunity of a given customer or transaction, in order to guide a decision. With advancements in computing speed, individual agent modeling systems have become capable of simulating human behaviour or reactions to given stimuli or scenarios.

The available sample units with known attributes and known performances is referred to as the "training sample". The units in other samples, with known attributes but unknown performances, are referred to as "out of [training] sample" units. The out of sample bear no chronological relation to the training sample units. For example, the training sample may consists of literary attributes of writings by Victorian authors, with known attribution, and the out-of sample unit may be newly found writing with unknown authorship; a predictive model may aid in attributing a work to a known author. Another example is given by analysis of blood splatter in simulated crime scenes in which the out of sample unit is the actual blood splatter pattern from a crime scene. The out of sample unit may be from the same time as the training units, from a previous time, or from a future time.

Descriptive Models

Descriptive models quantify relationships in data in a way that is often used to classify customers or prospects into groups. Unlike predictive models that focus on predicting a single customer behavior (such as credit risk), descriptive models identify many different relationships between customers or products. Descriptive models do not rank-order customers by their likelihood of taking a particular action the way predictive models do. Instead, descriptive models can be used, for example, to categorize customers by their product preferences and life stage. Descriptive modeling tools can be utilized to develop further models that can simulate large number of individualized agents and make predictions.

Decision Models

Decision models describe the relationship between all the elements of a decision—the known data (including results of predictive models), the decision, and the forecast results of the decision—in order to predict the results of decisions involving many variables. These models can be used in optimization, maximizing certain outcomes while minimizing others. Decision models are generally used to develop decision logic or a set of business rules that will produce the desired action for every customer or circumstance.

Applications

Although predictive analytics can be put to use in many applications, we outline a few examples where predictive analytics has shown positive impact in recent years.

Analytical Customer Relationship Management (CRM)

Analytical customer relationship management (CRM) is a frequent commercial application of predictive analysis. Methods of predictive analysis are applied to customer data to pursue CRM objectives, which involve constructing a holistic view of the customer no matter where their information resides in the company or the department involved. CRM uses predictive analysis in applications for marketing campaigns, sales, and customer services to name a few. These tools are required in order for a company to posture and focus their efforts effectively across the breadth of their customer base. They must analyze and understand the products in demand or have the potential for high demand, predict customers' buying habits in order to promote relevant products at multiple touch points, and proactively identify and mitigate issues that have the potential to lose customers or reduce their ability to gain new ones. Analytical customer relationship management can be applied throughout the customers lifecycle (acquisition, relationship growth, retention, and win-back). Several of the application areas described below (direct marketing, cross-sell, customer retention) are part of customer relationship management.

Child Protection

Over the last 5 years, some child welfare agencies have started using predictive analytics to flag high risk cases. The approach has been called "innovative" by the Commission to Eliminate Child Abuse and Neglect Fatalities (CECANF), and in Hillsborough County, Florida, where the lead child welfare agency uses a predictive modeling tool, there have been no abuse-related child deaths in the target population as of this writing.

Clinical Decision Support Systems

Experts use predictive analysis in health care primarily to determine which patients are at risk of developing certain conditions, like diabetes, asthma, heart disease, and other lifetime illnesses. Additionally, sophisticated clinical decision support systems incorporate predictive analytics to support medical decision making at the point of care. A working definition has been proposed by

Jerome A. Osheroff and colleagues: *Clinical decision support (**CDS**) provides clinicians, staff, patients, or other individuals with knowledge and person-specific information, intelligently filtered or presented at appropriate times, to enhance health and health care. It encompasses a variety of tools and interventions such as computerized alerts and reminders, clinical guidelines, order sets, patient data reports and dashboards, documentation templates, diagnostic support, and clinical workflow tools.*

A 2016 study of neurodegenerative disorders provides a powerful example of a CDS platform to diagnose, track, predict and monitor the progression of Parkinson's disease. Using large and multi-source imaging, genetics, clinical and demographic data, these investigators developed a decision support system that can predict the state of the disease with high accuracy, consistency and precision. They employed classical model-based and machine learning model-free methods to discriminate between different patient and control groups. Similar approaches may be used for predictive diagnosis and disease progression forecasting in many neurodegenerative disorders like Alzheimer's, Huntington's, Amyotrophic Lateral Sclerosis, as well as for other clinical and biomedical applications where Big Data is available.

Collection Analytics

Many portfolios have a set of delinquent customers who do not make their payments on time. The financial institution has to undertake collection activities on these customers to recover the amounts due. A lot of collection resources are wasted on customers who are difficult or impossible to recover. Predictive analytics can help optimize the allocation of collection resources by identifying the most effective collection agencies, contact strategies, legal actions and other strategies to each customer, thus significantly increasing recovery at the same time reducing collection costs.

Cross-sell

Often corporate organizations collect and maintain abundant data (e.g. customer records, sale transactions) as exploiting hidden relationships in the data can provide a competitive advantage. For an organization that offers multiple products, predictive analytics can help analyze customers' spending, usage and other behavior, leading to efficient cross sales, or selling additional products to current customers. This directly leads to higher profitability per customer and stronger customer relationships.

Customer Retention

With the number of competing services available, businesses need to focus efforts on maintaining continuous customer satisfaction, rewarding consumer loyalty and minimizing customer attrition. In addition, small increases in customer retention have been shown to increase profits disproportionately. One study concluded that a 5% increase in customer retention rates will increase profits by 25% to 95%. Businesses tend to respond to customer attrition on a reactive basis, acting only after the customer has initiated the process to terminate service. At this stage, the chance of changing the customer's decision is almost zero. Proper application of predictive analytics can lead to a more proactive retention strategy. By a frequent examination of a customer's past service usage, service performance, spending and other behavior patterns, predictive models can determine the likelihood of a customer terminating service sometime soon. An intervention with lucrative offers can increase the chance of retaining the customer. Silent attrition, the behavior of a customer to slowly but steadily reduce usage, is another problem that many companies face. Predictive analytics can also predict this behavior, so that the company can take proper actions to increase customer activity.

Direct Marketing

When marketing consumer products and services, there is the challenge of keeping up with competing products and consumer behavior. Apart from identifying prospects, predictive analytics can also help to identify the most effective combination of product versions, marketing material, communication channels and timing that should be used to target a given consumer. The goal of predictive analytics is typically to lower the cost per order or cost per action.

Fraud Detection

Fraud is a big problem for many businesses and can be of various types: inaccurate credit applications, fraudulent transactions (both offline and online), identity thefts and false insurance claims. These problems plague firms of all sizes in many industries. Some examples of likely victims are credit card issuers, insurance companies, retail merchants, manufacturers, business-to-business suppliers and even services providers. A predictive model can help weed out the “bads” and reduce a business’s exposure to fraud.

Predictive modeling can also be used to identify high-risk fraud candidates in business or the public sector. Mark Nigrini developed a risk-scoring method to identify audit targets. He describes the use of this approach to detect fraud in the franchisee sales reports of an international fast-food chain. Each location is scored using 10 predictors. The 10 scores are then weighted to give one final overall risk score for each location. The same scoring approach was also used to identify high-risk check kiting accounts, potentially fraudulent travel agents, and questionable vendors. A reasonably complex model was used to identify fraudulent monthly reports submitted by divisional controllers.

The Internal Revenue Service (IRS) of the United States also uses predictive analytics to mine tax returns and identify tax fraud.

Recent advancements in technology have also introduced predictive behavior analysis for web fraud detection. This type of solution utilizes heuristics in order to study normal web user behavior and detect anomalies indicating fraud attempts.

Portfolio, Product or Economy-level Prediction

Often the focus of analysis is not the consumer but the product, portfolio, firm, industry or even the economy. For example, a retailer might be interested in predicting store-level demand for inventory management purposes. Or the Federal Reserve Board might be interested in predicting the unemployment rate for the next year. These types of problems can be addressed by predictive analytics using time series techniques. They can also be addressed via machine learning approaches which transform the original time series into a feature vector space, where the learning algorithm finds patterns that have predictive power.

Project Risk Management

When employing risk management techniques, the results are always to predict and benefit from a future scenario. The capital asset pricing model (CAP-M) "predicts" the best portfolio to maximize return. Probabilistic risk assessment (PRA) when combined with mini-Delphi techniques and statistical approaches yields accurate forecasts. These are examples of approaches that can extend from project to market, and from near to long term. Underwriting and other business approaches identify risk management as a predictive method.

Underwriting

Many businesses have to account for risk exposure due to their different services and determine the cost needed to cover the risk. For example, auto insurance providers need to accurately determine the amount of premium to charge to cover each automobile and driver. A financial company needs to assess a borrower's potential and ability to pay before granting a loan. For a health insurance provider, predictive analytics can analyze a few years of past medical claims data, as well as lab, pharmacy and other records where available, to predict how expensive an enrollee is likely to be in the future. Predictive analytics can help underwrite these quantities by predicting the chances of illness, default, bankruptcy, etc. Predictive analytics can streamline the process of customer acquisition by predicting the future risk behavior of a customer using application level data. Predictive analytics in the form of credit scores have reduced the amount of time it takes for loan approvals, especially in the mortgage market where lending decisions are now made in a matter of hours rather than days or even weeks. Proper predictive analytics can lead to proper pricing decisions, which can help mitigate future risk of default.

Technology and Big Data Influences

Big data is a collection of data sets that are so large and complex that they become awkward to work with using traditional database management tools. The volume, variety and velocity of big data have introduced challenges across the board for capture, storage, search, sharing, analysis, and visualization. Examples of big data sources include web logs, RFID, sensor data, social networks, Internet search indexing, call detail records, military surveillance, and complex data in astronomic, biogeochemical, genomics, and atmospheric sciences. Big Data is the core of most predictive analytic services offered by IT organizations. Thanks to technological advances in computer hardware—faster CPUs, cheaper memory, and MPP architectures—and new technologies such as Hadoop, MapReduce, and in-database and text analytics for processing big data, it is now feasible to collect, analyze, and mine massive amounts of structured and unstructured data for new insights. It is also possible to run predictive algorithms on streaming data. Today, exploring big data and using predictive analytics is within reach of more organizations than ever before and new methods that are capable for handling such datasets are proposed

Analytical Techniques

The approaches and techniques used to conduct predictive analytics can broadly be grouped into regression techniques and machine learning techniques.

Regression Techniques

Regression models are the mainstay of predictive analytics. The focus lies on establishing a mathematical equation as a model to represent the interactions between the different variables in consideration. Depending on the situation, there are a wide variety of models that can be applied while performing predictive analytics. Some of them are briefly discussed below.

Linear Regression Model

The linear regression model analyzes the relationship between the response or dependent variable and a set of independent or predictor variables. This relationship is expressed as an equation that predicts the response variable as a linear function of the parameters. These parameters are adjusted so that a measure of fit is optimized. Much of the effort in model fitting is focused on minimizing the size of the residual, as well as ensuring that it is randomly distributed with respect to the model predictions.

The goal of regression is to select the parameters of the model so as to minimize the sum of the squared residuals. This is referred to as ordinary least squares (OLS) estimation and results in best linear unbiased estimates (BLUE) of the parameters if and only if the Gauss-Markov assumptions are satisfied.

Once the model has been estimated we would be interested to know if the predictor variables belong in the model—i.e. is the estimate of each variable's contribution reliable? To do this we can check the statistical significance of the model's coefficients which can be measured using the t-statistic. This amounts to testing whether the coefficient is significantly different from zero. How well the model predicts the dependent variable based on the value of the independent variables can be assessed by using the R^2 statistic. It measures predictive power of the model i.e. the proportion of the total variation in the dependent variable that is "explained" (accounted for) by variation in the independent variables.

Discrete Choice Models

Multivariate regression (above) is generally used when the response variable is continuous and has an unbounded range. Often the response variable may not be continuous but rather discrete. While mathematically it is feasible to apply multivariate regression to discrete ordered dependent variables, some of the assumptions behind the theory of multivariate linear regression no longer hold, and there are other techniques such as discrete choice models which are better suited for this type of analysis. If the dependent variable is discrete, some of those superior methods are logistic regression, multinomial logit and probit models. Logistic regression and probit models are used when the dependent variable is binary.

Logistic Regression

In a classification setting, assigning outcome probabilities to observations can be achieved through the use of a logistic model, which is basically a method which transforms information about the binary dependent variable into an unbounded continuous variable and estimates a regular multivariate model.

The Wald and likelihood-ratio test are used to test the statistical significance of each coefficient *b* in the model. A test assessing the goodness-of-fit of a classification model is the "percentage correctly predicted".

Multinomial Logistic Regression

An extension of the binary logit model to cases where the dependent variable has more than 2 categories is the multinomial logit model. In such cases collapsing the data into two categories might not make good sense or may lead to loss in the richness of the data. The multinomial logit model is the appropriate technique in these cases, especially when the dependent variable categories are not ordered (for examples colors like red, blue, green). Some authors have extended multinomial regression to include feature selection/importance methods such as random multinomial logit.

Probit Regression

Probit models offer an alternative to logistic regression for modeling categorical dependent variables. Even though the outcomes tend to be similar, the underlying distributions are different. Probit models are popular in social sciences like economics.

A good way to understand the key difference between probit and logit models is to assume that the dependent variable is driven by a latent variable z, which is a sum of a linear combination of explanatory variables and a random noise term.

We do not observe z but instead observe y which takes the value 0 (when $z < 0$) or 1 (otherwise). In the logit model we assume that the random noise term follows a logistic distribution with mean zero. In the probit model we assume that it follows a normal distribution with mean zero. Note that in social sciences (e.g. economics), probit is often used to model situations where the observed variable y is continuous but takes values between 0 and 1.

Logit Versus Probit

The probit model has been around longer than the logit model. They behave similarly, except that the logistic distribution tends to be slightly flatter tailed. One of the reasons the logit model was formulated was that the probit model was computationally difficult due to the requirement of numerically calculating integrals. Modern computing however has made this computation fairly simple. The coefficients obtained from the logit and probit model are fairly close. However, the odds ratio is easier to interpret in the logit model.

Practical reasons for choosing the probit model over the logistic model would be:

- There is a strong belief that the underlying distribution is normal
- The actual event is not a binary outcome (*e.g.*, bankruptcy status) but a proportion (*e.g.*, proportion of population at different debt levels).

Time Series Models

Time series models are used for predicting or forecasting the future behavior of variables. These models account for the fact that data points taken over time may have an internal structure (such as autocorrelation, trend or seasonal variation) that should be accounted for. As a result, standard regression techniques cannot be applied to time series data and methodology has been developed to decompose the trend, seasonal and cyclical component of the series. Modeling the dynamic path of a variable can improve forecasts since the predictable component of the series can be projected into the future.

Time series models estimate difference equations containing stochastic components. Two commonly used forms of these models are autoregressive models (AR) and moving-average (MA) models. The Box–Jenkins methodology (1976) developed by George Box and G.M. Jenkins combines the AR and MA models to produce the ARMA (autoregressive moving average) model which is the cornerstone of stationary time series analysis. ARIMA (autoregressive integrated moving average models) on the other hand are used to describe non-stationary time series. Box and Jenkins suggest differencing a non stationary time series to obtain a stationary series to which an ARMA model can be applied. Non stationary time series have a pronounced trend and do not have a constant long-run mean or variance.

Box and Jenkins proposed a three-stage methodology which includes: model identification, estimation and validation. The identification stage involves identifying if the series is stationary or not and the presence of seasonality by examining plots of the series, autocorrelation and partial autocorrelation functions. In the estimation stage, models are estimated using non-linear time series or maximum likelihood estimation procedures. Finally the validation stage involves diagnostic checking such as plotting the residuals to detect outliers and evidence of model fit.

In recent years time series models have become more sophisticated and attempt to model conditional heteroskedasticity with models such as ARCH (autoregressive conditional heteroskedasticity) and GARCH (generalized autoregressive conditional heteroskedasticity) models frequently used for financial time series. In addition time series models are also used to understand inter-relationships among economic variables represented by systems of equations using VAR (vector autoregression) and structural VAR models.

Survival or Duration Analysis

Survival analysis is another name for time to event analysis. These techniques were primarily developed in the medical and biological sciences, but they are also widely used in the social sciences like economics, as well as in engineering (reliability and failure time analysis).

Censoring and non-normality, which are characteristic of survival data, generate difficulty when trying to analyze the data using conventional statistical models such as multiple linear regression. The normal distribution, being a symmetric distribution, takes positive as well as negative values, but duration by its very nature cannot be negative and therefore normality cannot be assumed when dealing with duration/survival data. Hence the normality assumption of regression models is violated.

The assumption is that if the data were not censored it would be representative of the population of interest. In survival analysis, censored observations arise whenever the dependent variable of interest represents the time to a terminal event, and the duration of the study is limited in time.

An important concept in survival analysis is the hazard rate, defined as the probability that the event will occur at time t conditional on surviving until time t. Another concept related to the hazard rate is the survival function which can be defined as the probability of surviving to time t.

Most models try to model the hazard rate by choosing the underlying distribution depending on the shape of the hazard function. A distribution whose hazard function slopes upward is said to have positive duration dependence, a decreasing hazard shows negative duration dependence whereas constant hazard is a process with no memory usually characterized by the exponential distribution. Some of the distributional choices in survival models are: F, gamma, Weibull, log normal, inverse normal, exponential etc. All these distributions are for a non-negative random variable.

Duration models can be parametric, non-parametric or semi-parametric. Some of the models commonly used are Kaplan-Meier and Cox proportional hazard model (non parametric).

Classification and Regression Trees (CART)

Globally-optimal classification tree analysis (GO-CTA) (also called hierarchical optimal discriminant analysis) is a generalization of optimal discriminant analysis that may be used to identify the statistical model that has maximum accuracy for predicting the value of a categorical dependent variable for a dataset consisting of categorical and continuous variables. The output of HODA is a non-orthogonal tree that combines categorical variables and cut points for continuous variables that yields maximum predictive accuracy, an assessment of the exact Type I error rate, and an evaluation of potential cross-generalizability of the statistical model. Hierarchical optimal discriminant analysis may be thought of as a generalization of Fisher's linear discriminant analysis. Optimal discriminant analysis is an alternative to ANOVA (analysis of variance) and regression analysis, which attempt to express one dependent variable as a linear combination of other features or measurements. However, ANOVA and regression analysis give a dependent variable that is a numerical variable, while hierarchical optimal discriminant analysis gives a dependent variable that is a class variable.

Classification and regression trees (CART) are a non-parametric decision tree learning technique that produces either classification or regression trees, depending on whether the dependent variable is categorical or numeric, respectively.

Decision trees are formed by a collection of rules based on variables in the modeling data set:

- Rules based on variables' values are selected to get the best split to differentiate observations based on the dependent variable
- Once a rule is selected and splits a node into two, the same process is applied to each "child" node (i.e. it is a recursive procedure)
- Splitting stops when CART detects no further gain can be made, or some pre-set stopping rules are met. (Alternatively, the data are split as much as possible and then the tree is later pruned.)

Each branch of the tree ends in a terminal node. Each observation falls into one and exactly one terminal node, and each terminal node is uniquely defined by a set of rules.

A very popular method for predictive analytics is Leo Breiman's Random forests.

Multivariate Adaptive Regression Splines

Multivariate adaptive regression splines (MARS) is a non-parametric technique that builds flexible models by fitting piecewise linear regressions.

An important concept associated with regression splines is that of a knot. Knot is where one local regression model gives way to another and thus is the point of intersection between two splines.

In multivariate and adaptive regression splines, basis functions are the tool used for generalizing the search for knots. Basis functions are a set of functions used to represent the information contained in one or more variables. Multivariate and Adaptive Regression Splines model almost always creates the basis functions in pairs.

Multivariate and adaptive regression spline approach deliberately overfits the model and then prunes to get to the optimal model. The algorithm is computationally very intensive and in practice we are required to specify an upper limit on the number of basis functions.

Machine Learning Techniques

Machine learning, a branch of artificial intelligence, was originally employed to develop techniques to enable computers to learn. Today, since it includes a number of advanced statistical methods for regression and classification, it finds application in a wide variety of fields including medical diagnostics, credit card fraud detection, face and speech recognition and analysis of the stock market. In certain applications it is sufficient to directly predict the dependent variable without focusing on the underlying relationships between variables. In other cases, the underlying relationships can be very complex and the mathematical form of the dependencies unknown. For such cases, machine learning techniques emulate human cognition and learn from training examples to predict future events.

A brief discussion of some of these methods used commonly for predictive analytics is provided below. A detailed study of machine learning can be found in Mitchell (1997).

Neural Networks

Neural networks are nonlinear sophisticated modeling techniques that are able to model complex functions. They can be applied to problems of prediction, classification or control in a wide spectrum of fields such as finance, cognitive psychology/neuroscience, medicine, engineering, and physics.

Neural networks are used when the exact nature of the relationship between inputs and output is not known. A key feature of neural networks is that they learn the relationship between inputs and output through training. There are three types of training in neural networks used by different networks, supervised and unsupervised training, reinforcement learning, with supervised being the most common one.

Some examples of neural network training techniques are backpropagation, quick propagation, conjugate gradient descent, projection operator, Delta-Bar-Delta etc. Some unsupervised network architectures are multilayer perceptrons, Kohonen networks, Hopfield networks, etc.

Multilayer Perceptron (MLP)

The multilayer perceptron (MLP) consists of an input and an output layer with one or more hidden layers of nonlinearly-activating nodes or sigmoid nodes. This is determined by the weight vector and it is necessary to adjust the weights of the network. The backpropagation employs gradient fall to minimize the squared error between the network output values and desired values for those outputs. The weights adjusted by an iterative process of repetitive present of attributes. Small changes in the weight to get the desired values are done by the process called training the net and is done by the training set (learning rule).

Radial Basis Functions

A radial basis function (RBF) is a function which has built into it a distance criterion with respect to a center. Such functions can be used very efficiently for interpolation and for smoothing of data. Radial basis functions have been applied in the area of neural networks where they are used as a replacement for the sigmoidal transfer function. Such networks have 3 layers, the input layer, the hidden layer with the RBF non-linearity and a linear output layer. The most popular choice for the non-linearity is the Gaussian. RBF networks have the advantage of not being locked into local minima as do the feed-forward networks such as the multilayer perceptron.

Support Vector Machines

support vector machines (SVM) are used to detect and exploit complex patterns in data by clustering, classifying and ranking the data. They are learning machines that are used to perform binary classifications and regression estimations. They commonly use kernel based methods to apply linear classification techniques to non-linear classification problems. There are a number of types of SVM such as linear, polynomial, sigmoid etc.

Naïve Bayes

Naïve Bayes based on Bayes conditional probability rule is used for performing classification tasks. Naïve Bayes assumes the predictors are statistically independent which makes it an effective classification tool that is easy to interpret. It is best employed when faced with the problem of 'curse of dimensionality' i.e. when the number of predictors is very high.

k-nearest Neighbours

The nearest neighbour algorithm (KNN) belongs to the class of pattern recognition statistical methods. The method does not impose a priori any assumptions about the distribution from which the modeling sample is drawn. It involves a training set with both positive and negative values. A new sample is classified by calculating the distance to the nearest neighbouring training case. The sign of that point will determine the classification of the sample. In the k-nearest neighbour classifier, the k nearest points are considered and the sign of the majority is used to classify the sample. The performance of the kNN algorithm is influenced by three main factors: (1) the distance measure used to locate the nearest neighbours; (2) the decision rule used to derive a classification from the k-nearest neighbours; and (3) the number of neighbours used to classify the new sample. It can be proved that, unlike other methods, this method is universally asymptotically convergent, i.e.: as the size of the training set increases, if the observations are independent and identically distributed (i.i.d.), regardless of the distribution from which the sample is drawn, the predicted class will converge to the class assignment that minimizes misclassification error.

Geospatial Predictive Modeling

Conceptually, geospatial predictive modeling is rooted in the principle that the occurrences of events being modeled are limited in distribution. Occurrences of events are neither uniform nor random in distribution—there are spatial environment factors (infrastructure, sociocultural, topographic, etc.) that constrain and influence where the locations of events occur. Geospatial predictive modeling attempts to describe those constraints and influences by spatially correlating occurrences of historical geospatial locations with environmental factors that represent those constraints and influences. Geospatial predictive modeling is a process for analyzing events through a geographic filter in order to make statements of likelihood for event occurrence or emergence.

Tools

Historically, using predictive analytics tools—as well as understanding the results they delivered—required advanced skills. However, modern predictive analytics tools are no longer restricted to IT specialists. As more organizations adopt predictive analytics into decision-making processes and integrate it into their operations, they are creating a shift in the market toward business users as the primary consumers of the information. Business users want tools they can use on their own. Vendors are responding by creating new software that removes the mathematical complexity, provides user-friendly graphic interfaces and/or builds in short cuts that can, for example, recognize the kind of data available and suggest an appropriate predictive model. Predictive analytics tools have become sophisticated enough to adequately present and dissect data problems, so that any data-savvy information worker can utilize them to analyze data and retrieve meaningful, useful results. For example, modern tools present findings using simple charts, graphs, and scores that indicate the likelihood of possible outcomes.

There are numerous tools available in the marketplace that help with the execution of predictive analytics. These range from those that need very little user sophistication to those that are designed for the expert practitioner. The difference between these tools is often in the level of customization and heavy data lifting allowed.

Notable open source predictive analytic tools include:

- Apache Mahout
- GNU Octave
- KNIME
- OpenNN
- Orange
- R
- scikit-learn

- Weka

Notable commercial predictive analytic tools include:

- Alpine Data Labs
- Alteryx
- Angoss KnowledgeSTUDIO
- BIRT Analytics
- IBM SPSS Statistics and IBM SPSS Modeler
- KXEN Modeler
- Mathematica
- MATLAB
- Minitab
- LabVIEW
- Neural Designer
- Oracle Advanced Analytics
- Pervasive
- Predixion Software
- RapidMiner
- RCASE
- Revolution Analytics
- SAP HANA and SAP BusinessObjects Predictive Analytics
- SAS and SAS Enterprise Miner
- STATA
- Statgraphics
- STATISTICA
- TeleRetail
- TIBCO

Beside these software packages, specific tools have also been developed for industrial applications. For example, Watchdog Agent Toolbox has been developed and optimized for predictive analysis in prognostics and health management applications and is available for MATLAB and LabVIEW.

The most popular commercial predictive analytics software packages according to the Rexer Analytics Survey for 2013 are IBM SPSS Modeler, SAS Enterprise Miner, and Dell Statistica.

PMML

In an attempt to provide a standard language for expressing predictive models, the Predictive Model Markup Language (PMML) has been proposed. Such an XML-based language provides a way for the different tools to define predictive models and to share these between PMML compliant applications. PMML 4.0 was released in June, 2009.

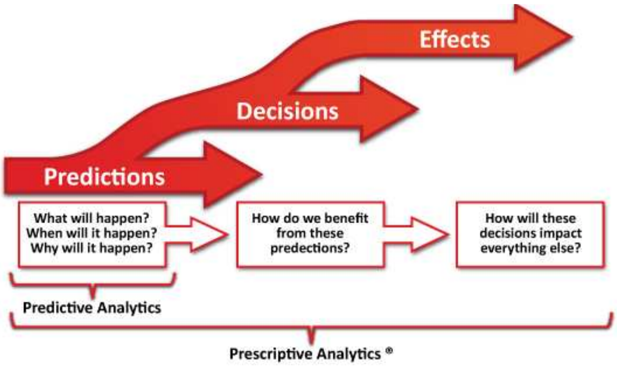
Criticism

There are plenty of skeptics when it comes to computers and algorithms abilities to predict the future, including Gary King, a professor from Harvard University and the director of the Institute for Quantitative Social Science. People are influenced by their environment in innumerable ways. Trying to understand what people will do next assumes that all the influential variables can be known and measured accurately. “People’s environments change even more quickly than they themselves do. Everything from the weather to their relationship with their mother can change the way people think and act. All of those variables are unpredictable. How they will impact a person is even less predictable. If put in the exact same situation tomorrow, they may make a completely different decision. This means that a statistical prediction is only valid in sterile laboratory conditions, which suddenly isn’t as useful as it seemed before.”

Prescriptive Analytics

Prescriptive analytics is the third and final phase of analytics (BA) which also includes descriptive and predictive analytics.

Referred to as the “final frontier of analytic capabilities,” prescriptive analytics entails the application of mathematical and computational sciences suggests decision options to take advantage of the results of descriptive and predictive analytics. The first stage of business analytics is descriptive analytics, which still accounts for the majority of all business analytics today. Descriptive analytics looks at past performance and understands that performance by mining historical data to look for the reasons behind past success or failure. Most management reporting - such as sales, marketing, operations, and finance - uses this type of post-mortem analysis.



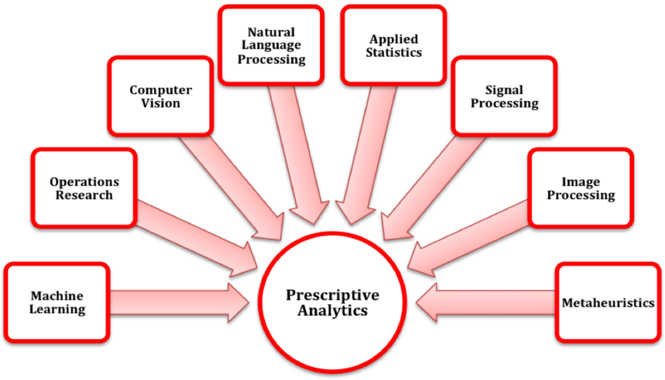
Prescriptive Analytics extends beyond predictive analytics by specifying both the actions necessary to achieve predicted outcomes, and the interrelated effects of each decision

The next phase is predictive analytics. Predictive analytics answers the question what is likely to happen. This is when historical data is combined with rules, algorithms, and occasionally external data to determine the probable future outcome of an event or the likelihood of a situation occurring. The final phase is prescriptive analytics, which goes beyond predicting future outcomes by also suggesting actions to benefit from the predictions and showing the implications of each decision option.

Prescriptive analytics not only anticipates what will happen and when it will happen, but also why it will happen. Further, prescriptive analytics suggests decision options on how to take advantage of a future opportunity or mitigate a future risk and shows the implication of each decision option. Prescriptive analytics can continually take in new data to re-predict and re-prescribe, thus automatically improving prediction accuracy and prescribing better decision options. Prescriptive analytics ingests hybrid data, a combination of structured (numbers, categories) and unstructured data (videos, images, sounds, texts), and business rules to predict what lies ahead and to prescribe how to take advantage of this predicted future without compromising other priorities.

All three phases of analytics can be performed through professional services or technology or a combination. In order to scale, prescriptive analytics technologies need to be adaptive to take into account the growing volume, velocity, and variety of data that most mission critical processes and their environments may produce.

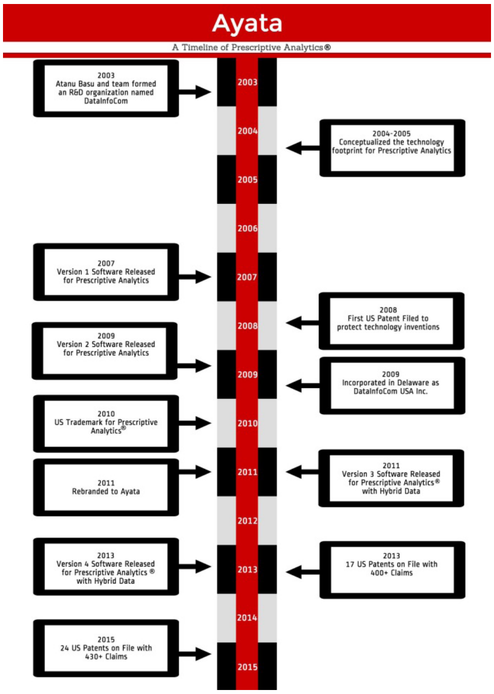
One criticism of prescriptive analytics is that its distinction from predictive analytics is ill-defined and therefore ill-conceived.



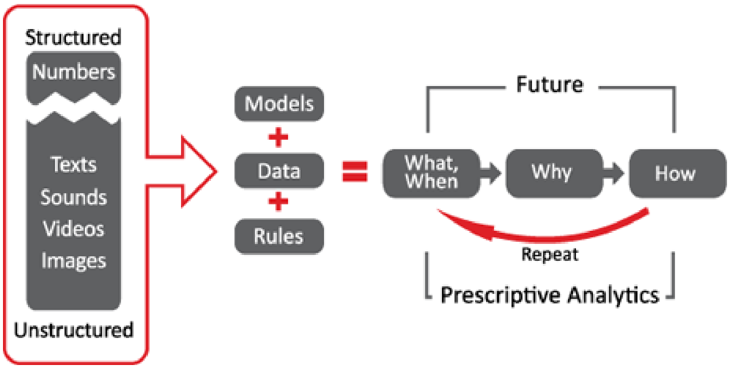
The scientific disciplines that comprise Prescriptive Analytics

History

While the term Prescriptive Analytics, first coined by IBM and later trademarked by Ayata, the underlying concepts have been around for hundreds of years. The technology behind prescriptive analytics synergistically combines hybrid data, business rules with mathematical models and computational models. The data inputs to prescriptive analytics may come from multiple sources: internal, such as inside a corporation; and external, also known as environmental data. The data may be structured, which includes numbers and categories, as well as unstructured data, such as texts, images, sounds, and videos. Unstructured data differs from structured data in that its format varies widely and cannot be stored in traditional relational databases without significant effort at data transformation. More than 80% of the world's data today is unstructured, according to IBM.


Ayata
A Timeline of Prescriptive Analytics®
2003 Atanu Basu and team formed an R&D organization named DataInfoCom
2004-2005 Conceptualized the technology footprint for Prescriptive Analytics
2007 Version 1 Software Released for Prescriptive Analytics
2008 First US Patent Filed to protect technology inventions
2009 Version 2 Software Released for Prescriptive Analytics
2009 Incorporated in Delaware as DataInfoCom USA Inc.
2010 US Trademark for Prescriptive Analytics®
2011 Rebranded to Ayata
2011 Version 3 Software Released for Prescriptive Analytics® with Hybrid Data
2013 Version 4 Software Released for Prescriptive Analytics® with Hybrid Data
2013 17 US Patents on File with 400+ Claims
2015 24 US Patents on File with 430+ Claims
2003
2004
2005
2006
2007
2008
2009
2010
2011
2012
2013
2014
2015

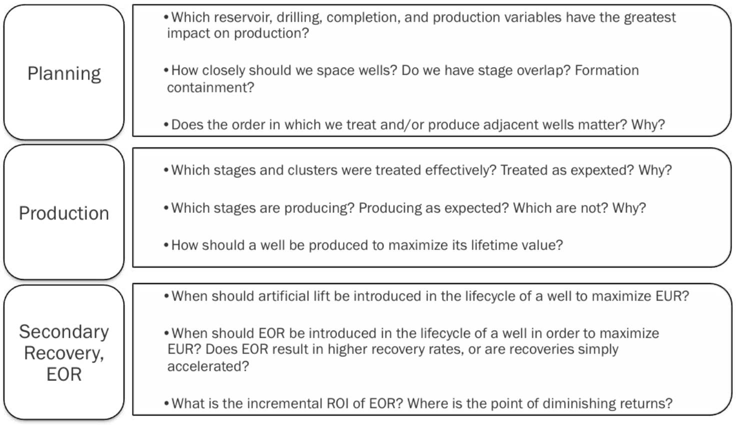
Timeline tracing evolution of Prescriptive Analytics capability and software


Structured
Numbers
Texts
Sounds
Videos
Images
Unstructured
Models
+
Data
+
Rules
=
Future
What, When
Why
How
Repeat
Prescriptive Analytics

Prescriptive Analytics incorporates both structured and unstructured data, and uses a combination of advanced analytic techniques and disciplines to predict, prescribe, and adapt.

In addition to this variety of data types and growing data volume, incoming data can also evolve with respect to velocity, that is, more data being generated at a faster or a variable pace. Business rules define the business process and include objectives constraints, preferences, policies, best practices, and boundaries. Mathematical models and computational models are techniques derived from mathematical sciences, computer science and related disciplines such as applied statistics, machine learning, operations research, natural language processing, computer vision, pattern recognition, image processing, speech recognition, and signal processing. The correct application of all these methods and the verification of their results implies the need for resources on a massive scale including human, computational and temporal for every Prescriptive Analytic project. In order to spare the expense of dozens of people, high performance machines and weeks of work one must consider the reduction of resources and therefore a reduction in the accuracy or reliability of the outcome. The preferable route is a reduction that produces a probabilistic result within acceptable limits.

Applications in Oil and Gas


Planning
Production
Secondary Recovery, EOR

Planning
- Which reservoir, drilling, completion, and production variables have the greatest impact on production?
- How closely should we space wells? Do we have stage overlap? Formation containment?
- Does the order in which we treat and/or produce adjacent wells matter? Why?

Production
- Which stages and clusters were treated effectively? Treated as expected? Why?
- Which stages are producing? Producing as expected? Which are not? Why?
- How should a well be produced to maximize its lifetime value?

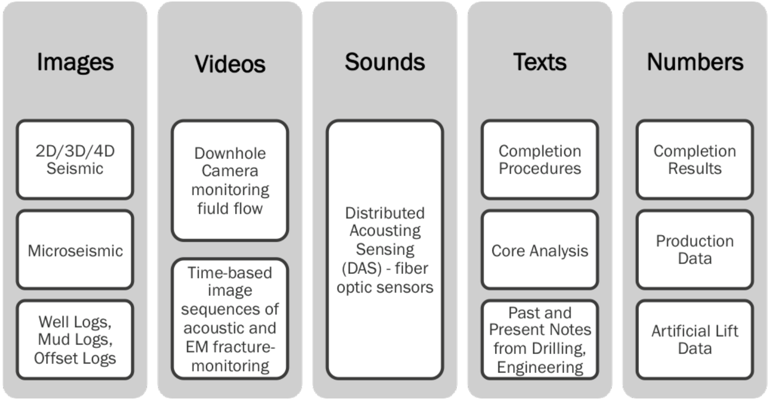
Secondary Recovery, EOR
- When should artificial lift be introduced in the lifecycle of a well to maximize EUR?
- When should EOR be introduced in the lifecycle of a well in order to maximize EUR? Does EOR result in higher recovery rates, or are recoveries simply accelerated?
- What is the incremental ROI of EOR? Where is the point of diminishing returns?

Key Questions Prescriptive Analytics software answers for oil and gas producers

Energy is the largest industry in the world ($6 trillion in size). The processes and decisions related to oil and natural gas exploration, development and production generate large amounts of data. Many types of captured data are used to create models and images of the Earth's structure and layers 5,000 - 35,000 feet below the surface and to describe activities around the wells themselves, such as depositional characteristics, machinery performance, oil flow rates, reservoir temperatures and pressures. Prescriptive analytics software can help with both locating and producing hydrocarbons

by taking in seismic data, well log data, production data, and other related data sets to prescribe specific recipes for how and where to drill, complete, and produce wells in order to optimize recovery, minimize cost, and reduce environmental footprint.

Unconventional Resource Development


Images
2D/3D/4D Seismic
Microseismic
Well Logs, Mud Logs, Offset Logs
Videos
Downhole Camera monitoring fiuld flow
Time-based image sequences of acoustic and EM fracture-monitoring
Sounds
Distributed Acousting Sensing (DAS) - fiber optic sensors
Texts
Completion Procedures
Core Analysis
Past and Present Notes from Drilling, Engineering
Numbers
Completion Results
Production Data
Artificial Lift Data

Examples of structured and unstructured data sets generated and by the oil and gas companies and their ecosystem of service providers that can be analyzed together using Prescriptive Analytics software

With the value of the end product determined by global commodity economics, the basis of competition for operators in upstream E&P is the ability to effectively deploy capital to locate and extract resources more efficiently, effectively, predictably, and safely than their peers. In unconventional resource plays, operational efficiency and effectiveness is diminished by reservoir inconsistencies, and decision-making impaired by high degrees of uncertainty. These challenges manifest themselves in the form of low recovery factors and wide performance variations.

Prescriptive Analytics software can accurately predict production and prescribe optimal configurations of controllable drilling, completion, and production variables by modeling numerous internal and external variables simultaneously, regardless of source, structure, size, or format. Prescriptive analytics software can also provide decision options and show the impact of each decision option so the operations managers can proactively take appropriate actions, on time, to guarantee future exploration and production performance, and maximize the economic value of assets at every point over the course of their serviceable lifetimes.

Oilfield Equipment Maintenance

In the realm of oilfield equipment maintenance, Prescriptive Analytics can optimize configuration, anticipate and prevent unplanned downtime, optimize field scheduling, and improve maintenance planning. According to General Electric, there are more than 130,000 electric submersible pumps (ESP's) installed globally, accounting for 60% of the world's oil production. Prescriptive Analytics has been deployed to predict when and why an ESP will fail, and recommend the necessary actions to prevent the failure.

In the area of Health, Safety, and Environment, prescriptive analytics can predict and preempt incidents that can lead to reputational and financial loss for oil and gas companies.

Pricing

Pricing is another area of focus. Natural gas prices fluctuate dramatically depending upon supply, demand, econometrics, geopolitics, and weather conditions. Gas producers, pipeline transmission companies and utility firms have a keen interest in more accurately predicting gas prices so that they can lock in favorable terms while hedging downside risk. Prescriptive analytics software can accurately predict prices by modeling internal and external variables simultaneously and also provide decision options and show the impact of each decision option.

Applications in Healthcare

Multiple factors are driving healthcare providers to dramatically improve business processes and operations as the United States healthcare industry embarks on the necessary migration from a largely fee-for service, volume-based system to a fee-for-performance, value-based system. Prescriptive analytics is playing a key role to help improve the performance in a number of areas involving various stakeholders: payers, providers and pharmaceutical companies.

Prescriptive analytics can help providers improve effectiveness of their clinical care delivery to the population they manage and in the process achieve better patient satisfaction and retention. Providers can do better population health management by identifying appropriate intervention models for risk stratified population combining data from the in-facility care episodes and home based telehealth.

Prescriptive analytics can also benefit healthcare providers in their capacity planning by using analytics to leverage operational and usage data combined with data of external factors such as economic data, population demographic trends and population health trends, to more accurately plan for future capital investments such as new facilities and equipment utilization as well as understand the trade-offs between adding additional beds and expanding an existing facility versus building a new one.

Prescriptive analytics can help pharmaceutical companies to expedite their drug development by identifying patient cohorts that are most suitable for the clinical trials worldwide - patients who are expected to be compliant and will not drop out of the trial due to complications. Analytics can tell companies how much time and money they can save if they choose one patient cohort in a specific country vs. another.

In provider-payer negotiations, providers can improve their negotiating position with health insurers by developing a robust understanding of future service utilization. By accurately predicting utilization, providers can also better allocate personnel.

Social Media Analytics

Social Media Analytics as a part of social analytics is the process of gathering data from stakeholder conversations on digital media and processing into structured insights leading to more information-driven business decisions and increased customer centrality for brands and businesses.

Social media analytics can also be referred as social media listening, social media monitoring or social media intelligence.

Digital media sources for social media analytics include social media channels, blogs, forums, image sharing sites, video sharing sites, aggregators, classifieds, complaints, Q&A, reviews, Wikipedia and others.

Social media analytics is an industry agnostic practice and is commonly used in different approaches on business decisions, marketing, customer service, reputation management, sales and others. There is an array of tools that offers the social media analysis, varying from the level of business requirement. Logic behind algorithms that are designed for these tools is selection, data pre-processing, transformation, mining and hidden pattern evaluation.

In order to make the complete process of social media analysis a success it is important that key performance indicators (KPIs) for objectively evaluating the data is defined.

Social media analytics is important when one needs to understand the patterns that are hidden in large amount of social data related to particular brands.

Homophily is used as a apart of analytics, it is a tendency that a contact between similar people occurs at a higher rate than among dissimilar people. According to research, two users who follow reciprocally share topical interests by mining their thousands of links. All these are used for taking major business decision in social media sectors.

The success of social media monitoring (SMM) tools may vary from one company to another. According to Soleman and Cohard (2016), beyond technical factors related to social media monitoring (SMM) (quality of sources, functionalities, quality of the tool), organizations must also take into account the need for new capabilities, human, managerial and organizational skills to take advantage of their SMM tools.

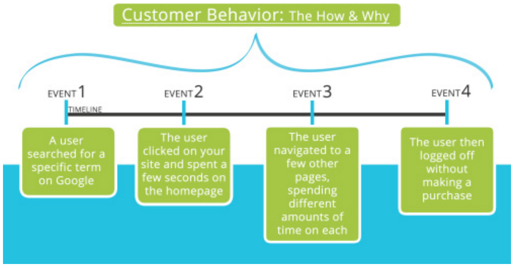
Behavioral Analytics

Behavioral analytics is a recent advancement in business analytics that reveals new insights into the behavior of consumers on eCommerce platforms, online games, web and mobile applications, and IoT. The rapid increase in the volume of raw event data generated by the digital world enables methods that go beyond typical analysis by demographics and other traditional metrics that tell us what kind of people took what actions in the past. Behavioral analysis focuses on understanding how consumers act and why, enabling accurate predictions about how they are likely to act in the future. It enables marketers to make the right offers to the right consumer segments at the right time.

Behavioral analytics utilizes the massive volumes of raw user event data captured during sessions in which consumers use application, game, or website, including traffic data like navigation path, clicks, social media interactions, purchasing decisions and marketing responsiveness. Also, the event-data can include advertising metrics like click-to-conversion time, as well as comparisons between other metrics like the monetary value of an order and the amount of time spent on the site. These data points are then compiled and analyzed, whether by looking at session progression from when a user first entered the platform until a sale was made, or what other products a user bought or looked at before this purchase. Behavioral analysis allows future actions and trends to be predicted based on the collection of such data.

While business analytics has a more broad focus on the who, what, where and when of business intelligence, behavioral analytics narrows that scope, allowing one to take seemingly unrelated data points in order to extrapolate, predict and determine errors and future trends. It takes a more holistic and human view of data, connecting individual data points to tell us not only what is happening, but also how and why it is happening.

Examples and Real World Applications


Customer Behavior: The How & Why
EVENT 1
EVENT 2
EVENT 3
EVENT 4
TIMELINE
A user searched for a specific term on Google
The user clicked on your site and spent a few seconds on the homepage
The user navigated to a few other pages, spending different amounts of time on each
The user then logged off without making a purchase

Visual Representation of Events that Make Up Behavioral Analysis

Data shows that a large percentage of users using a certain eCommerce platform found it by searching for “Thai food” on Google. After landing on the homepage, most people spent some time on the “Asian Food” page and then logged off without placing an order. Looking at each of these events as separate data points does not represent what is really going on and why people did not make a purchase. However, viewing these data points as a representation of overall user behavior enables one to interpolate how and why users acted in this particular case.

Behavioral analytics looks at all site traffic and page views as a timeline of connected events that did not lead to orders. Since most users left after viewing the "Asian Food" page, there could be a disconnect between what they are searching for on Google and what the "Asian Food" page displays. Knowing this, a quick look at the "Asian Food" page reveals that it does not display Thai food prominently and thus people do not think it is actually offered, even though it is.

Behavioral analytics is becoming increasingly popular in commercial environments. Amazon.com is a leader in using behavioral analytics to recommend additional products that customers are likely to buy based on their previous purchasing patterns on the site. Behavioral analytics is also used by Target to suggest products to customers in their retail stores, while political campaigns use it to determine how potential voters should be approached. In addition to retail and political applications, behavioral analytics is also used by banks and manufacturing firms to prioritize leads generated by their websites. Behavioral analytics also allow developers to manage users in online-gaming and web applications.

Types

- Ecommerce and retail – Product recommendations and predicting future sales trends
- Online gaming – Predicting usage trends, load, and user preferences in future releases
- Application development – Determining how users use an application to predict future usage and preferences.
- Cohort analysis – Breaking users down into similar groups to gain a more focused understanding of their behavior.
- Security – Detecting compromised credentials and insider threats by locating anomalous behavior.
- Suggestions – People who liked this also liked...
- Presentation of relevant content based on user behavior.

Components of Behavioral Analytics

An ideal behavioral analytics solution would include:

- Real-time capture of vast volumes of raw event data across all relevant digital devices and applications used during sessions
- Automatic aggregation of raw event data into relevant data sets for rapid access, filtering and analysis
- Ability to query data in an unlimited number of ways, enabling users to ask any business question
- Extensive library of built-in analysis functions such as cohort, path and funnel analysis

- A visualization component

Subsets of Behavioral Analytics

Path Analysis (Computing)

Path analysis, is the analysis of a path, which is a portrayal of a chain of consecutive events that a given user or cohort performs during a set period of time while using a website, online game, or eCommerce platform. As a subset of behavioral analytics, path analysis is a way to understand user behavior in order to gain actionable insights into the data. Path analysis provides a visual portrayal of every event a user or cohort performs as part of a path during a set period of time.

While it is possible to track a user's path through the site, and even show that path as a visual representation, the real question is how to gain these actionable insights. If path analysis simply outputs a "pretty" graph, while it may look nice, it does not provide anything concrete to act upon.

Examples

In order to get the most out of path analysis the first step would be to determine what needs to be analyzed and what are the goals of the analysis. A company might be trying to figure out why their site is running slow, are certain types of users interested in certain pages or products, or if their user interface is set up in a logical way.

Now that the goal has been set there are a few ways of performing the analysis. If a large percentage of a certain cohort, people between the ages of 18-25, logs into an online game, creates a profile and then spends the next 10 minutes wandering around the menu page, then it may be that the user interface is not logical. By seeing this group of users following the path that they did a developer will be able to analyze the data and realize that after creating a profile, the "play game" button does not appear. Thus, path analysis was able to provide actionable data for the company to act on and fix an error.

In eCommerce, path analysis can help customize a shopping experience to each user. By looking at what products other customers in a certain cohort looked at before buying one, a company can suggest "items you may also like" to the next customer and increase the chances of them making a purchase. Also, path analysis can help solve performance issues on a platform. For example, a company looks at a path and realizes that their site freezes up after a certain combinations of events. By analyzing the path and the progression of events that led to the error, the company can pinpoint the error and fix it.

Evolution

Historically path analysis fell under the broad category of website analytics, and related only to the analysis of paths through websites. Path analysis in website analytics is a process of determining a sequence of pages visited in a visitor session prior to some desired event, such as the visitor purchasing an item or requesting a newsletter. The precise order of pages visited may or may not be important and may or may not be specified. In practice, this analysis is done in aggregate, ranking the paths (sequences of pages) visited prior to the desired event, by descending frequency of use. The idea is to determine what features of the website encourage the desired result. "Fallout analysis,” a subset of path analysis, looks at “black holes” on the site, or paths that lead to a dead end most frequently, paths or features that confuse or lose potential customers.

With the advent of big data along with web based applications, online games, and eCommerce platforms, path analysis has come to include much more than just web path analysis. Understanding how users move through an app, game, or other web platform are all part of modern-day path analysis.

Understanding Visitors

In the real world when you visit a shop the shelves and products are not placed in a random order. The shop owner carefully analyzes the visitors and path they walk through the shop, especially when they are selecting or buying products. Next the shop owner will reorder the shelves and products to optimize sales by putting everything in the most logical order for the visitors. In a supermarket this will typically result in the wine shelf next to a variety of cookies, chips, nuts, etc. Simply because people drink wine and eat nuts with it.

In most web sites there is a same logic that can be applied. Visitors who have questions about a product will go to the product information or support section of a web site. From there they make a logical step to the frequently asked questions page if they have a specific question. A web site owner also wants to analyze visitor behavior. For example, if a web site offers products for sale, the owner wants to convert as many visitors to a completed purchase. If there is a sign-up form with multiple pages, web site owners want to guide visitors to the final sign-up page.

Path analysis answers typical questions like:
Where do most visitors go after they enter my home page?
Is there a strong visitor relation between product A and product B on my web site?.
Questions that can’t be answered by page hits and unique visitors statistics.

Funnels and Goals

Google Analytics provides a path function with funnels and goals. A predetermined path of web site pages is specified and every visitor walking the path is a goal. This approach is very helpful when analyzing how many visitors reach a certain destination page, called an end point analysis.

Using Maps

The paths visitors walk in a web site can lead to an endless number of unique paths. As a result, there is no point in analyzing each path, but to look for the strongest paths. These strongest paths are typically shown in a graphical map or in text like: Page A --> Page B --> Page D --> Exit.

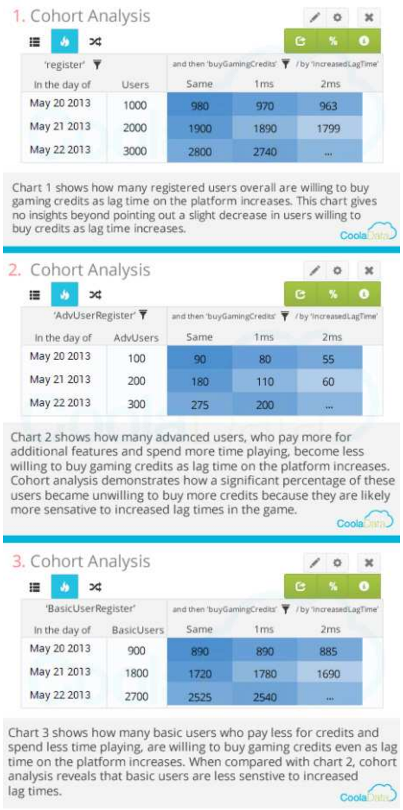
Cohort Analysis

Cohort analysis is a subset of behavioral analytics that takes the data from a given dataset (e.g. an eCommerce platform, web application, or online game) and rather than looking at all users as one unit, it breaks them into related groups for analysis. These related groups, or cohorts, usually share common characteristics or experiences within a defined time-span. Cohort analysis allows a company to “see patterns clearly across the life-cycle of a customer (or user), rather than slicing across all customers blindly without accounting for the natural cycle that a customer undergoes.” By seeing these patterns of time, a company can adapt and tailor its service to those specific cohorts. While cohort analysis is sometimes associated with a cohort study, they are different and should not be viewed as one and the same. Cohort analysis has come to describe specifically the analysis of cohorts in regards to big data and business analytics, while a cohort study is a more general umbrella term that describes a type of study in which data is broken down into similar groups.

Examples

The goal of a business analytic tool is to analyze and present actionable information. In order for a company to act on such info it must be relevant to the situation at hand. A database full of thousands or even millions of entries of all user data makes it tough to gain actionable data, as those data span many different categories and time periods. Actionable cohort analysis allows for the ability to drill down to the users of each specific cohort to gain a better understanding of their behaviors, such as if users checked out, and how much did they pay. In cohort analysis “each new group [cohort] provides the opportunity to start with a fresh set of users,” allowing the company to look at only the data that is relevant to the current query and act on it.

In e-Commerce, a firm may only be interested in customers who signed up in the last two weeks and who made a purchase, which is an example of a specific cohort. A software developer may only care about the data from users who sign up after a certain upgrade, or who use certain features of the platform.



1. Cohort Analysis

'register'		and then 'buyGamingCredits' / by 'increasedLagTime'		
In the day of	Users	Same	1ms	2ms
May 20 2013	1000	980	970	963
May 21 2013	2000	1900	1890	1799
May 22 2013	3000	2800	2740	...

Chart 1 shows how many registered users overall are willing to buy gaming credits as lag time on the platform increases. This chart gives no insights beyond pointing out a slight decrease in users willing to buy credits as lag time increases.

2. Cohort Analysis

'AdvUserRegister'		and then 'buyGamingCredits' / by 'increasedLagTime'		
In the day of	AdvUsers	Same	1ms	2ms
May 20 2013	100	90	80	55
May 21 2013	200	180	110	60
May 22 2013	300	275	200	...

Chart 2 shows how many advanced users, who pay more for additional features and spend more time playing, become less willing to buy gaming credits as lag time on the platform increases. Cohort analysis demonstrates how a significant percentage of these users became unwilling to buy more credits because they are likely more sensative to increased lag times in the game.

3. Cohort Analysis

'BasicUserRegister'		and then 'buyGamingCredits' / by 'increasedLagTime'		
In the day of	BasicUsers	Same	1ms	2ms
May 20 2013	900	890	890	885
May 21 2013	1800	1720	1780	1690
May 22 2013	2700	2525	2540	...

Chart 3 shows how many basic users who pay less for credits and spend less time playing, are willing to buy gaming credits even as lag time on the platform increases. When compared with chart 2, cohort analysis reveals that basic users are less senstive to increased lag times.

An example of cohort analysis of gamers on a certain platform: Expert gamers, cohort 1, will care more about advanced features and lag time compared to new sign-ups, cohort 2. With these two cohorts determined, and the analysis run, the gaming company would be presented with a visual representation of the data specific to the two cohorts. It could then see that a slight lag in load times has been translating into a significant loss of revenue from advanced gamers, while new sign-ups have not even noticed the lag. Had the company simply looked at its overall revenue reports for all customers, it would not have been able to see the differences between these two cohorts. Cohort analysis allows a company to pick up on patterns and trends and make the changes necessary to keep both advanced and new gamers happy.

Deep Actionable Cohort Analytics

"An actionable metric is one that ties specific and repeatable actions to observed results [like user registration, or checkout]. The opposite of actionable metrics are vanity metrics (like web hits or number of downloads) which only serve to document the current state of the product but offer no insight into how we got here or what to do next." Without actionable analytics the information that is being presented may not have any practical application, as the only data points represent vanity metrics that do not translate into any specific outcome. While it is useful for a company to know how many people are on their site, that metric is useless on its own. For it to be actionable it needs to relate a "repeatable action to [an] observed result".

Performing Cohort Analysis

In order to perform a proper cohort analysis, there are four main stages:

- Determine what question you want to answer. The point of the analysis is to come up with actionable information on which to act in order to improve business, product, user experience, turnover, etc. To ensure that happens, it is important that the right question is asked. In the gaming example above, the company was unsure why they were losing revenue as lag time increased, despite the fact that users were still signing up and playing games.
- Define the metrics that will be able to help you answer the question. A proper cohort analysis requires the identification of an event, such as a user checking out, and specific properties, like how much the user paid. The gaming example measured a customer's willingness to buy gaming credits based on how much lag time there was on the site.
- Define the specific cohorts that are relevant. In creating a cohort, one must either analyze all the users and target them or perform attribute contribution in order to find the relevant differences between each of them, ultimately to discover and explain their behavior as a specific cohort. The above example splits users into "basic" and "advanced" users as each group differs in actions, pricing structure sensitivities, and usage levels.
- Perform the cohort analysis. The analysis above was done using data visualization which allowed the gaming company to realize that their revenues were falling because their higher-paying advanced users were not using the system as the lag time increased. Since the advanced users were such a large portion of the company's revenue, the additional basic user signups were not covering the financial losses from losing the advanced users. In order to fix this, the company improved their lag times and began catering more to their advanced users.

References

- Coker, Frank (2014). Pulse: Understanding the Vital Signs of Your Business (1st ed.). Bellevue, WA: Ambient Light Publishing. pp. 30, 39, 42,more. ISBN 978-0-9893086-0-1.
- Finlay, Steven (2014). Predictive Analytics, Data Mining and Big Data. Myths, Misconceptions and Methods (1st ed.). Basingstoke: Palgrave Macmillan. p. 237. ISBN 1137379278.
- Siegel, Eric (2013). Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die (1st ed.). Wiley. ISBN 978-1-1183-5685-2.
- Nigrini, Mark (June 2011). “Forensic Analytics: Methods and Techniques for Forensic Accounting Investigations”. Hoboken, NJ: John Wiley & Sons Inc. ISBN 978-0-470-89046-2.
- Lindert, Bryan (October 2014). “Eckerd Rapid Safety Feedback Bringing Business Intelligence to Child Welfare” (PDF). Policy & Practice. Retrieved March 3, 2016.
- “New Strategies Long Overdue on Measuring Child Welfare Risk - The Chronicle of Social Change”. The Chronicle of Social Change. Retrieved 2016-04-04.
- “Eckerd Rapid Safety Feedback® Highlighted in National Report of Commission to Eliminate Child Abuse and Neglect Fatalities”. Eckerd Kids. Retrieved 2016-04-04.
- “A National Strategy to Eliminate Child Abuse and Neglect Fatalities” (PDF). Commission to Eliminate Child Abuse and Neglect Fatalities. (2016). Retrieved April 4, 2016.
- “Predictive Big Data Analytics: A Study of Parkinson’s Disease using Large, Complex, Heterogeneous, Incongruent, Multi-source and Incomplete Observations”. PLoS ONE. Retrieved 2016-08-10.
- “2014 Embedded Business Intelligence Market Study Now Available From Dresner Advisory Services”. Market Wired. Retrieved August 2015.

3

Data Mining: An Overview

The process of understanding the patterns found in large data sets is known as data mining. Some of the aspects of data mining that have been elucidated in the following section are association rule learning, cluster analysis, regression analysis, automatic summarization and examples of data mining. The chapter on data mining offers an insightful focus, keeping in mind the complex subject matter.

Data Mining

Data mining is an interdisciplinary subfield of computer science. It is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and online updating. Data mining is the analysis step of the "knowledge discovery in databases" process, or KDD.

The term is a misnomer, because the goal is the extraction of patterns and knowledge from large amounts of data, not the extraction (*mining*) of data itself. It also is a buzzword and is frequently applied to any form of large-scale data or information processing (collection, extraction, warehousing, analysis, and statistics) as well as any application of computer decision support system, including artificial intelligence, machine learning, and business intelligence. The book *Data mining: Practical machine learning tools and techniques with Java* (which covers mostly machine learning material) was originally to be named just *Practical machine learning*, and the term *data mining* was only added for marketing reasons. Often the more general terms (*large scale*) *data analysis* and *analytics* – or, when referring to actual methods, *artificial intelligence* and *machine learning* – are more appropriate.

The actual data mining task is the automatic or semi-automatic analysis of large quantities of data to extract previously unknown, interesting patterns such as groups of data records (cluster analysis), unusual records (anomaly detection), and dependencies (association rule mining). This usually involves using database techniques such as spatial indices. These patterns can then be seen as a kind of summary of the input data, and may be used in further analysis or, for example, in machine learning and predictive analytics. For example, the data mining step might identify multiple groups in the data, which can then be used to obtain more accurate prediction results by a decision support system. Neither the data collection, data preparation, nor result interpretation and reporting is part of the data mining step, but do belong to the overall KDD process as additional steps.

The related terms *data dredging*, *data fishing*, and *data snooping* refer to the use of data mining methods to sample parts of a larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These methods can, however, be used in creating new hypotheses to test against the larger data populations.

Etymology

In the 1960s, statisticians used terms like “Data Fishing” or “Data Dredging” to refer to what they considered the bad practice of analyzing data without an a-priori hypothesis. The term “Data Mining” appeared around 1990 in the database community. For a short time in 1980s, a phrase “database mining”, was used, but since it was trademarked by HNC, a San Diego-based company, to pitch their Database Mining Workstation; researchers consequently turned to “data mining”. Other terms used include Data Archaeology, Information Harvesting, Information Discovery, Knowledge Extraction, etc. Gregory Piatetsky-Shapiro coined the term “Knowledge Discovery in Databases” for the first workshop on the same topic (KDD-1989) and this term became more popular in AI and Machine Learning Community. However, the term data mining became more popular in the business and press communities. Currently, Data Mining and Knowledge Discovery are used interchangeably. Since about 2007, “Predictive Analytics” and since 2011, “Data Science” terms were also used to describe this field.

In the Academic community, the major forums for research started in 1995 when the First International Conference on Data Mining and Knowledge Discovery (KDD-95) was started in Montreal under AAAI sponsorship. It was co-chaired by Usama Fayyad and Ramasamy Uthurusamy. A year later, in 1996, Usama Fayyad launched the journal by Kluwer called Data Mining and Knowledge Discovery as its founding Editor-in-Chief. Later he started the SIGKDDD Newsletter SIGKDD Explorations. The KDD International conference became the primary highest quality conference in Data Mining with an acceptance rate of research paper submissions below 18%. The Journal Data Mining and Knowledge Discovery is the primary research journal of the field.

Background

The manual extraction of patterns from data has occurred for centuries. Early methods of identifying patterns in data include Bayes’ theorem (1700s) and regression analysis (1800s). The proliferation, ubiquity and increasing power of computer technology has dramatically increased data collection, storage, and manipulation ability. As data sets have grown in size and complexity, direct “hands-on” data analysis has increasingly been augmented with indirect, automated data processing, aided by other discoveries in computer science, such as neural networks, cluster analysis, genetic algorithms (1950s), decision trees and decision rules (1960s), and support vector machines (1990s). Data mining is the process of applying these methods with the intention of uncovering hidden patterns in large data sets. It bridges the gap from applied statistics and artificial intelligence (which usually provide the mathematical background) to database management by exploiting the way data is stored and indexed in databases to execute the actual learning and discovery algorithms more efficiently, allowing such methods to be applied to ever larger data sets.

Process

The Knowledge Discovery in Databases (KDD) process is commonly defined with the stages:

(1) Selection

(2) Pre-processing

(3) Transformation

(4) *Data Mining*

(5) Interpretation/Evaluation.

It exists, however, in many variations on this theme, such as the Cross Industry Standard Process for Data Mining (CRISP-DM) which defines six phases:

(1) Business Understanding

(2) Data Understanding

(3) Data Preparation

(4) Modeling

(5) Evaluation

(6) Deployment

or a simplified process such as (1) pre-processing, (2) data mining, and (3) results validation.

Polls conducted in 2002, 2004, 2007 and 2014 show that the CRISP-DM methodology is the leading methodology used by data miners. The only other data mining standard named in these polls was SEMMA. However, 3–4 times as many people reported using CRISP-DM. Several teams of researchers have published reviews of data mining process models, and Azevedo and Santos conducted a comparison of CRISP-DM and SEMMA in 2008.

Pre-processing

Before data mining algorithms can be used, a target data set must be assembled. As data mining can only uncover patterns actually present in the data, the target data set must be large enough to contain these patterns while remaining concise enough to be mined within an acceptable time limit. A common source for data is a data mart or data warehouse. Pre-processing is essential to analyze the multivariate data sets before data mining. The target set is then cleaned. Data cleaning removes the observations containing noise and those with missing data.

Data Mining

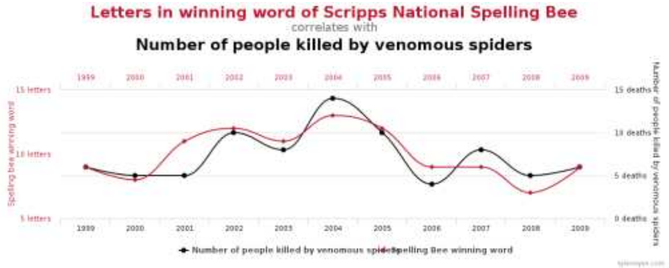
Data mining involves six common classes of tasks:

- Anomaly detection (Outlier/change/deviation detection) – The identification of unusual data records, that might be interesting or data errors that require further investigation.
- Association rule learning (Dependency modelling) – Searches for relationships between variables. For example, a supermarket might gather data on customer purchasing habits. Using association rule learning, the supermarket can determine which products are frequently bought together and use this information for marketing purposes. This is some-

times referred to as market basket analysis.

- Clustering – is the task of discovering groups and structures in the data that are in some way or another "similar", without using known structures in the data.
- Classification – is the task of generalizing known structure to apply to new data. For example, an e-mail program might attempt to classify an e-mail as "legitimate" or as "spam".
- Regression – attempts to find a function which models the data with the least error.
- Summarization – providing a more compact representation of the data set, including visualization and report generation.

Results Validation



An example of data produced by data dredging through a bot operated by statistician Tyler Viglen, apparently showing a close link between the best word winning a spelling bee competition and the number of people in the United States killed by venomous spiders. The similarity in trends is obviously a coincidence.

Data mining can unintentionally be misused, and can then produce results which appear to be significant; but which do not actually predict future behaviour and cannot be reproduced on a new sample of data and bear little use. Often this results from investigating too many hypotheses and not performing proper statistical hypothesis testing. A simple version of this problem in machine learning is known as overfitting, but the same problem can arise at different phases of the process and thus a train/test split - when applicable at all - may not be sufficient to prevent this from happening.

The final step of knowledge discovery from data is to verify that the patterns produced by the data mining algorithms occur in the wider data set. Not all patterns found by the data mining algorithms are necessarily valid. It is common for the data mining algorithms to find patterns in the training set which are not present in the general data set. This is called overfitting. To overcome this, the evaluation uses a test set of data on which the data mining algorithm was not trained. The learned patterns are applied to this test set, and the resulting output is compared to the desired output. For example, a data mining algorithm trying to distinguish "spam" from "legitimate" emails would be trained on a training set of sample e-mails. Once trained, the learned patterns would be applied to the test set of e-mails on which it had *not* been trained. The accuracy of the patterns can then be measured from how many e-mails they correctly classify. A number of statistical methods may be used to evaluate the algorithm, such as ROC curves.

If the learned patterns do not meet the desired standards, subsequently it is necessary to re-evaluate and change the pre-processing and data mining steps. If the learned patterns do meet the desired standards, then the final step is to interpret the learned patterns and turn them into knowledge.

Research

The premier professional body in the field is the Association for Computing Machinery's (ACM) Special Interest Group (SIG) on Knowledge Discovery and Data Mining (SIGKDD). Since 1989 this ACM SIG has hosted an annual international conference and published its proceedings, and since 1999 it has published a biannual academic journal titled "SIGKDD Explorations".

Computer science conferences on data mining include:

- CIKM Conference – ACM Conference on Information and Knowledge Management
- DMIN Conference – International Conference on Data Mining
- DMKD Conference – Research Issues on Data Mining and Knowledge Discovery
- DSAA Conference – IEEE International Conference on Data Science and Advanced Analytics
- ECDM Conference – European Conference on Data Mining
- ECML-PKDD Conference – European Conference on Machine Learning and Principles and Practice of Knowledge Discovery in Databases
- EDM Conference – International Conference on Educational Data Mining
- INFOCOM Conference – IEEE INFOCOM
- ICDM Conference – IEEE International Conference on Data Mining
- KDD Conference – ACM SIGKDD Conference on Knowledge Discovery and Data Mining
- MLDM Conference – Machine Learning and Data Mining in Pattern Recognition
- PAKDD Conference – The annual Pacific-Asia Conference on Knowledge Discovery and Data Mining
- PAW Conference – Predictive Analytics World
- SDM Conference – SIAM International Conference on Data Mining (SIAM)
- SSTD Symposium – Symposium on Spatial and Temporal Databases
- WSDM Conference – ACM Conference on Web Search and Data Mining

Data mining topics are also present on many data management/database conferences such as the ICDE Conference, SIGMOD Conference and International Conference on Very Large Data Bases

Standards

There have been some efforts to define standards for the data mining process, for example the 1999 European Cross Industry Standard Process for Data Mining (CRISP-DM 1.0) and the 2004 Java Data Mining standard (JDM 1.0). Development on successors to these processes (CRISP-DM 2.0 and JDM 2.0) was active in 2006, but has stalled since. JDM 2.0 was withdrawn without reaching a final draft.

For exchanging the extracted models – in particular for use in predictive analytics – the key standard is the Predictive Model Markup Language (PMML), which is an XML-based language developed by the Data Mining Group (DMG) and supported as exchange format by many data mining applications. As the name suggests, it only covers prediction models, a particular data mining task of high importance to business applications. However, extensions to cover (for example) subspace clustering have been proposed independently of the DMG.

Notable Uses

Data mining is used wherever there is digital data available today. Notable examples of data mining can be found throughout business, medicine, science, and surveillance.

Privacy Concerns and Ethics

While the term "data mining" itself has no ethical implications, it is often associated with the mining of information in relation to peoples' behavior (ethical and otherwise).

The ways in which data mining can be used can in some cases and contexts raise questions regarding privacy, legality, and ethics. In particular, data mining government or commercial data sets for national security or law enforcement purposes, such as in the Total Information Awareness Program or in ADVISE, has raised privacy concerns.

Data mining requires data preparation which can uncover information or patterns which may compromise confidentiality and privacy obligations. A common way for this to occur is through data aggregation. Data aggregation involves combining data together (possibly from various sources) in a way that facilitates analysis (but that also might make identification of private, individual-level data deducible or otherwise apparent). This is not data mining *per se*, but a result of the preparation of data before – and for the purposes of – the analysis. The threat to an individual's privacy comes into play when the data, once compiled, cause the data miner, or anyone who has access to the newly compiled data set, to be able to identify specific individuals, especially when the data were originally anonymous.

It is recommended that an individual is made aware of the following before data are collected:

- the purpose of the data collection and any (known) data mining projects;
- how the data will be used;
- who will be able to mine the data and use the data and their derivatives;
- the status of security surrounding access to the data;
- how collected data can be updated.

Data may also be modified so as to *become* anonymous, so that individuals may not readily be identified. However, even "de-identified"/"anonymized" data sets can potentially contain enough information to allow identification of individuals, as occurred when journalists were able to find several individuals based on a set of search histories that were inadvertently released by AOL.

The inadvertent revelation of personally identifiable information leading to the provider violates

Fair Information Practices. This indiscretion can cause financial, emotional, or bodily harm to the indicated individual. In one instance of privacy violation, the patrons of Walgreens filed a lawsuit against the company in 2011 for selling prescription information to data mining companies who in turn provided the data to pharmaceutical companies.

Situation in Europe

Europe has rather strong privacy laws, and efforts are underway to further strengthen the rights of the consumers. However, the U.S.-E.U. Safe Harbor Principles currently effectively expose European users to privacy exploitation by U.S. companies. As a consequence of Edward Snowden's Global surveillance disclosure, there has been increased discussion to revoke this agreement, as in particular the data will be fully exposed to the National Security Agency, and attempts to reach an agreement have failed.

Situation in The United States

In the United States, privacy concerns have been addressed by the US Congress via the passage of regulatory controls such as the Health Insurance Portability and Accountability Act (HIPAA). The HIPAA requires individuals to give their "informed consent" regarding information they provide and its intended present and future uses. According to an article in *Biotech Business Week', "'[i]n practice, HIPAA may not offer any greater protection than the longstanding regulations in the research arena,' says the AAHC. More importantly, the rule's goal of protection through informed consent is undermined by the complexity of consent forms that are required of patients and participants, which approach a level of incomprehensibility to average individuals." This underscores the necessity for data anonymity in data aggregation and mining practices.*

U.S. information privacy legislation such as HIPAA and the Family Educational Rights and Privacy Act (FERPA) applies only to the specific areas that each such law addresses. Use of data mining by the majority of businesses in the U.S. is not controlled by any legislation.

Copyright Law

Situation in Europe

Due to a lack of flexibilities in European copyright and database law, the mining of in-copyright works such as web mining without the permission of the copyright owner is not legal. Where a database is pure data in Europe there is likely to be no copyright, but database rights may exist so data mining becomes subject to regulations by the Database Directive. On the recommendation of the Hargreaves review this led to the UK government to amend its copyright law in 2014 to allow content mining as a limitation and exception. Only the second country in the world to do so after Japan, which introduced an exception in 2009 for data mining. However, due to the restriction of the Copyright Directive, the UK exception only allows content mining for non-commercial purposes. UK copyright law also does not allow this provision to be overridden by contractual terms and conditions. The European Commission facilitated stakeholder discussion on text and data mining in 2013, under the title of Licences for Europe. The focus on the solution to this legal issue being licences and not limitations and exceptions led to representatives of universities, researchers, libraries, civil society groups and open access publishers to leave the stakeholder dialogue in May 2013.

Situation in The United States

By contrast to Europe, the flexible nature of US copyright law, and in particular fair use means that content mining in America, as well as other fair use countries such as Israel, Taiwan and South Korea is viewed as being legal. As content mining is transformative, that is it does not supplant the original work, it is viewed as being lawful under fair use. For example, as part of the Google Book settlement the presiding judge on the case ruled that Google's digitisation project of in-copyright books was lawful, in part because of the transformative uses that the digitisation project displayed - one being text and data mining.

Software

Free Open-source Data Mining Software and Applications

The following applications are available under free/open source licenses. Public access to application sourcecode is also available.

- Carrot2: Text and search results clustering framework.
- Chemicalize.org: A chemical structure miner and web search engine.
- ELKI: A university research project with advanced cluster analysis and outlier detection methods written in the Java language.
- GATE: a natural language processing and language engineering tool.
- KNIME: The Konstanz Information Miner, a user friendly and comprehensive data analytics framework.
- Massive Online Analysis (MOA): a real-time big data stream mining with concept drift tool in the Java programming language.
- ML-Flex: A software package that enables users to integrate with third-party machine-learning packages written in any programming language, execute classification analyses in parallel across multiple computing nodes, and produce HTML reports of classification results.
- MLPACK library: a collection of ready-to-use machine learning algorithms written in the C++ language.
- MEPX - cross platform tool for regression and classification problems based on a Genetic Programming variant.
- NLTK (Natural Language Toolkit): A suite of libraries and programs for symbolic and statistical natural language processing (NLP) for the Python language.
- OpenNN: Open neural networks library.
- Orange: A component-based data mining and machine learning software suite written in the Python language.
- R: A programming language and software environment for statistical computing, data mining, and graphics. It is part of the GNU Project.

- scikit-learn is an open source machine learning library for the Python programming language
- Torch: An open source deep learning library for the Lua programming language and scientific computing framework with wide support for machine learning algorithms.
- UIMA: The UIMA (Unstructured Information Management Architecture) is a component framework for analyzing unstructured content such as text, audio and video – originally developed by IBM.
- Weka: A suite of machine learning software applications written in the Java programming language.

Proprietary Data-mining Software and Applications

The following applications are available under proprietary licenses.

- Angoss KnowledgeSTUDIO: data mining tool provided by Angoss.
- Clarabridge: enterprise class text analytics solution.
- HP Vertica Analytics Platform: data mining software provided by HP.
- IBM SPSS Modeler: data mining software provided by IBM.
- KXEN Modeler: data mining tool provided by KXEN.
- LIONsolver: an integrated software application for data mining, business intelligence, and modeling that implements the Learning and Intelligent OptimizatioN (LION) approach.
- Megaputer Intelligence: data and text mining software is called PolyAnalyst.
- Microsoft Analysis Services: data mining software provided by Microsoft.
- NetOwl: suite of multilingual text and entity analytics products that enable data mining.
- OpenText™ Big Data Analytics: Visual Data Mining & Predictive Analysis by Open Text Corporation
- Oracle Data Mining: data mining software by Oracle.
- PSeven: platform for automation of engineering simulation and analysis, multidisciplinary optimization and data mining provided by DATADVANCE.
- Qlucore Omics Explorer: data mining software provided by Qlucore.
- RapidMiner: An environment for machine learning and data mining experiments.
- SAS Enterprise Miner: data mining software provided by the SAS Institute.
- STATISTICA Data Miner: data mining software provided by StatSoft.
- Tanagra: A visualisation-oriented data mining software, also for teaching.

Marketplace Surveys

Several researchers and organizations have conducted reviews of data mining tools and surveys of data miners. These identify some of the strengths and weaknesses of the software packages. They also provide an overview of the behaviors, preferences and views of data miners. Some of these reports include:

- Hurwitz Victory Index: Report for Advanced Analytics as a market research assessment tool, it highlights both the diverse uses for advanced analytics technology and the vendors who make those applications possible.Recent-research
- 2011 Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery
- Rexer Analytics Data Miner Surveys (2007–2013)
- Forrester Research 2010 Predictive Analytics and Data Mining Solutions report
- Gartner 2008 “Magic Quadrant” report
- Robert A. Nisbet’s 2006 Three Part Series of articles “Data Mining Tools: Which One is Best For CRM?”
- Haughton et al.’s 2003 Review of Data Mining Software Packages in *The American Statistician*
- Goebel & Gruenwald 1999 “A Survey of Data Mining a Knowledge Discovery Software Tools” in SIGKDD Explorations

Anomaly Detection

In data mining, anomaly detection (also outlier detection) is the identification of items, events or observations which do not conform to an expected pattern or other items in a dataset. Typically the anomalous items will translate to some kind of problem such as bank fraud, a structural defect, medical problems or errors in a text. Anomalies are also referred to as outliers, novelties, noise, deviations and exceptions.

In particular, in the context of abuse and network intrusion detection, the interesting objects are often not *rare* objects, but unexpected *bursts* in activity. This pattern does not adhere to the common statistical definition of an outlier as a rare object, and many outlier detection methods (in particular unsupervised methods) will fail on such data, unless it has been aggregated appropriately. Instead, a cluster analysis algorithm may be able to detect the micro clusters formed by these patterns.

Three broad categories of anomaly detection techniques exist. Unsupervised anomaly detection techniques detect anomalies in an unlabeled test data set under the assumption that the majority of the instances in the data set are normal by looking for instances that seem to fit least to the remainder of the data set. Supervised anomaly detection techniques require a data set that has been labeled as “normal” and “abnormal” and involves training a classifier (the key difference to many other statistical classification problems is the inherent unbalanced nature of outlier detection).

Semi-supervised anomaly detection techniques construct a model representing normal behavior from a given *normal* training data set, and then testing the likelihood of a test instance to be generated by the learnt model.

Applications

Anomaly detection is applicable in a variety of domains, such as intrusion detection, fraud detection, fault detection, system health monitoring, event detection in sensor networks, and detecting Eco-system disturbances. It is often used in preprocessing to remove anomalous data from the dataset. In supervised learning, removing the anomalous data from the dataset often results in a statistically significant increase in accuracy.

Popular Techniques

Several anomaly detection techniques have been proposed in literature. Some of the popular techniques are:

- Density-based techniques (k-nearest neighbor, local outlier factor, and many more variations of this concept).
- Subspace- and correlation-based outlier detection for high-dimensional data.
- One class support vector machines.
- Replicator neural networks.
- Cluster analysis-based outlier detection.
- Deviations from association rules and frequent itemsets.
- Fuzzy logic based outlier detection.
- Ensemble techniques, using feature bagging, score normalization and different sources of diversity.

The performance of different methods depends a lot on the data set and parameters, and methods have little systematic advantages over another when compared across many data sets and parameters.

Application to Data Security

Anomaly detection was proposed for intrusion detection systems (IDS) by Dorothy Denning in 1986. Anomaly detection for IDS is normally accomplished with thresholds and statistics, but can also be done with soft computing, and inductive learning. Types of statistics proposed by 1999 included profiles of users, workstations, networks, remote hosts, groups of users, and programs based on frequencies, means, variances, covariances, and standard deviations. The counterpart of anomaly detection in intrusion detection is misuse detection.

Software

ELKI is an open-source Java data mining toolkit that contains several anomaly detection algorithms, as well as index acceleration for them.

Association Rule Learning

Association rule learning is a rule-based machine learning method for discovering interesting relations between variables in large databases. It is intended to identify strong rules discovered in databases using some measures of interestingness. Based on the concept of strong rules, Rakesh Agrawal et al. introduced association rules for discovering regularities between products in large-scale transaction data recorded by point-of-sale (POS) systems in supermarkets. For example, the rule $\{\text{onions}, \text{potatoes}\} \Rightarrow \{\text{burger}\}$ found in the sales data of a supermarket would indicate that if a customer buys onions and potatoes together, they are likely to also buy hamburger meat. Such information can be used as the basis for decisions about marketing activities such as, e.g., promotional pricing or product placements. In addition to the above example from market basket analysis association rules are employed today in many application areas including Web usage mining, intrusion detection, Continuous production, and bioinformatics. In contrast with sequence mining, association rule learning typically does not consider the order of items either within a transaction or across transactions.

Definition

Example database with 5 transactions and 5 items					
transaction ID	milk	bread	butter	beer	diapers
1	1	1	0	0	0
2	0	0	1	0	0
3	0	0	0	1	1
4	1	1	1	0	0
5	0	1	0	0	0

Following the original definition by Agrawal et al. the problem of association rule mining is defined as:

Let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n binary attributes called *items*.

Let $D = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the *database*.

Each *transaction* in D has a unique transaction ID and contains a subset of the items in I .

A *rule* is defined as an implication of the form:

$$X \Rightarrow Y$$

Where $X, Y \subseteq I$ and $X \cap Y = \emptyset$..

Every rule is composed by two different sets of items, also known as *itemsets*, X and Y, where X is called *antecedent* or left-hand-side (LHS) and Y *consequent* or right-hand-side (RHS).

To illustrate the concepts, we use a small example from the supermarket domain. The set of items is $I = \{\text{milk}, \text{bread}, \text{butter}, \text{beer}, \text{diapers}\}$ and in the table is shown a small database containing the items, where, in each entry, the value 1 means the presence of the item in the corresponding transaction, and the value 0 represents the absence of an item in that transaction.

An example rule for the supermarket could be $\{\text{butter}, \text{bread}\} \Rightarrow \{\text{milk}\}$ meaning that if butter and bread are bought, customers also buy milk.

Note: this example is extremely small. In practical applications, a rule needs a support of several hundred transactions before it can be considered statistically significant, and data-sets often contain thousands or millions of transactions.

Useful Concepts

In order to select interesting rules from the set of all possible rules, constraints on various measures of significance and interest are used. The best-known constraints are minimum thresholds on support and confidence.

Let X be an item-set, $X \Rightarrow Y$ an association rule and T a set of transactions of a given database.

Support

Support is an indication of how frequently the item-set appears in the database.

The support value of X with respect to T is defined as the proportion of transactions in the database which contains the item-set X. In formula: $\text{supp}(X)/N$

In the example database, the item-set $\{\text{beer}, \text{diapers}\}$ has a support of since it occurs in 20% of all transactions (1 out of 5 transactions). The argument of $\text{supp}()$ is a set of preconditions, and thus becomes more restrictive as it grows (instead of more inclusive).

Confidence

Confidence is an indication of how often the rule has been found to be true.

The *confidence* value of a rule, $X \Rightarrow Y$, with respect to a set of transactions T, is the proportion of the transactions that contains X which also contains Y.

Confidence is defined as:

$$\text{conf}(X \Rightarrow Y) = \text{supp}(X \cup Y)/\text{supp}(X).$$

For example, the rule $\{\text{butter}, \text{bread}\} \Rightarrow \{\text{milk}\}$ has a confidence of $0.2 >$ in the database, which means that for 100% of the transactions containing butter and bread the rule is correct (100% of the times a customer buys butter and bread, milk is bought as well).

Note that $\text{supp}(X \cup Y)$ means the support of the union of the items in X and Y. This is somewhat confusing since we normally think in terms of probabilities of events and not sets of items. We can rewrite $\text{supp}(X \cup Y)$ as the joint probability $P(E_X \cap E_Y)$, where E_X and E_Y are the events that a transaction contains itemset X or Y, respectively.

Thus confidence can be interpreted as an estimate of the conditional probability $P(E_Y \mid E_X)$, the probability of finding the RHS of the rule in transactions under the condition that these transactions also contain the LHS.

Lift

The *lift* of a rule is defined as:

$$\mathrm{lift}(X \Rightarrow Y) = \frac{\mathrm{supp}(X \cup Y)}{\mathrm{supp}(X) \times \mathrm{supp}(Y)}$$

or the ratio of the observed support to that expected if X and Y were independent.

For example, the rule $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ has a lift of $\frac{0.2}{0.4 \times 0.4} = 1.25$.

If the rule had a lift of 1, it would imply that the probability of occurrence of the antecedent and that of the consequent are independent of each other. When two events are independent of each other, no rule can be drawn involving those two events.

If the lift is > 1, that lets us know the degree to which those two occurrences are dependent on one another, and makes those rules potentially useful for predicting the consequent in future data sets.

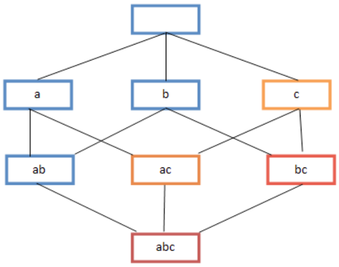
The value of lift is that it considers both the confidence of the rule and the overall data set.

Conviction

The *conviction* of a rule is defined as $\mathrm{conv}(X \Rightarrow Y) = \frac{1 - \mathrm{supp}(Y)}{1 - \mathrm{conf}(X \Rightarrow Y)}$.

For example, the rule $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ has a conviction of $\frac{1 - 0.4}{1 - 0.5} = 1.2$, and can be interpreted as the ratio of the expected frequency that X occurs without Y (that is to say, the frequency that the rule makes an incorrect prediction) if X and Y were independent divided by the observed frequency of incorrect predictions. In this example, the conviction value of 1.2 shows that the rule $\{\mathrm{milk}, \mathrm{bread}\} \Rightarrow \{\mathrm{butter}\}$ would be incorrect 20% more often (1.2 times as often) if the association between X and Y was purely random chance.

Process



Frequent itemset lattice, where the color of the box indicates how many transactions contain the combination of items. Note that lower levels of the lattice can contain at most the minimum number of their parents' items; e.g. {ac} can have only at most $min(a,c)$ items. This is called the *downward-closure property*.

Association rules are usually required to satisfy a user-specified minimum support and a user-specified minimum confidence at the same time. Association rule generation is usually split up into two separate steps:

1. A minimum support threshold is applied to find all *frequent item-sets* in a database.
2. A minimum confidence constraint is applied to these frequent item-sets in order to form rules.

While the second step is straightforward, the first step needs more attention.

Finding all frequent item-sets in a database is difficult since it involves searching all possible item-sets (item combinations). The set of possible item-sets is the power set over I and has size $2^n - 1$ (excluding the empty set which is not a valid item-set). Although the size of the power-set grows exponentially in the number of items n in I, efficient search is possible using the *downward-closure property* of support (also called *anti-monotonicity*) which guarantees that for a frequent itemset, all its subsets are also frequent and thus for an infrequent item-set, all its super-sets must also be infrequent. Exploiting this property, efficient algorithms (e.g., Apriori and Eclat) can find all frequent item-sets.

History

The concept of association rules was popularised particularly due to the 1993 article of Agrawal et al., which has acquired more than 18,000 citations according to Google Scholar, as of August 2015, and is thus one of the most cited papers in the Data Mining field. However, it is possible that what is now called "association rules" is similar to what appears in the 1966 paper on GUHA, a general data mining method developed by Petr Hájek et al.

An early (circa 1989) use of minimum support and confidence to find all association rules is the Feature Based Modeling framework, which found all rules with $\mathrm{supp}(X)$ and $\mathrm{conf}(X \Rightarrow Y)$ greater than user defined constraints.

Alternative Measures of Interestingness

In addition to confidence, other measures of *interestingness* for rules have been proposed. Some popular measures are:

- All-confidence
- Collective strength
- Conviction
- Leverage
- Lift (originally called interest)

Several more measures are presented and compared by Tan et al. and by Hahsler. Looking for techniques that can model what the user has known (and using these models as interestingness measures) is currently an active research trend under the name of "Subjective Interestingness."

Statistically Sound Associations

One limitation of the standard approach to discovering associations is that by searching massive numbers of possible associations to look for collections of items that appear to be associated, there is a large risk of finding many spurious associations. These are collections of items that co-occur with unexpected frequency in the data, but only do so by chance. For example, suppose we are considering a collection of 10,000 items and looking for rules containing two items in the left-hand-side and 1 item in the right-hand-side. There are approximately 1,000,000,000,000 such rules. If we apply a statistical test for independence with a significance level of 0.05 it means there is only a 5% chance of accepting a rule if there is no association. If we assume there are no associations, we should nonetheless expect to find 50,000,000,000 rules. Statistically sound association discovery controls this risk, in most cases reducing the risk of finding *any* spurious associations to a user-specified significance levels.

Algorithms

Many algorithms for generating association rules were presented over time.

Some well known algorithms are Apriori, Eclat and FP-Growth, but they only do half the job, since they are algorithms for mining frequent itemsets. Another step needs to be done after to generate rules from frequent itemsets found in a database.

Apriori Algorithm

Apriori uses a breadth-first search strategy to count the support of itemsets and uses a candidate generation function which exploits the downward closure property of support.

Eclat Algorithm

Eclat (alt. ECLAT, stands for Equivalence Class Transformation) is a depth-first search algorithm using set intersection. It is a naturally elegant algorithm suitable for both sequential as well as parallel execution with locality enhancing properties. It was first introduced by Zaki, Parthasarathy, Li and Ogihara in a series of papers written in 1997.

Mohammed Javeed Zaki, Srinivasan Parthasarathy, M. Ogihara, Wei Li: New Algorithms for Fast Discovery of Association Rules. KDD 1997.

Mohammed Javeed Zaki, Srinivasan Parthasarathy, Mitsunori Ogihara, Wei Li: Parallel Algorithms for Discovery of Association Rules. Data Min. Knowl. Discov. 1(4): 343-373 (1997)

FP-growth Algorithm

FP stands for frequent pattern.

In the first pass, the algorithm counts occurrence of items (attribute-value pairs) in the dataset, and stores them to 'header table'. In the second pass, it builds the FP-tree structure by inserting instances. Items in each instance have to be sorted by descending order of their frequency in the dataset, so that the tree can be processed quickly. Items in each instance that do not meet minimum coverage threshold are discarded. If many instances share most frequent items, FP-tree provides high compression close to tree root.

Recursive processing of this compressed version of main dataset grows large item sets directly, instead of generating candidate items and testing them against the entire database. Growth starts from the bottom of the header table (having longest branches), by finding all instances matching given condition. New tree is created, with counts projected from the original tree corresponding to the set of instances that are conditional on the attribute, with each node getting sum of its children counts. Recursive growth ends when no individual items conditional on the attribute meet minimum support threshold, and processing continues on the remaining header items of the original FP-tree.

Once the recursive process has completed, all large item sets with minimum coverage have been found, and association rule creation begins.

Others

AprioriDP

AprioriDP utilizes Dynamic Programming in Frequent itemset mining. The working principle is to eliminate the candidate generation like FP-tree, but it stores support count in specialized data structure instead of tree.

Context Based Association Rule Mining Algorithm

CBPNARM is the newly developed algorithm which is developed in 2013 to mine association rules on the basis of context. It uses context variable on the basis of which the support of an itemset is changed on the basis of which the rules are finally populated to the rule set.

Node-set-based Algorithms

FIN, PrePost and PPV are three algorithms based on node sets. They use nodes in a coding FP-tree to represent itemsets, and employ a depth-first search strategy to discovery frequent itemsets using “intersection” of node sets.

GUHA Procedure ASSOC

GUHA is a general method for exploratory data analysis that has theoretical foundations in observational calculi.

The ASSOC procedure is a GUHA method which mines for generalized association rules using fast bitstrings operations. The association rules mined by this method are more general than those output by apriori, for example “items” can be connected both with conjunction and disjunctions and the relation between antecedent and consequent of the rule is not restricted to setting minimum support and confidence as in apriori: an arbitrary combination of supported interest measures can be used.

OPUS Search

OPUS is an efficient algorithm for rule discovery that, in contrast to most alternatives, does not require either monotone or anti-monotone constraints such as minimum support. Initially used to find rules for a fixed consequent it has subsequently been extended to find rules with any item as a consequent. OPUS search is the core technology in the popular Magnum Opus association discovery system.

Lore

A famous story about association rule mining is the "beer and diaper" story. A purported survey of behavior of supermarket shoppers discovered that customers (presumably young men) who buy diapers tend also to buy beer. This anecdote became popular as an example of how unexpected association rules might be found from everyday data. There are varying opinions as to how much of the story is true. Daniel Powers says:

In 1992, Thomas Blischok, manager of a retail consulting group at Teradata, and his staff prepared an analysis of 1.2 million market baskets from about 25 Osco Drug stores. Database queries were developed to identify affinities. The analysis "did discover that between 5:00 and 7:00 p.m. that consumers bought beer and diapers". Osco managers did NOT exploit the beer and diapers relationship by moving the products closer together on the shelves.

Other Types of Association Mining

Multi-Relation Association Rules: Multi-Relation Association Rules (MRAR) is a new class of association rules which in contrast to primitive, simple and even multi-relational association rules (that are usually extracted from multi-relational databases), each rule item consists of one entity but several relations. These relations indicate indirect relationship between the entities. Consider the following MRAR where the first item consists of three relations *live in*, *nearby* and *humid*: "Those who *live in* a place which is *near by* a city with *humid* climate type and also are *younger* than 20 -> their *health condition* is good". Such association rules are extractable from RDBMS data or semantic web data.

Context Based Association Rules is a form of association rule. Context Based Association Rules claims more accuracy in association rule mining by considering a hidden variable named context variable which changes the final set of association rules depending upon the value of context variables. For example the baskets orientation in market basket analysis reflects an odd pattern in the early days of month.This might be because of abnormal context i.e. salary is drawn at the start of the month

Contrast set learning is a form of associative learning. Contrast set learners use rules that differ meaningfully in their distribution across subsets.

Weighted class learning is another form of associative learning in which weight may be assigned to classes to give focus to a particular issue of concern for the consumer of the data mining results.

High-order pattern discovery facilitate the capture of high-order (polythetic) patterns or event associations that are intrinsic to complex real-world data.

K-optimal pattern discovery provides an alternative to the standard approach to association rule learning that requires that each pattern appear frequently in the data.

Approximate Frequent Itemset mining is a relaxed version of Frequent Itemset mining that allows some of the items in some of the rows to be 0.

Generalized Association Rules hierarchical taxonomy (concept hierarchy)

Quantitative Association Rules categorical and quantitative data

Interval Data Association Rules e.g. partition the age into 5-year-increment ranged

Maximal Association Rules

Sequential pattern mining discovers subsequences that are common to more than minsup sequences in a sequence database, where minsup is set by the user. A sequence is an ordered list of transactions.

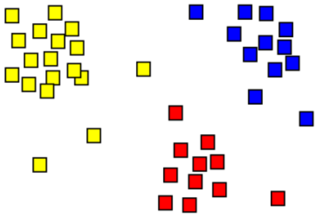
Sequential Rules discovering relationships between items while considering the time ordering. It is generally applied on a sequence database. For example, a sequential rule found in database of sequences of customer transactions can be that customers who bought a computer and CD-Roms, later bought a webcam, with a given confidence and support.

Subspace Clustering, a specific type of Clustering high-dimensional data, is in many variants also based on the downward-closure property for specific clustering models.

Warmr is shipped as part of the ACE data mining suite. It allows association rule learning for first order relational rules.

Expected float entropy minimisation simultaneously uncovers relationships between nodes (referred to as items on this page) of a system and also between the different states that the nodes can be in. The theory links associations inherent to systems such as the brain to associations present in perception.

Cluster Analysis



The result of a cluster analysis shown as the coloring of the squares into three clusters.

Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). It is a main task of exploratory data mining, and a common technique for statistical data analysis, used in many fields, including machine learning, pattern recognition, image analysis, information retrieval, bioinformatics, data compression, and computer graphics.

Cluster analysis itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It is often necessary to modify data preprocessing and model parameters until the result achieves the desired properties.

Besides the term *clustering*, there are a number of terms with similar meanings, including *automatic classification*, *numerical taxonomy*, *botryology* and *typological analysis*. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification the resulting discriminative power is of interest.

Cluster analysis was originated in anthropology by Driver and Kroeber in 1932 and introduced to psychology by Zubin in 1938 and Robert Tryon in 1939 and famously used by Cattell beginning in 1943 for trait theory classification in personality psychology.

Definition

The notion of a "cluster" cannot be precisely defined, which is one of the reasons why there are so many clustering algorithms. There is a common denominator: a group of data objects. However, different researchers employ different cluster models, and for each of these cluster models again different algorithms can be given. The notion of a cluster, as found by different algorithms, varies significantly in its properties. Understanding these "cluster models" is key to understanding the differences between the various algorithms. Typical cluster models include:

- Connectivity models: for example, hierarchical clustering builds models based on distance connectivity.
- Centroid models: for example, the k-means algorithm represents each cluster by a single mean vector.
- Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.
- Density models: for example, DBSCAN and OPTICS defines clusters as connected dense regions in the data space.
- Subspace models: in Biclustering (also known as Co-clustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.
- Group models: some algorithms do not provide a refined model for their results and just provide the grouping information.
- Graph-based models: a clique, that is, a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasi-cliques, as in the HCS clustering algorithm.

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example, a hierarchy of clusters embedded in each other. Clusterings can be roughly distinguished as:

- hard clustering: each object belongs to a cluster or not
- soft clustering (also: fuzzy clustering): each object belongs to each cluster to a certain degree (for example, a likelihood of belonging to the cluster)

There are also finer distinctions possible, for example:

- strict partitioning clustering: here each object belongs to exactly one cluster
- strict partitioning clustering with outliers: objects can also belong to no cluster, and are considered outliers.
- overlapping clustering (also: alternative clustering, multi-view clustering): while usually a hard clustering, objects may belong to more than one cluster.
- hierarchical clustering: objects that belong to a child cluster also belong to the parent cluster
- subspace clustering: while an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap.

Algorithms

Clustering algorithms can be categorized based on their cluster model, as listed above. The following overview will only list the most prominent examples of clustering algorithms, as there are possibly over 100 published clustering algorithms. Not all provide models for their clusters and can thus not easily be categorized.

There is no objectively "correct" clustering algorithm, but as it was noted, "clustering is in the eye of the beholder." The most appropriate clustering algorithm for a particular problem often needs to be chosen experimentally, unless there is a mathematical reason to prefer one cluster model over another. It should be noted that an algorithm that is designed for one kind of model has no chance on a data set that contains a radically different kind of model. For example, k-means cannot find non-convex clusters.

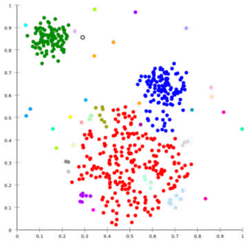
Connectivity-based Clustering (Hierarchical Clustering)

Connectivity based clustering, also known as *hierarchical clustering*, is based on the core idea of objects being more related to nearby objects than to objects farther away. These algorithms connect "objects" to form "clusters" based on their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form, which can be represented using a dendrogram, which explains where the common name "hierarchical clustering" comes from: these algorithms do not provide a single partitioning of the data set, but instead provide an extensive hierarchy of clusters that merge with each other at certain distances. In a dendrogram, the y-axis marks the distance at which the clusters merge, while the objects are placed along the x-axis such that the clusters don't mix.

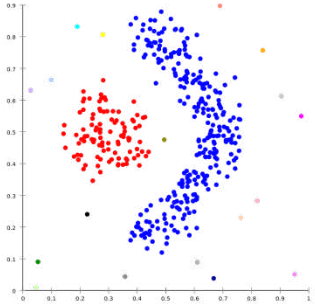
Connectivity based clustering is a whole family of methods that differ by the way distances are computed. Apart from the usual choice of distance functions, the user also needs to decide on the linkage criterion (since a cluster consists of multiple objects, there are multiple candidates to compute the distance to) to use. Popular choices are known as single-linkage clustering (the minimum of object distances), complete linkage clustering (the maximum of object distances) or UPGMA ("Unweighted Pair Group Method with Arithmetic Mean", also known as average linkage clustering). Furthermore, hierarchical clustering can be agglomerative (starting with single elements and aggregating them into clusters) or divisive (starting with the complete data set and dividing it into partitions).

These methods will not produce a unique partitioning of the data set, but a hierarchy from which the user still needs to choose appropriate clusters. They are not very robust towards outliers, which will either show up as additional clusters or even cause other clusters to merge (known as "chaining phenomenon", in particular with single-linkage clustering). In the general case, the complexity is $\mathcal{O}(n^3)$ for agglomerative clustering and $\mathcal{O}(2^{n-1})$ for divisive clustering, which makes them too slow for large data sets. For some special cases, optimal efficient methods (of complexity $\mathcal{O}(n^2)$) are known: SLINK for single-linkage and CLINK for complete-linkage clustering. In the data mining community these methods are recognized as a theoretical foundation of cluster analysis, but often considered obsolete. They did however provide inspiration for many later methods such as density based clustering.

Linkage clustering examples



Single-linkage on Gaussian data. At 35 clusters, the biggest cluster starts fragmenting into smaller parts, while before it was still connected to the second largest due to the single-link effect.



Single-linkage on density-based clusters. 20 clusters extracted, most of which contain single elements, since linkage clustering does not have a notion of "noise".

Centroid-based Clustering

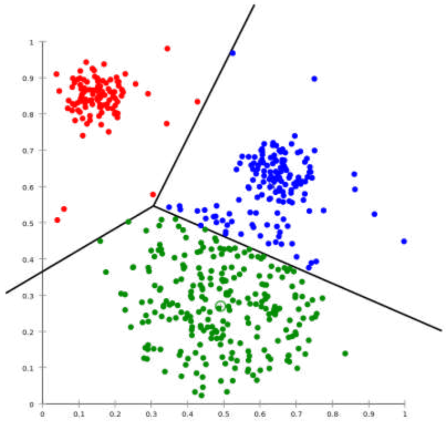
In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the *k* cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

The optimization problem itself is known to be NP-hard, and thus the common approach is to search only for approximate solutions. A particularly well known approximative method is Lloyd's algorithm, often actually referred to as "*k-means algorithm*". It does however only find a local optimum, and is commonly run multiple times with different random initializations. Variations of k-means often include such optimizations as choosing the best of multiple runs, but also restricting the centroids to members of the data set (k-medoids), choosing medians (k-medians clustering), choosing the initial centers less randomly (K-means++) or allowing a fuzzy cluster assignment (Fuzzy c-means).

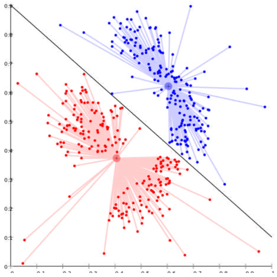
Most k-means-type algorithms require the number of clusters - *k* - to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Furthermore, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. This often leads to incorrectly cut borders in between of clusters (which is not surprising, as the algorithm optimized cluster centers, not cluster borders).

K-means has a number of interesting theoretical properties. First, it partitions the data space into a structure known as a Voronoi diagram. Second, it is conceptually close to nearest neighbor classification, and as such is popular in machine learning. Third, it can be seen as a variation of model based classification, and Lloyd's algorithm as a variation of the Expectation-maximization algorithm for this model discussed below.

k-Means clustering examples



K-means separates data into Voronoi-cells, which assumes equal-sized clusters (not adequate here)



K-means cannot represent density-based clusters

Distribution-based Clustering

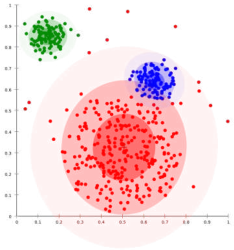
The clustering model most closely related to statistics is based on distribution models. Clusters can then easily be defined as objects belonging most likely to the same distribution. A convenient property of this approach is that this closely resembles the way artificial data sets are generated: by sampling random objects from a distribution.

While the theoretical foundation of these methods is excellent, they suffer from one key problem known as overfitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

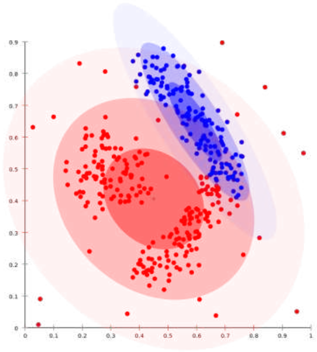
One prominent method is known as Gaussian mixture models (using the expectation-maximization algorithm). Here, the data set is usually modelled with a fixed (to avoid overfitting) number of Gaussian distributions that are initialized randomly and whose parameters are iteratively optimized to fit better to the data set. This will converge to a local optimum, so multiple runs may produce different results. In order to obtain a hard clustering, objects are often then assigned to the Gaussian distribution they most likely belong to; for soft clusterings, this is not necessary.

Distribution-based clustering produces complex models for clusters that can capture correlation and dependence between attributes. However, these algorithms put an extra burden on the user: for many real data sets, there may be no concisely defined mathematical model (e.g. assuming Gaussian distributions is a rather strong assumption on the data).

Expectation-Maximization (EM) clustering examples



On Gaussian-distributed data, EM works well, since it
uses Gaussians for modelling clusters



Density-based clusters cannot be modeled using Gaussian distributions

Density-based Clustering

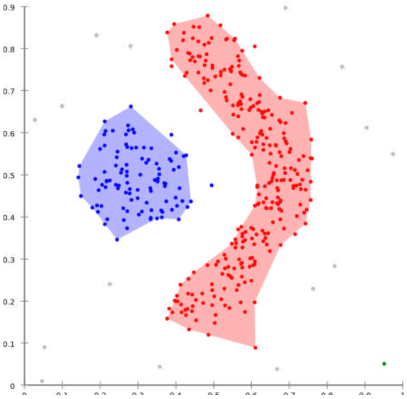
In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points.

The most popular density based clustering method is DBSCAN. In contrast to many newer methods, it features a well-defined cluster model called "density-reachability". Similar to linkage based clustering, it is based on connecting points within certain distance thresholds. However, it only connects points that satisfy a density criterion, in the original variant defined as a minimum number of other objects within this radius. A cluster consists of all density-connected objects (which can form a cluster of an arbitrary shape, in contrast to many other methods) plus all objects that are within these objects' range. Another interesting property of DBSCAN is that its complexity is fairly low - it requires a linear number of range queries on the database - and that it will discover essentially the same results (it is deterministic for core and noise points, but not for border points) in each run, therefore there is no need to run it multiple times. OPTICS is a generalization of DBSCAN that removes the need to choose an appropriate value for the range parameter ε, and produces a hierarchical result related to that of linkage clustering. DeLi-Clu, Density-Link-Clustering combines ideas from single-linkage clustering and OPTICS, eliminating the ε parameter entirely and offering performance improvements over OPTICS by using an R-tree index.

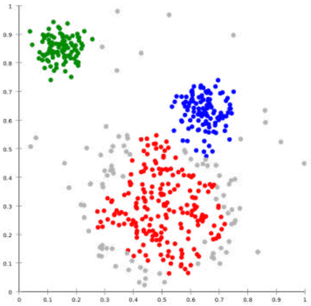
The key drawback of DBSCAN and OPTICS is that they expect some kind of density drop to detect cluster borders. Moreover, they cannot detect intrinsic cluster structures which are prevalent in the majority of real life data. A variation of DBSCAN, EnDBSCAN, efficiently detects such kinds of structures. On data sets with, for example, overlapping Gaussian distributions - a common use case in artificial data - the cluster borders produced by these algorithms will often look arbitrary, because the cluster density decreases continuously. On a data set consisting of mixtures of Gaussians, these algorithms are nearly always outperformed by methods such as EM clustering that are able to precisely model this kind of data.

Mean-shift is a clustering approach where each object is moved to the densest area in its vicinity, based on kernel density estimation. Eventually, objects converge to local maxima of density. Similar to k-means clustering, these "density attractors" can serve as representatives for the data set, but mean-shift can detect arbitrary-shaped clusters similar to DBSCAN. Due to the expensive iterative procedure and density estimation, mean-shift is usually slower than DBSCAN or k-Means.

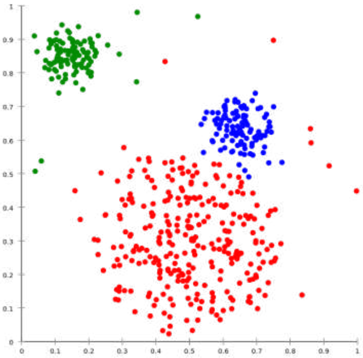
Density-based clustering examples



Density-based clustering with DBSCAN.



DBSCAN assumes clusters of similar density, and may have problems separating nearby clusters



OPTICS is a DBSCAN variant that handles different densities much better

Recent Developments

In recent years considerable effort has been put into improving the performance of existing algorithms. Among them are *CLARANS* (Ng and Han, 1994), and *BIRCH* (Zhang et al., 1996). With the recent need to process larger and larger data sets (also known as big data), the willingness to trade semantic meaning of the generated clusters for performance has been increasing. This led to the development of pre-clustering methods such as canopy clustering, which can process huge data sets efficiently, but the resulting "clusters" are merely a rough pre-partitioning of the data set to then analyze the partitions with existing slower methods such as k-means clustering. Various other approaches to clustering have been tried such as seed based clustering.

For high-dimensional data, many of the existing methods fail due to the curse of dimensionality, which renders particular distance functions problematic in high-dimensional spaces. This led to new clustering algorithms for high-dimensional data that focus on subspace clustering (where only some attributes are used, and cluster models include the relevant attributes for the cluster) and correlation clustering that also looks for arbitrary rotated ("correlated") subspace clusters that can be modeled by giving a correlation of their attributes. Examples for such clustering algorithms are CLIQUE and SUBCLU.

Ideas from density-based clustering methods (in particular the DBSCAN/OPTICS family of algorithms) have been adopted to subspace clustering (HiSC, hierarchical subspace clustering and DiSH) and correlation clustering (HiCO, hierarchical correlation clustering, 4C using "correlation connectivity" and ERiC exploring hierarchical density-based correlation clusters).

Several different clustering systems based on mutual information have been proposed. One is Marina Meilă's *variation of information* metric; another provides hierarchical clustering. Using genetic algorithms, a wide range of different fit-functions can be optimized, including mutual information. Also message passing algorithms, a recent development in Computer Science and Statistical Physics, has led to the creation of new types of clustering algorithms.

Other Methods

Basic sequential algorithmic scheme (BSAS)

Evaluation and Assessment

Evaluation of clustering results sometimes is referred to as cluster validation.

There have been several suggestions for a measure of similarity between two clusterings. Such a measure can be used to compare how well different data clustering algorithms perform on a set of data. These measures are usually tied to the type of criterion being considered in assessing the quality of a clustering method.

Internal Evaluation

When a clustering result is evaluated based on the data that was clustered itself, this is called internal evaluation. These methods usually assign the best score to the algorithm that produces clusters with high similarity within a cluster and low similarity between clusters. One drawback of using internal criteria in cluster evaluation is that high scores on an internal measure do not necessarily result in effective information retrieval applications. Additionally, this evaluation is biased towards algorithms that use the same cluster model. For example, k-Means clustering naturally optimizes object distances, and a distance-based internal criterion will likely overrate the resulting clustering.

Therefore, the internal evaluation measures are best suited to get some insight into situations where one algorithm performs better than another, but this shall not imply that one algorithm produces more valid results than another. Validity as measured by such an index depends on the claim that this kind of structure exists in the data set. An algorithm designed for some kind of models has no chance if the data set contains a radically different set of models, or if the evaluation measures a radically different criterion. For example, k-means clustering can only find convex clusters, and many evaluation indexes assume convex clusters. On a data set with non-convex clusters neither the use of k-means, nor of an evaluation criterion that assumes convexity, is sound.

The following methods can be used to assess the quality of clustering algorithms based on internal criterion:

- Davies–Bouldin index

  The Davies–Bouldin index can be calculated by the following formula:

  $$DB = \frac{1}{n}\sum_{i=1}^{n} \max_{j \neq i} \left(\frac{\sigma_i + \sigma_j}{d(c_i, c_j)} \right)$$

  where n is the number of clusters, c_x is the centroid of cluster x, σ_x is the average distance of all elements in cluster x to centroid c_x, and $d(c_i, c_j)$ is the distance between centroids c_i and c_j. Since algorithms that produce clusters with low intra-cluster distances (high intra-cluster similarity) and high inter-cluster distances (low inter-cluster similarity) will have a low Davies–Bouldin index, the clustering algorithm that produces a collection of clusters with the smallest Davies–Bouldin index is considered the best algorithm based on this criterion.

- Dunn index

  The Dunn index aims to identify dense and well-separated clusters. It is defined as the ratio between the minimal inter-cluster distance to maximal intra-cluster distance. For each cluster partition, the Dunn index can be calculated by the following formula:

  $$D = \frac{\min_{1 \leq i < j \leq n} d(i, j)}{\max_{1 \leq k \leq n} d^{'}(k)},$$

  where $d(i,j)$ represents the distance between clusters i and j, and d '(k) measures the intra-cluster distance of cluster k. The inter-cluster distance $d(i,j)$ between two clusters may be any number of distance measures, such as the distance between the centroids of the clusters. Similarly, the intra-cluster distance d '(k) may be measured in a variety ways, such as the maximal distance between any pair of elements in cluster k. Since internal criterion seek clusters with high intra-cluster similarity and low inter-cluster similarity, algorithms that produce clusters with high Dunn index are more desirable.

- Silhouette coefficient

  The silhouette coefficient contrasts the average distance to elements in the same cluster with the average distance to elements in other clusters. Objects with a high silhouette value are considered well clustered, objects with a low value may be outliers. This index works well with k-means clustering, and is also used to determine the optimal number of clusters.

External Evaluation

In external evaluation, clustering results are evaluated based on data that was not used for clustering, such as known class labels and external benchmarks. Such benchmarks consist of a set of pre-classified items, and these sets are often created by (expert) humans. Thus, the benchmark sets can be thought of as a gold standard for evaluation. These types of evaluation methods measure how close the clustering is to the predetermined benchmark classes. However, it has recently been discussed whether this is adequate for real data, or only on synthetic data sets with a factual ground truth, since classes can contain internal structure, the attributes present may not allow separation of clusters or the classes may contain anomalies. Additionally, from a knowledge discovery point of view, the reproduction of known knowledge may not necessarily be the intended result. In the special scenario of constrained clustering, where meta information (such as class labels) is used already in the clustering process, the hold-out of information for evaluation purposes is non-trivial.

A number of measures are adapted from variants used to evaluate classification tasks. In place of counting the number of times a class was correctly assigned to a single data point (known as true positives), such *pair counting* metrics assess whether each pair of data points that is truly in the same cluster is predicted to be in the same cluster.

Some of the measures of quality of a cluster algorithm using external criterion include:

- Rand measure (William M. Rand 1971)

  The Rand index computes how similar the clusters (returned by the clustering algorithm) are to the benchmark classifications. One can also view the Rand index as a measure of the percentage of correct decisions made by the algorithm. It can be computed using the following formula:

$$RI = \frac{TP + TN}{TP + FP + FN + TN}$$

  where TP is the number of true positives, TN is the number of true negatives, FP is the number of false positives, and FN is the number of false negatives. One issue with the Rand index is that false positives and false negatives are equally weighted. This may be an undesirable characteristic for some clustering applications. The F-measure addresses this concern, as does the chance-corrected adjusted Rand index.

- F-measure

The F-measure can be used to balance the contribution of false negatives by weighting recall through a parameter $\beta \geq 0$. Let precision and recall be defined as follows:

$$P = \frac{TP}{TP + FP}$$

$$R = \frac{TP}{TP + FN}$$

where P is the precision rate and R is the recall rate. We can calculate the F-measure by using the following formula:

$$F_{\beta} = \frac{(\beta^2 + 1) \cdot P \cdot R}{\beta^2 \cdot P + R}$$

Notice that when $\beta = 0$, $F_0 = P$. In other words, recall has no impact on the F-measure when $\beta = 0$, and increasing β allocates an increasing amount of weight to recall in the final F-measure.

- Jaccard index

The Jaccard index is used to quantify the similarity between two datasets. The Jaccard index takes on a value between 0 and 1. An index of 1 means that the two dataset are identical, and an index of 0 indicates that the datasets have no common elements. The Jaccard index is defined by the following formula:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{TP}{TP + FP + FN}$$

This is simply the number of unique elements common to both sets divided by the total number of unique elements in both sets.

- Fowlkes–Mallows index (E. B. Fowlkes & C. L. Mallows 1983)

The Fowlkes-Mallows index computes the similarity between the clusters returned by the clustering algorithm and the benchmark classifications. The higher the value of the Fowlkes-Mallows index the more similar the clusters and the benchmark classifications are. It can be computed using the following formula:

$$FM = \sqrt{\frac{TP}{TP + FP} \cdot \frac{TP}{TP + FN}}$$

where TP is the number of true positives, FP is the number of false positives, and FN is the number of false negatives. The FM index is the geometric mean of the precision and recall P and R, while the F-measure is their harmonic mean. Moreover, precision and recall are also known as Wallace's indices B^I and B^{II}.

- The Mutual Information is an information theoretic measure of how much information is shared between a clustering and a ground-truth classification that can detect a non-linear similarity between two clusterings. Adjusted mutual information is the corrected-for-chance variant of this that has a reduced bias for varying cluster numbers.
- Confusion matrix

  A confusion matrix can be used to quickly visualize the results of a classification (or clustering) algorithm. It shows how different a cluster is from the gold standard cluster.

Applications

Biology, computational biology and bioinformatics

Plant and animal ecology

cluster analysis is used to describe and to make spatial and temporal comparisons of communities (assemblages) of organisms in heterogeneous environments; it is also used in plant systematics to generate artificial phylogenies or clusters of organisms (individuals) at the species, genus or higher level that share a number of attributes

Transcriptomics

clustering is used to build groups of genes with related expression patterns (also known as coexpressed genes) as in HCS clustering algorithm . Often such groups contain functionally related proteins, such as enzymes for a specific pathway, or genes that are co-regulated. High throughput experiments using expressed sequence tags (ESTs) or DNA microarrays can be a powerful tool for genome annotation, a general aspect of genomics.

Sequence analysis

clustering is used to group homologous sequences into gene families. This is a very important concept in bioinformatics, and evolutionary biology in general.

High-throughput genotyping platforms

clustering algorithms are used to automatically assign genotypes.

Human genetic clustering

The similarity of genetic data is used in clustering to infer population structures.

Medicine

Medical imaging

On PET scans, cluster analysis can be used to differentiate between different types of tissue in a three-dimensional image for many different purposes.

Business and marketing

Market research

Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers, and for use in market segmentation, Product positioning, New product development and Selecting test markets.

Grouping of shopping items

Clustering can be used to group all the shopping items available on the web into a set of unique products. For example, all the items on eBay can be grouped into unique products. (eBay doesn't have the concept of a SKU)

World wide web

Social network analysis

In the study of social networks, clustering may be used to recognize communities within large groups of people.

Search result grouping

In the process of intelligent grouping of the files and websites, clustering may be used to create a more relevant set of search results compared to normal search engines like Google. There are currently a number of web based clustering tools such as Clusty.

Slippy map optimization

Flickr's map of photos and other map sites use clustering to reduce the number of markers on a map. This makes it both faster and reduces the amount of visual clutter.

Computer science

Software evolution

Clustering is useful in software evolution as it helps to reduce legacy properties in code by reforming functionality that has become dispersed. It is a form of restructuring and hence is a way of direct preventative maintenance.

Image segmentation

Clustering can be used to divide a digital image into distinct regions for border detection or object recognition.

Evolutionary algorithms

Clustering may be used to identify different niches within the population of an evolutionary algorithm so that reproductive opportunity can be distributed more evenly amongst the evolving species or subspecies.

Recommender systems

Recommender systems are designed to recommend new items based on a user's tastes. They sometimes use clustering algorithms to predict a user's preferences based on the preferences of other users in the user's cluster.

Markov chain Monte Carlo methods

Clustering is often utilized to locate and characterize extrema in the target distribution.

Anomaly detection

Anomalies/outliers are typically - be it explicitly or implicitly - defined with respect to clustering structure in data.

Social science

Crime analysis

Cluster analysis can be used to identify areas where there are greater incidences of particular types of crime. By identifying these distinct areas or "hot spots" where a similar crime has happened over a period of time, it is possible to manage law enforcement resources more effectively.

Educational data mining

Cluster analysis is for example used to identify groups of schools or students with similar properties.

Typologies

From poll data, projects such as those undertaken by the Pew Research Center use cluster analysis to discern typologies of opinions, habits, and demographics that may be useful in politics and marketing.

Others

Field robotics

Clustering algorithms are used for robotic situational awareness to track objects and detect outliers in sensor data.

Mathematical chemistry

To find structural similarity, etc., for example, 3000 chemical compounds were clustered in the space of 90 topological indices.

Climatology

To find weather regimes or preferred sea level pressure atmospheric patterns.

Petroleum geology

Cluster analysis is used to reconstruct missing bottom hole core data or missing log curves in order to evaluate reservoir properties.

Physical geography

The clustering of chemical properties in different sample locations.

Statistical Classification

In machine learning and statistics, classification is the problem of identifying to which of a set of categories (sub-populations) a new observation belongs, on the basis of a training set of data containing observations (or instances) whose category membership is known.

An example would be assigning a given email into "spam" or "non-spam" classes or assigning a diagnosis to a given patient as described by observed characteristics of the patient (gender, blood pressure, presence or absence of certain symptoms, etc.). Classification is an example of pattern recognition.

In the terminology of machine learning, classification is considered an instance of supervised learning, i.e. learning where a training set of correctly identified observations is available. The corresponding unsupervised procedure is known as clustering, and involves grouping data into categories based on some measure of inherent similarity or distance.

Often, the individual observations are analyzed into a set of quantifiable properties, known variously as explanatory variables or *features*. These properties may variously be categorical (e.g. "A", "B", "AB" or "O", for blood type), ordinal (e.g. "large", "medium" or "small"), integer-valued (e.g. the number of occurrences of a particular word in an email) or real-valued (e.g. a measurement of blood pressure). Other classifiers work by comparing observations to previous observations by means of a similarity or distance function.

An algorithm that implements classification, especially in a concrete implementation, is known as a classifier. The term "classifier" sometimes also refers to the mathematical function, implemented by a classification algorithm, that maps input data to a category.

Terminology across fields is quite varied. In statistics, where classification is often done with logistic regression or a similar procedure, the properties of observations are termed explanatory variables (or independent variables, regressors, etc.), and the categories to be predicted are known as outcomes, which are considered to be possible values of the dependent variable. In machine learning, the observations are often known as *instances*, the explanatory variables are termed *features* (grouped into a feature vector), and the possible categories to be predicted are *classes*. Other fields may use different terminology: e.g. in community ecology, the term "classification" normally refers to cluster analysis, i.e. a type of unsupervised learning, rather than the supervised learning described in this article.

Relation to Other Problems

Classification and clustering are examples of the more general problem of pattern recognition, which is the assignment of some sort of output value to a given input value. Other examples are regression, which assigns a real-valued output to each input; sequence labeling, which assigns a class to each member of a sequence of values (for example, part of speech tagging, which assigns a part of speech to each word in an input sentence); parsing, which assigns a parse tree to an input sentence, describing the syntactic structure of the sentence; etc.

A common subclass of classification is probabilistic classification. Algorithms of this nature use statistical inference to find the best class for a given instance. Unlike other algorithms, which simply output a "best" class, probabilistic algorithms output a probability of the instance being a member of each of the possible classes. The best class is normally then selected as the one with the highest probability. However, such an algorithm has numerous advantages over non-probabilistic classifiers:

- It can output a confidence value associated with its choice (in general, a classifier that can do this is known as a *confidence-weighted classifier*).
- Correspondingly, it can *abstain* when its confidence of choosing any particular output is too low.
- Because of the probabilities which are generated, probabilistic classifiers can be more effectively incorporated into larger machine-learning tasks, in a way that partially or completely avoids the problem of *error propagation*.

Frequentist Procedures

Early work on statistical classification was undertaken by Fisher, in the context of two-group problems, leading to Fisher's linear discriminant function as the rule for assigning a group to a new observation. This early work assumed that data-values within each of the two groups had a multivariate normal distribution. The extension of this same context to more than two-groups has also been considered with a restriction imposed that the classification rule should be linear. Later work for the multivariate normal distribution allowed the classifier to be nonlinear: several classification rules can be derived based on slight different adjustments of the Mahalanobis distance, with a new observation being assigned to the group whose centre has the lowest adjusted distance from the observation.

Bayesian Procedures

Unlike frequentist procedures, Bayesian classification procedures provide a natural way of taking into account any available information about the relative sizes of the sub-populations associated with the different groups within the overall population. Bayesian procedures tend to be computationally expensive and, in the days before Markov chain Monte Carlo computations were developed, approximations for Bayesian clustering rules were devised.

Some Bayesian procedures involve the calculation of group membership probabilities: these can be viewed as providing a more informative outcome of a data analysis than a simple attribution of a single group-label to each new observation.

Binary and Multiclass Classification

Classification can be thought of as two separate problems – binary classification and multiclass classification. In binary classification, a better understood task, only two classes are involved, whereas multiclass classification involves assigning an object to one of several classes. Since many classification methods have been developed specifically for binary classification, multiclass classification often requires the combined use of multiple binary classifiers.

Feature Vectors

Most algorithms describe an individual instance whose category is to be predicted using a feature vector of individual, measurable properties of the instance. Each property is termed a feature, also known in statistics as an explanatory variable (or independent variable, although features may or may not be statistically independent). Features may variously be binary (e.g. “male” or “female”); categorical (e.g. “A”, “B”, “AB” or “O”, for blood type); ordinal (e.g. “large”, “medium” or “small”); integer-valued (e.g. the number of occurrences of a particular word in an email); or real-valued (e.g. a measurement of blood pressure). If the instance is an image, the feature values might correspond to the pixels of an image; if the instance is a piece of text, the feature values might be occurrence frequencies of different words. Some algorithms work only in terms of discrete data and require that real-valued or integer-valued data be *discretized* into groups (e.g. less than 5, between 5 and 10, or greater than 10)

Linear Classifiers

A large number of algorithms for classification can be phrased in terms of a linear function that assigns a score to each possible category *k* by combining the feature vector of an instance with a vector of weights, using a dot product. The predicted category is the one with the highest score. This type of score function is known as a linear predictor function and has the following general form:

$$score(\mathbf{X}_i, k) = \beta_k \cdot \mathbf{X}_i,$$

where X_i is the feature vector for instance *i*, β_k is the vector of weights corresponding to category *k*, and score(X_i, *k*) is the score associated with assigning instance *i* to category *k*. In discrete choice theory, where instances represent people and categories represent choices, the score is considered the utility associated with person *i* choosing category *k*.

Algorithms with this basic setup are known as linear classifiers. What distinguishes them is the procedure for determining (training) the optimal weights/coefficients and the way that the score is interpreted.

Examples of such algorithms are

- Logistic regression and Multinomial logistic regression
- Probit regression
- The perceptron algorithm
- Support vector machines
- Linear discriminant analysis.

Algorithms

Examples of classification algorithms include:

- Linear classifiers
    - Fisher's linear discriminant
    - Logistic regression
    - Naive Bayes classifier
    - Perceptron
- Support vector machines
    - Least squares support vector machines
- Quadratic classifiers
- Kernel estimation
    - k-nearest neighbor
- Boosting (meta-algorithm)
- Decision trees
    - Random forests
- Neural networks
- FMM Neural Networks
- Learning vector quantization

Evaluation

Classifier performance depends greatly on the characteristics of the data to be classified. There is no single classifier that works best on all given problems (a phenomenon that may be explained by the no-free-lunch theorem). Various empirical tests have been performed to compare classifier performance and to find the characteristics of data that determine classifier performance. Determining a suitable classifier for a given problem is however still more an art than a science.

The measures precision and recall are popular metrics used to evaluate the quality of a classification system. More recently, receiver operating characteristic (ROC) curves have been used to evaluate the tradeoff between true- and false-positive rates of classification algorithms.

As a performance metric, the uncertainty coefficient has the advantage over simple accuracy in that it is not affected by the relative sizes of the different classes. Further, it will not penalize an algorithm for simply *rearranging* the classes.

Application Domains

Classification has many applications. In some of these it is employed as a data mining procedure, while in others more detailed statistical modeling is undertaken.

- Computer vision
    - Medical imaging and medical image analysis
    - Optical character recognition
    - Video tracking
- Drug discovery and development
    - Toxicogenomics
    - Quantitative structure-activity relationship
- Geostatistics
- Speech recognition
- Handwriting recognition
- Biometric identification
- Biological classification
- Statistical natural language processing
- Document classification
- Internet search engines
- Credit scoring
- Pattern recognition
- Micro-array classification

Regression Analysis

In statistical modeling, regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average value of the dependent variable when the independent variables are fixed. Less commonly, the focus is on a quantile, or other location parameter of the conditional distribution of the dependent variable given the independent variables. In all cases, the estimation target is a function of the independent variables called the regression function. In regression analysis, it is also of interest to characterize the variation of the dependent variable around the regression function which can be described by a probability distribution. A related but distinct approach is necessary condition analysis (NCA), which estimates the maximum (rather than average) value of the dependent variable for a given value of the independent variable (ceiling line rather than central line) in order to identify what value of the independent variable is necessary but not sufficient for a given value of the dependent variable.

Regression analysis is widely used for prediction and forecasting, where its use has substantial overlap with the field of machine learning. Regression analysis is also used to understand which among the independent variables are related to the dependent variable, and to explore the forms of these relationships. In restricted circumstances, regression analysis can be used to infer causal relationships between the independent and dependent variables. However this can lead to illusions or false relationships, so caution is advisable; for example, correlation does not imply causation.

Many techniques for carrying out regression analysis have been developed. Familiar methods such as linear regression and ordinary least squares regression are parametric, in that the regression function is defined in terms of a finite number of unknown parameters that are estimated from the data. Nonparametric regression refers to techniques that allow the regression function to lie in a specified set of functions, which may be infinite-dimensional.

The performance of regression analysis methods in practice depends on the form of the data generating process, and how it relates to the regression approach being used. Since the true form of the data-generating process is generally not known, regression analysis often depends to some extent on making assumptions about this process. These assumptions are sometimes testable if a sufficient quantity of data is available. Regression models for prediction are often useful even when the assumptions are moderately violated, although they may not perform optimally. However, in many applications, especially with small effects or questions of causality based on observational data, regression methods can give misleading results.

In a narrower sense, regression may refer specifically to the estimation of continuous response variables, as opposed to the discrete response variables used in classification. The case of a continuous output variable may be more specifically referred to as metric regression to distinguish it from related problems.

History

The earliest form of regression was the method of least squares, which was published by Legendre in 1805, and by Gauss in 1809. Legendre and Gauss both applied the method to the problem of determining, from astronomical observations, the orbits of bodies about the Sun (mostly comets, but also later the then newly discovered minor planets). Gauss published a further development of the theory of least squares in 1821, including a version of the Gauss–Markov theorem.

The term “regression” was coined by Francis Galton in the nineteenth century to describe a biological phenomenon. The phenomenon was that the heights of descendants of tall ancestors tend to regress down towards a normal average (a phenomenon also known as regression toward the mean). For Galton, regression had only this biological meaning, but his work was later extended by Udny Yule and Karl Pearson to a more general statistical context. In the work of Yule and Pearson, the joint distribution of the response and explanatory variables is assumed to be Gaussian. This assumption was weakened by R.A. Fisher in his works of 1922 and 1925. Fisher assumed that the conditional distribution of the response variable is Gaussian, but the joint distribution need not be. In this respect, Fisher's assumption is closer to Gauss's formulation of 1821.

In the 1950s and 1960s, economists used electromechanical desk calculators to calculate regressions. Before 1970, it sometimes took up to 24 hours to receive the result from one regression.

Regression methods continue to be an area of active research. In recent decades, new methods have been developed for robust regression, regression involving correlated responses such as time series and growth curves, regression in which the predictor (independent variable) or response variables are curves, images, graphs, or other complex data objects, regression methods accommodating various types of missing data, nonparametric regression, Bayesian methods for regression, regression in which the predictor variables are measured with error, regression with more predictor variables than observations, and causal inference with regression.

Regression Models

Regression models involve the following variables:

- The unknown parameters, denoted as β, which may represent a scalar or a vector.
- The independent variables, X.
- The dependent variable, Y.

In various fields of application, different terminologies are used in place of dependent and independent variables.

A regression model relates Y to a function of X and β.

$$Y \approx f(X, \beta)$$

The approximation is usually formalized as $E(Y \mid X) = f(X, \beta)$. To carry out regression analysis, the form of the function f must be specified. Sometimes the form of this function is based on knowledge about the relationship between Y and X that does not rely on the data. If no such knowledge is available, a flexible or convenient form for f is chosen.

Assume now that the vector of unknown parameters β is of length k. In order to perform a regression analysis the user must provide information about the dependent variable Y:

- If N data points of the form (Y, X) are observed, where $N < k$, most classical approaches to regression analysis cannot be performed: since the system of equations defining the regression model is underdetermined, there are not enough data to recover β.
- If exactly $N = k$ data points are observed, and the function f is linear, the equations $Y = f(X, \beta)$ can be solved exactly rather than approximately. This reduces to solving a set of N equations with N unknowns (the elements of β), which has a unique solution as long as the X are linearly independent. If f is nonlinear, a solution may not exist, or many solutions may exist.
- The most common situation is where $N > k$ data points are observed. In this case, there is

enough information in the data to estimate a unique value for β that best fits the data in some sense, and the regression model when applied to the data can be viewed as an overdetermined system in β.

In the last case, the regression analysis provides the tools for:

1. Finding a solution for unknown parameters β that will, for example, minimize the distance between the measured and predicted values of the dependent variable Y (also known as method of least squares).
2. Under certain statistical assumptions, the regression analysis uses the surplus of information to provide statistical information about the unknown parameters β and predicted values of the dependent variable Y.

Necessary Number of Independent Measurements

Consider a regression model which has three unknown parameters, β_0, β_1, and β_2. Suppose an experimenter performs 10 measurements all at exactly the same value of independent variable vector X (which contains the independent variables X_1, X_2, and X_3). In this case, regression analysis fails to give a unique set of estimated values for the three unknown parameters; the experimenter did not provide enough information. The best one can do is to estimate the average value and the standard deviation of the dependent variable Y. Similarly, measuring at two different values of X would give enough data for a regression with two unknowns, but not for three or more unknowns.

If the experimenter had performed measurements at three different values of the independent variable vector X, then regression analysis would provide a unique set of estimates for the three unknown parameters in β.

In the case of general linear regression, the above statement is equivalent to the requirement that the matrix X^TX is invertible.

Statistical Assumptions

When the number of measurements, N, is larger than the number of unknown parameters, k, and the measurement errors ε_i are normally distributed then *the excess of information* contained in $(N - k)$ measurements is used to make statistical predictions about the unknown parameters. This excess of information is referred to as the degrees of freedom of the regression.

Underlying Assumptions

Classical assumptions for regression analysis include:

- The sample is representative of the population for the inference prediction.
- The error is a random variable with a mean of zero conditional on the explanatory variables.
- The independent variables are measured with no error. (Note: If this is not so, modeling may be done instead using errors-in-variables model techniques).

- The independent variables (predictors) are linearly independent, i.e. it is not possible to express any predictor as a linear combination of the others.
- The errors are uncorrelated, that is, the variance–covariance matrix of the errors is diagonal and each non-zero element is the variance of the error.
- The variance of the error is constant across observations (homoscedasticity). If not, weighted least squares or other methods might instead be used.

These are sufficient conditions for the least-squares estimator to possess desirable properties; in particular, these assumptions imply that the parameter estimates will be unbiased, consistent, and efficient in the class of linear unbiased estimators. It is important to note that actual data rarely satisfies the assumptions. That is, the method is used even though the assumptions are not true. Variation from the assumptions can sometimes be used as a measure of how far the model is from being useful. Many of these assumptions may be relaxed in more advanced treatments. Reports of statistical analyses usually include analyses of tests on the sample data and methodology for the fit and usefulness of the model.

Assumptions include the geometrical support of the variables. Independent and dependent variables often refer to values measured at point locations. There may be spatial trends and spatial autocorrelation in the variables that violate statistical assumptions of regression. Geographic weighted regression is one technique to deal with such data. Also, variables may include values aggregated by areas. With aggregated data the modifiable areal unit problem can cause extreme variation in regression parameters. When analyzing data aggregated by political boundaries, postal codes or census areas results may be very distinct with a different choice of units.

Linear Regression

In linear regression, the model specification is that the dependent variable, y_i is a linear combination of the *parameters* (but need not be linear in the *independent variables*). For example, in simple linear regression for modeling n data points there is one independent variable: x_i, and two parameters, β_0 and β_1 :

$$\text{straight line: } y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \quad i = 1,\ldots,n.$$

In multiple linear regression, there are several independent variables or functions of independent variables.

Adding a term in x_i^2 to the preceding regression gives:

$$\text{parabola: } y_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \varepsilon_i,\ i = 1,\ldots,n.$$

This is still linear regression; although the expression on the right hand side is quadratic in the independent variable x_i, it is linear in the parameters β_0, β_1 and β_2.

In both cases, ε_i is an error term and the subscript i indexes a particular observation.

Returning our attention to the straight line case: Given a random sample from the population, we estimate the population parameters and obtain the sample linear regression model:

$$\widehat{y_i} = \hat{\beta}_0 + \hat{\beta}_1 x_i.$$

The residual, $e_i = y_i - \hat{y}_i$, is the difference between the value of the dependent variable predicted by the model, $\widehat{y}_i$,, and the true value of the dependent variable, y_i. One method of estimation is ordinary least squares. This method obtains parameter estimates that minimize the sum of squared residuals, SSE, also sometimes denoted RSS:

$$SSE = \sum_{i=1}^{n} e_i^2.$$

Minimization of this function results in a set of normal equations, a set of simultaneous linear equations in the parameters, which are solved to yield the parameter estimators, $\widehat{\beta_0}$, $\widehat{\beta_1}$

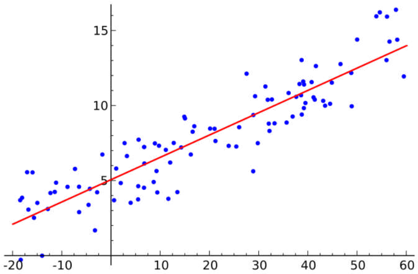


Illustration of linear regression on a data set.

In the case of simple regression, the formulas for the least squares estimates are

$$\widehat{\beta_1} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sum (x_i - \bar{x})^2} \text{ and } \widehat{\beta_0} = \bar{y} - \widehat{\beta_1}\bar{x}$$

where $\bar{x}$ is the mean (average) of the x values and $\bar{y}$ is the mean of the y values.

Under the assumption that the population error term has a constant variance, the estimate of that variance is given by:

$$\hat{\sigma}_\varepsilon^2 = \frac{SSE}{n-2}.$$

This is called the mean square error (MSE) of the regression. The denominator is the sample size reduced by the number of model parameters estimated from the same data, (n-p) for p regressors or (n-p-1) if an intercept is used. In this case, p=1 so the denominator is n-2.

The standard errors of the parameter estimates are given by

$$\hat{\sigma}_{\beta_0} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum (x_i - \bar{x})^2}}$$

$$\hat{\sigma}_{\beta_1} = \hat{\sigma}_\varepsilon \sqrt{\frac{1}{\sum (x_i - \bar{x})^2}}.$$

Under the further assumption that the population error term is normally distributed, the researcher can use these estimated standard errors to create confidence intervals and conduct hypothesis tests about the population parameters.

General Linear Model

In the more general multiple regression model, there are p independent variables:

$$y_i = \beta_1 x_{i1} + \beta_2 x_{i2} + \cdots + \beta_p x_{ip} + \varepsilon_i,$$

where x_{ij} is the i^{th} observation on the j^{th} independent variable. If the first independent variable takes the value 1 for all i, $x_{i1} = 1$, then β_1 is called the regression intercept.

The least squares parameter estimates are obtained from p normal equations. The residual can be written as

$$\varepsilon_i = y_i - \hat{\beta}_1 x_{i1} - \cdots - \hat{\beta}_p x_{ip}.$$

The normal equations are

$$\sum_{i=1}^{n} \sum_{k=1}^{p} X_{ij} X_{ik} \hat{\beta}_k = \sum_{i=1}^{n} X_{ij} y_i,\ j = 1, \ldots, p.$$

In matrix notation, the normal equations are written as

$$(X^\top X)\hat{\beta} = X^\top Y,$$

where the ij element of X is x_{ij}, the i element of the column vector Y is y_i, and the j element of $\hat{\beta}$ is $\hat{\beta}_j$. Thus X is $n \times p$, Y is $n \times 1$, and $\hat{\beta}$ is $p \times 1$. The solution is

$$\hat{\beta} = (X^\top X)^{-1} X^\top Y.$$

Diagnostics

Once a regression model has been constructed, it may be important to confirm the goodness of fit of the model and the statistical significance of the estimated parameters. Commonly used checks of goodness of fit include the R-squared, analyses of the pattern of residuals and hypothesis testing. Statistical significance can be checked by an F-test of the overall fit, followed by t-tests of individual parameters.

Interpretations of these diagnostic tests rest heavily on the model assumptions. Although examination of the residuals can be used to invalidate a model, the results of a t-test or F-test are sometimes more difficult to interpret if the model's assumptions are violated. For example, if the error term does not have a normal distribution, in small samples the estimated parameters will not follow normal distributions and complicate inference. With relatively large samples, however, a central limit theorem can be invoked such that hypothesis testing may proceed using asymptotic approximations.

“Limited Dependent” Variables

The phrase “limited dependent” is used in econometric statistics for categorical and constrained variables.

The response variable may be non-continuous (“limited” to lie on some subset of the real line). For binary (zero or one) variables, if analysis proceeds with least-squares linear regression, the model is called the linear probability model. Nonlinear models for binary dependent variables include the probit and logit model. The multivariate probit model is a standard method of estimating a joint relationship between several binary dependent variables and some independent variables. For categorical variables with more than two values there is the multinomial logit. For ordinal variables with more than two values, there are the ordered logit and ordered probit models. Censored regression models may be used when the dependent variable is only sometimes observed, and Heckman correction type models may be used when the sample is not randomly selected from the population of interest. An alternative to such procedures is linear regression based on polychoric correlation (or polyserial correlations) between the categorical variables. Such procedures differ in the assumptions made about the distribution of the variables in the population. If the variable is positive with low values and represents the repetition of the occurrence of an event, then count models like the Poisson regression or the negative binomial model may be used instead.

Interpolation and Extrapolation

Regression models predict a value of the Y variable given known values of the X variables. Prediction *within* the range of values in the dataset used for model-fitting is known informally as interpolation. Prediction *outside* this range of the data is known as extrapolation. Performing extrapolation relies strongly on the regression assumptions. The further the extrapolation goes outside the data, the more room there is for the model to fail due to differences between the assumptions and the sample data or the true values.

It is generally advised that when performing extrapolation, one should accompany the estimated value of the dependent variable with a prediction interval that represents the uncertainty. Such intervals tend to expand rapidly as the values of the independent variable(s) moved outside the range covered by the observed data.

For such reasons and others, some tend to say that it might be unwise to undertake extrapolation.

However, this does not cover the full set of modelling errors that may be being made: in particular, the assumption of a particular form for the relation between Y and X. A properly conducted regression analysis will include an assessment of how well the assumed form is matched by the observed data, but it can only do so within the range of values of the independent variables actually available. This means that any extrapolation is particularly reliant on the assumptions being made about the structural form of the regression relationship. Best-practice advice here is that a linear-in-variables and linear-in-parameters relationship should not be chosen simply for computational convenience, but that all available knowledge should be deployed in constructing a regression model. If this knowledge includes the fact that the dependent variable cannot go outside a certain range of values, this can be made use of in selecting the model – even if the observed dataset has no values particularly near such bounds. The implications of this step of choosing an appropriate functional form for the regression can be great when extrapolation is considered. At a minimum, it can ensure that any extrapolation arising from a fitted model is "realistic" (or in accord with what is known).

Nonlinear Regression

When the model function is not linear in the parameters, the sum of squares must be minimized by an iterative procedure. This introduces many complications which are summarized in Differences between linear and non-linear least squares

Power and Sample Size Calculations

There are no generally agreed methods for relating the number of observations versus the number of independent variables in the model. One rule of thumb suggested by Good and Hardin is $N = m^n$, , where N is the sample size, n is the number of independent variables and m is the number of observations needed to reach the desired precision if the model had only one independent variable. For example, a researcher is building a linear regression model using a dataset that contains 1000 patients (N). If the researcher decides that five observations are needed to precisely define a straight line (m), then the maximum number of independent variables the model can support is 4, because

$$\frac{\log 1000}{\log 5} = 4.29.$$

Other Methods

Although the parameters of a regression model are usually estimated using the method of least squares, other methods which have been used include:

- Bayesian methods, e.g. Bayesian linear regression
- Percentage regression, for situations where reducing *percentage* errors is deemed more appropriate.
- Least absolute deviations, which is more robust in the presence of outliers, leading to quantile regression
- Nonparametric regression, requires a large number of observations and is computationally intensive
- Distance metric learning, which is learned by the search of a meaningful distance metric in a given input space.

Software

All major statistical software packages perform least squares regression analysis and inference. Simple linear regression and multiple regression using least squares can be done in some spreadsheet applications and on some calculators. While many statistical software packages can perform various types of nonparametric and robust regression, these methods are less standardized; different software packages implement different methods, and a method with a given name may be implemented differently in different packages. Specialized regression software has been developed for use in fields such as survey analysis and neuroimaging.

Automatic Summarization

Automatic summarization is the process of reducing a text document with a computer program in order to create a summary that retains the most important points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a representative subset of the data, which contains the *information* of the entire set. Summarization technologies are used in a large number of sectors in industry today. An example of the use of summarization technology is search engines such as Google. Other examples include document summarization, image collection summarization and video summarization. Document summarization, tries to automatically create a *representative summary* or *abstract* of the entire document, by finding the most *informative* sentences. Similarly, in image summarization the system finds the most representative and important (or salient) images. Similarly, in consumer videos one would want to remove the boring or repetitive scenes, and extract out a much shorter and concise version of the video. This is also important, say for surveillance videos, where one might want to extract only important events in the recorded video, since most part of the video may be uninteresting with nothing going on. As the problem of information overload grows, and as the amount of data increases, the interest in automatic summarization is also increasing.

Generally, there are two approaches to automatic summarization: *extraction* and *abstraction*. Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. In contrast, abstractive methods build an internal semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. Research into abstractive methods is an increasingly important and active research area, however due to complexity constraints, research to date has focused primarily on extractive methods. In some application domains, extractive summarization makes more sense. Examples of these include image collection summarization and video summarization.

Extraction-based Summarization

In this summarization task, the automatic system extracts objects from the entire collection, without modifying the objects themselves. Examples of this include keyphrase extraction, where the goal is to select individual words or phrases to “tag” a document, and document summarization, where the goal is to select whole sentences (without modifying them) to create a short paragraph summary. Similarly, in image collection summarization, the system extracts images from the collection without modifying the images themselves.

Abstraction-based Summarization

Extraction techniques merely copy the information deemed most important by the system to the summary (for example, key clauses, sentences or paragraphs), while abstraction involves paraphrasing sections of the source document. In general, abstraction can condense a text more strongly than extraction, but the programs that can do this are harder to develop as they require use of natural language generation technology, which itself is a growing field.

While some work has been done in abstractive summarization (creating an abstract synopsis like that of a human), the majority of summarization systems are extractive (selecting a subset of sentences to place in a summary).

Aided Summarization

Machine learning techniques from closely related fields such as information retrieval or text mining have been successfully adapted to help automatic summarization.

Apart from Fully Automated Summarizers (FAS), there are systems that aid users with the task of summarization (MAHS = Machine Aided Human Summarization), for example by highlighting candidate passages to be included in the summary, and there are systems that depend on post-processing by a human (HAMS = Human Aided Machine Summarization).

Applications and Systems for Summarization

There are broadly two types of extractive summarization tasks depending on what the summarization program focuses on. The first is *generic summarization*, which focuses on obtaining a generic summary or abstract of the collection (whether documents, or sets of images, or videos, news stories etc.). The second is *query relevant summarization*, sometimes called *query-based summarization*, which summarizes objects specific to a query. Summarization systems are able to create both query relevant text summaries and generic machine-generated summaries depending on what the user needs.

An example of a summarization problem is document summarization, which attempts to automatically produce an abstract from a given document. Sometimes one might be interested in generating a summary from a single source document, while others can use multiple source documents (for example, a cluster of articles on the same topic). This problem is called multi-document summarization. A related application is summarizing news articles. Imagine a system, which automatically pulls together news articles on a given topic (from the web), and concisely represents the latest news as a summary.

Image collection summarization is another application example of automatic summarization. It consists in selecting a representative set of images from a larger set of images. A summary in this context is useful to show the most representative images of results in an image collection exploration system. Video summarization is a related domain, where the system automatically creates a trailer of a long video. This also has applications in consumer or personal videos, where one might want to skip the boring or repetitive actions. Similarly, in surveillance videos, one would want to extract important and suspicious activity, while ignoring all the boring and redundant frames captured.

At a very high level, summarization algorithms try to find subsets of objects (like set of sentences, or a set of images), which cover information of the entire set. This is also called the *core-set*. These algorithms model notions like diversity, coverage, information and representativeness of the summary. Query based summarization techniques, additionally model for relevance of the summary with the query. Some techniques and algorithms which naturally model summarization problems are TextRank and PageRank, Submodular set function, Determinantal point process, maximal marginal relevance (MMR) etc.

Keyphrase Extraction

The task is the following. You are given a piece of text, such as a journal article, and you must produce a list of keywords or key[phrase]s that capture the primary topics discussed in the text. In the case of research articles, many authors provide manually assigned keywords, but most text lacks pre-existing keyphrases. For example, news articles rarely have keyphrases attached, but it would be useful to be able to automatically do so for a number of applications discussed below. Consider the example text from a news article:

> "The Army Corps of Engineers, rushing to meet President Bush's promise to protect New Orleans by the start of the 2006 hurricane season, installed defective flood-control pumps last year despite warnings from its own expert that the equipment would fail during a storm, according to documents obtained by The Associated Press".

A keyphrase extractor might select "Army Corps of Engineers", "President Bush", "New Orleans", and "defective flood-control pumps" as keyphrases. These are pulled directly from the text. In contrast, an abstractive keyphrase system would somehow internalize the content and generate keyphrases that do not appear in the text, but more closely resemble what a human might produce, such as "political negligence" or "inadequate protection from floods". Abstraction requires a deep understanding of the text, which makes it difficult for a computer system. Keyphrases have many applications. They can enable document browsing by providing a short summary, improve information retrieval (if documents have keyphrases assigned, a user could search by keyphrase to produce more reliable hits than a full-text search), and be employed in generating index entries for a large text corpus.

Depending on the different literature and the definition of key terms, words or phrases, highly related theme is certainly the Keyword extraction.

Supervised Learning Approaches

Beginning with the work of Turney, many researchers have approached keyphrase extraction as a supervised machine learning problem. Given a document, we construct an example for each unigram, bigram, and trigram found in the text (though other text units are also possible, as discussed below). We then compute various features describing each example (e.g., does the phrase begin with an upper-case letter?). We assume there are known keyphrases available for a set of training documents. Using the known keyphrases, we can assign positive or negative labels to the examples. Then we learn a classifier that can discriminate between positive and negative examples as a function of the features. Some classifiers make a binary classification for a test example, while others assign a probability of being a keyphrase. For instance, in the above text, we might learn a rule that says phrases with initial capital letters are likely to be keyphrases. After training a learner, we can select keyphrases for test documents in the following manner. We apply the same example-generation strategy to the test documents, then run each example through the learner. We can determine the keyphrases by looking at binary classification decisions or probabilities returned from our learned model. If probabilities are given, a threshold is used to select the keyphrases. Keyphrase extractors are generally evaluated using precision and recall. Precision measures how many of the proposed keyphrases are actually correct. Recall measures how many of the true keyphrases your system proposed. The two measures can be combined in an F-score, which is the harmonic mean of the two ($F = 2PR/(P + R)$). Matches between the proposed keyphrases and the known keyphrases can be checked after stemming or applying some other text normalization.

Designing a supervised keyphrase extraction system involves deciding on several choices (some of these apply to unsupervised, too). The first choice is exactly how to generate examples. Turney and others have used all possible unigrams, bigrams, and trigrams without intervening punctuation and after removing stopwords. Hulth showed that you can get some improvement by selecting examples to be sequences of tokens that match certain patterns of part-of-speech tags. Ideally, the mechanism for generating examples produces all the known labeled keyphrases as candidates, though this is often not the case. For example, if we use only unigrams, bigrams, and trigrams, then we will never be able to extract a known keyphrase containing four words. Thus, recall may suffer. However, generating too many examples can also lead to low precision.

We also need to create features that describe the examples and are informative enough to allow a learning algorithm to discriminate keyphrases from non- keyphrases. Typically features involve various term frequencies (how many times a phrase appears in the current text or in a larger corpus), the length of the example, relative position of the first occurrence, various boolean syntactic features (e.g., contains all caps), etc. The Turney paper used about 12 such features. Hulth uses a reduced set of features, which were found most successful in the KEA (Keyphrase Extraction Algorithm) work derived from Turney's seminal paper.

In the end, the system will need to return a list of keyphrases for a test document, so we need to have a way to limit the number. Ensemble methods (i.e., using votes from several classifiers) have been used to produce numeric scores that can be thresholded to provide a user-provided number of keyphrases. This is the technique used by Turney with C4.5 decision trees. Hulth used a single binary classifier so the learning algorithm implicitly determines the appropriate number.

Once examples and features are created, we need a way to learn to predict keyphrases. Virtually any supervised learning algorithm could be used, such as decision trees, Naive Bayes, and rule induction. In the case of Turney's GenEx algorithm, a genetic algorithm is used to learn parameters for a domain-specific keyphrase extraction algorithm. The extractor follows a series of heuristics to identify keyphrases. The genetic algorithm optimizes parameters for these heuristics with respect to performance on training documents with known key phrases.

Unsupervised Approach: Textrank

Another keyphrase extraction algorithm is TextRank. While supervised methods have some nice properties, like being able to produce interpretable rules for what features characterize a keyphrase, they also require a large amount of training data. Many documents with known keyphrases are needed. Furthermore, training on a specific domain tends to customize the extraction process to that domain, so the resulting classifier is not necessarily portable, as some of Turney's results demonstrate. Unsupervised keyphrase extraction removes the need for training data. It approaches the problem from a different angle. Instead of trying to learn explicit features that characterize keyphrases, the TextRank algorithm exploits the structure of the text itself to determine keyphrases that appear "central" to the text in the same way that PageRank selects important Web pages. Recall this is based on the notion of "prestige" or "recommendation" from social networks. In this way, TextRank does not rely on any previous training data at all, but rather can be run on any arbitrary piece of text, and it can produce output simply based on the text's intrinsic properties. Thus the algorithm is easily portable to new domains and languages.

TextRank is a general purpose graph-based ranking algorithm for NLP. Essentially, it runs PageRank on a graph specially designed for a particular NLP task. For keyphrase extraction, it builds a graph using some set of text units as vertices. Edges are based on some measure of semantic or lexical similarity between the text unit vertices. Unlike PageRank, the edges are typically undirected and can be weighted to reflect a degree of similarity. Once the graph is constructed, it is used to form a stochastic matrix, combined with a damping factor (as in the "random surfer model"), and the ranking over vertices is obtained by finding the eigenvector corresponding to eigenvalue 1 (i.e., the stationary distribution of the random walk on the graph).

The vertices should correspond to what we want to rank. Potentially, we could do something similar to the supervised methods and create a vertex for each unigram, bigram, trigram, etc. However, to keep the graph small, the authors decide to rank individual unigrams in a first step, and then include a second step that merges highly ranked adjacent unigrams to form multi-word phrases. This has a nice side effect of allowing us to produce keyphrases of arbitrary length. For example, if we rank unigrams and find that "advanced", "natural", "language", and "processing" all get high ranks, then we would look at the original text and see that these words appear consecutively and create a final keyphrase using all four together. Note that the unigrams placed in the graph can be filtered by part of speech. The authors found that adjectives and nouns were the best to include. Thus, some linguistic knowledge comes into play in this step.

Edges are created based on word co-occurrence in this application of TextRank. Two vertices are connected by an edge if the unigrams appear within a window of size N in the original text. N is typically around 2–10. Thus, "natural" and "language" might be linked in a text about NLP. "Natural" and "processing" would also be linked because they would both appear in the same string of N words. These edges build on the notion of "text cohesion" and the idea that words that appear near each other are likely related in a meaningful way and "recommend" each other to the reader.

Since this method simply ranks the individual vertices, we need a way to threshold or produce a limited number of keyphrases. The technique chosen is to set a count T to be a user-specified fraction of the total number of vertices in the graph. Then the top T vertices/unigrams are selected based on their stationary probabilities. A post- processing step is then applied to merge adjacent instances of these T unigrams. As a result, potentially more or less than T final keyphrases will be produced, but the number should be roughly proportional to the length of the original text.

It is not initially clear why applying PageRank to a co-occurrence graph would produce useful keyphrases. One way to think about it is the following. A word that appears multiple times throughout a text may have many different co-occurring neighbors. For example, in a text about machine learning, the unigram "learning" might co-occur with "machine", "supervised", "un-supervised", and “semi-supervised” in four different sentences. Thus, the “learning” vertex would be a central “hub” that connects to these other modifying words. Running PageRank/TextRank on the graph is likely to rank “learning” highly. Similarly, if the text contains the phrase “supervised classification”, then there would be an edge between “supervised” and “classification”. If “classification” appears several other places and thus has many neighbors, its importance would contribute to the importance of “supervised”. If it ends up with a high rank, it will be selected as one of the top T unigrams, along with “learning” and probably “classification”. In the final post-processing step, we would then end up with keyphrases “supervised learning” and “supervised classification”.

In short, the co-occurrence graph will contain densely connected regions for terms that appear often and in different contexts. A random walk on this graph will have a stationary distribution that assigns large probabilities to the terms in the centers of the clusters. This is similar to densely connected Web pages getting ranked highly by PageRank. This approach has also been used in document summarization, considered below.

Document Summarization

Like keyphrase extraction, document summarization aims to identify the essence of a text. The only real difference is that now we are dealing with larger text units—whole sentences instead of words and phrases.

Before getting into the details of some summarization methods, we will mention how summarization systems are typically evaluated. The most common way is using the so-called ROUGE (Recall-Oriented Understudy for Gisting Evaluation) measure. This is a recall-based measure that determines how well a system-generated summary covers the content present in one or more human-generated model summaries known as references. It is recall-based to encourage systems to include all the important topics in the text. Recall can be computed with respect to unigram, bigram, trigram, or 4-gram matching. For example, ROUGE-1 is computed as division of count of unigrams in reference that appear in system and count of unigrams in reference summary.

If there are multiple references, the ROUGE-1 scores are averaged. Because ROUGE is based only on content overlap, it can determine if the same general concepts are discussed between an automatic summary and a reference summary, but it cannot determine if the result is coherent or the sentences flow together in a sensible manner. High-order n-gram ROUGE measures try to judge fluency to some degree. Note that ROUGE is similar to the BLEU measure for machine translation, but BLEU is precision- based, because translation systems favor accuracy.

A promising line in document summarization is adaptive document/text summarization. The idea of adaptive summarization involves preliminary recognition of document/text genre and subsequent application of summarization algorithms optimized for this genre. First summarizes that perform adaptive summarization have been created.

Supervised Learning Approaches

Supervised text summarization is very much like supervised keyphrase extraction. Basically, if you have a collection of documents and human-generated summaries for them, you can learn features of sentences that make them good candidates for inclusion in the summary. Features might include the position in the document (i.e., the first few sentences are probably important), the number of words in the sentence, etc. The main difficulty in supervised extractive summarization is that the known summaries must be manually created by extracting sentences so the sentences in an original training document can be labeled as "in summary" or "not in summary". This is not typically how people create summaries, so simply using journal abstracts or existing summaries is usually not sufficient. The sentences in these summaries do not necessarily match up with sentences in the original text, so it would be difficult to assign labels to examples for training. Note, however, that these natural summaries can still be used for evaluation purposes, since ROUGE-1 only cares about unigrams.

Maximum Entropy-based Summarization

During the DUC 2001 and 2002 evaluation workshops, TNO developed a sentence extraction system for multi-document summarization in the news domain. The system was based on a hybrid system using a naive Bayes classifier and statistical language models for modeling salience. Although the system exhibited good results, the researchers wanted to explore the effectiveness of a maximum entropy (ME) classifier for the meeting summarization task, as ME is known to be robust against feature dependencies. Maximum entropy has also been applied successfully for summarization in the broadcast news domain.

TextRank and LexRank

The unsupervised approach to summarization is also quite similar in spirit to unsupervised keyphrase extraction and gets around the issue of costly training data. Some unsupervised summarization approaches are based on finding a "centroid" sentence, which is the mean word vector of all the sentences in the document. Then the sentences can be ranked with regard to their similarity to this centroid sentence.

A more principled way to estimate sentence importance is using random walks and eigenvector centrality. LexRank is an algorithm essentially identical to TextRank, and both use this approach for document summarization. The two methods were developed by different groups at the same time, and LexRank simply focused on summarization, but could just as easily be used for keyphrase extraction or any other NLP ranking task.

In both LexRank and TextRank, a graph is constructed by creating a vertex for each sentence in the document.

The edges between sentences are based on some form of semantic similarity or content overlap. While LexRank uses cosine similarity of TF-IDF vectors, TextRank uses a very similar measure based on the number of words two sentences have in common (normalized by the sentences' lengths). The LexRank paper explored using unweighted edges after applying a threshold to the cosine values, but also experimented with using edges with weights equal to the similarity score. TextRank uses continuous similarity scores as weights.

In both algorithms, the sentences are ranked by applying PageRank to the resulting graph. A summary is formed by combining the top ranking sentences, using a threshold or length cutoff to limit the size of the summary.

It is worth noting that TextRank was applied to summarization exactly as described here, while LexRank was used as part of a larger summarization system (MEAD) that combines the LexRank score (stationary probability) with other features like sentence position and length using a linear combination with either user-specified or automatically tuned weights. In this case, some training documents might be needed, though the TextRank results show the additional features are not absolutely necessary.

Another important distinction is that TextRank was used for single document summarization, while LexRank has been applied to multi-document summarization. The task remains the same in both cases—only the number of sentences to choose from has grown. However, when summarizing multiple documents, there is a greater risk of selecting duplicate or highly redundant sentences to place in the same summary. Imagine you have a cluster of news articles on a particular event, and you want to produce one summary. Each article is likely to have many similar sentences, and you would only want to include distinct ideas in the summary. To address this issue, LexRank applies a heuristic post-processing step that builds up a summary by adding sentences in rank order, but discards any sentences that are too similar to ones already placed in the summary. The method used is called Cross-Sentence Information Subsumption (CSIS).

These methods work based on the idea that sentences “recommend” other similar sentences to the reader. Thus, if one sentence is very similar to many others, it will likely be a sentence of great importance. The importance of this sentence also stems from the importance of the sentences “recommending” it. Thus, to get ranked highly and placed in a summary, a sentence must be similar to many sentences that are in turn also similar to many other sentences. This makes intuitive sense and allows the algorithms to be applied to any arbitrary new text. The methods are domain-independent and easily portable. One could imagine the features indicating important sentences in the news domain might vary considerably from the biomedical domain. However, the unsupervised “recommendation”-based approach applies to any domain.

Multi-document Summarization

Multi-document summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. Resulting summary report allows individual users, such as professional information consumers, to quickly familiarize themselves with information contained in a large cluster of documents. In such a way, multi-document summarization systems are complementing the news aggregators performing the next step down the road of coping with information overload. Multi-document summarization may also be done in response to a question.

Multi-document summarization creates information reports that are both concise and comprehensive. With different opinions being put together and outlined, every topic is described from multiple perspectives within a single document. While the goal of a brief summary is to simplify information search and cut the time by pointing to the most relevant source documents, comprehensive multi-document summary should itself contain the required information, hence limiting the need for accessing original files to cases when refinement is required. Automatic summaries present information extracted from multiple sources algorithmically, without any editorial touch or subjective human intervention, thus making it completely unbiased.

Incorporating Diversity

Multi-document extractive summarization faces a problem of potential redundancy. Ideally, we would like to extract sentences that are both "central" (i.e., contain the main ideas) and "diverse" (i.e., they differ from one another). LexRank deals with diversity as a heuristic final stage using CSIS, and other systems have used similar methods, such as Maximal Marginal Relevance (MMR), in trying to eliminate redundancy in information retrieval results. There is a general purpose graph-based ranking algorithm like Page/Lex/TextRank that handles both "centrality" and "diversity" in a unified mathematical framework based on absorbing Markov chain random walks. (An absorbing random walk is like a standard random walk, except some states are now absorbing states that act as "black holes" that cause the walk to end abruptly at that state.) The algorithm is called GRASSHOPPER. In addition to explicitly promoting diversity during the ranking process, GRASSHOPPER incorporates a prior ranking (based on sentence position in the case of summarization).

The state of the art results for multi-document summarization, however, are obtained using mixtures of submodular functions. These methods have achieved the state of the art results for Document Summarization Corpora, DUC 04 - 07. Similar results were also achieved with the use of determinantal point processes (which are a special case of submodular functions) for DUC-04.

Submodular Functions as Generic Tools for Summarization

The idea of a Submodular set function has recently emerged as a powerful modeling tool for various summarization problems. Submodular functions naturally model notions of *coverage*, *information*, *representation* and *diversity*. Moreover, several important combinatorial optimization problems occur as special instances of submodular optimization. For example, the set cover problem is a special case of submodular optimization, since the set cover function is submodular. The set cover function attempts to find a subset of objects which *cover* a given set of concepts. For example, in document summarization, one would like the summary to cover all important and relevant concepts in the document. This is an instance of set cover. Similarly, the facility location problem is a special case of submodular functions. The Facility Location function also naturally models coverage and diversity. Another example of a submodular optimization problem is using a Determinantal point process to model diversity. Similarly, the Maximum-Marginal-Relevance procedure can also be seen as an instance of submodular optimization. All these important models encouraging coverage, diversity and information are all submodular. Moreover, submodular functions can be efficiently combined together, and the resulting function is still submodular. Hence, one could combine one submodular function which models diversity, another one which models coverage and use human supervision to learn a right model of a submodular function for the problem.

While submodular functions are fitting problems for summarization, they also admit very efficient algorithms for optimization. For example, a simple greedy algorithm admits a constant factor guarantee. Moreover, the greedy algorithm is extremely simple to implement and can scale to large datasets, which is very important for summarization problems.

Submodular functions have achieved state-of-the-art for almost all summarization problems. For example, work by Lin and Bilmes, 2012 shows that submodular functions achieve the best results to date on DUC-04, DUC-05, DUC-06 and DUC-07 systems for document summarization. Similarly, work by Lin and Bilmes, 2011, shows that many existing systems for automatic summarization are instances of submodular functions. This was a break through result establishing submodular functions as the right models for summarization problems.

Submodular Functions have also been used for other summarization tasks. Tschiatschek et al., 2014 show that mixtures of submodular functions achieve state-of-the-art results for image collection summarization. Similarly, Bairi et al., 2015 show the utility of submodular functions for summarizing multi-document topic hierarchies. Submodular Functions have also successfully been used for summarizing machine learning datasets.

Evaluation Techniques

The most common way to evaluate the informativeness of automatic summaries is to compare them with human-made model summaries.

Evaluation techniques fall into intrinsic and extrinsic, inter-textual and intra-textual.

Intrinsic and Extrinsic Evaluation

An intrinsic evaluation tests the summarization system in and of itself while an extrinsic evaluation tests the summarization based on how it affects the completion of some other task. Intrinsic evaluations have assessed mainly the coherence and informativeness of summaries. Extrinsic evaluations, on the other hand, have tested the impact of summarization on tasks like relevance assessment, reading comprehension, etc.

Inter-textual and Intra-textual

Intra-textual methods assess the output of a specific summarization system, and the inter-textual ones focus on contrastive analysis of outputs of several summarization systems.

Human judgement often has wide variance on what is considered a “good” summary, which means that making the evaluation process automatic is particularly difficult. Manual evaluation can be used, but this is both time and labor-intensive as it requires humans to read not only the summaries but also the source documents. Other issues are those concerning coherence and coverage.

One of the metrics used in NIST’s annual Document Understanding Conferences, in which research groups submit their systems for both summarization and translation tasks, is the ROUGE metric (Recall-Oriented Understudy for Gisting Evaluation). It essentially calculates n-gram overlaps between automatically generated summaries and previously-written human summaries. A high level of overlap should indicate a high level of shared concepts between the two summaries. Note that overlap metrics like this are unable to provide any feedback on a summary’s coherence. Anaphor resolution remains another problem yet to be fully solved. Similarly, for image summarization, Tschiatschek et al., developed a Visual-ROUGE score which judges the performance of algorithms for image summarization.

Current Challenges in Evaluating Summaries Automatically

Evaluating summaries, either manually or automatically, is a hard task. The main difficulty in evaluation comes from the impossibility of building a fair gold-standard against which the results of the systems can be compared. Furthermore, it is also very hard to determine what a correct summary is, because there is always the possibility of a system to generate a good summary that is quite different from any human summary used as an approximation to the correct output.

Content selection is not a deterministic problem. People are subjective, and different authors would choose different sentences. And individuals may not be consistent. A particular person may chose different sentences at different times. Two distinct sentences expressed in different words can express the same meaning. This phenomenon is known as paraphrasing. We can find an approach to automatically evaluating summaries using paraphrases (ParaEval).

Most summarization systems perform an extractive approach, selecting and copying important sentences from the source documents. Although humans can also cut and paste relevant information of a text, most of the times they rephrase sentences when necessary, or they join different related information into one sentence.

Domain Specific Versus Domain Independent Summarization Techniques

Domain independent summarization techniques generally apply sets of general features which can be used to identify information-rich text segments. Recent research focus has drifted to domain-specific summarization techniques that utilize the available knowledge specific to the domain of text. For example, automatic summarization research on medical text generally attempts to utilize the various sources of codified medical knowledge and ontologies.

Evaluating Summaries Qualitatively

The main drawback of the evaluation systems existing so far is that we need at least one reference summary, and for some methods more than one, to be able to compare automatic summaries with models. This is a hard and expensive task. Much effort has to be done in order to have corpus of texts and their corresponding summaries. Furthermore, for some methods, not only do we need to have human-made summaries available for comparison, but also manual annotation has to be performed in some of them (e.g. SCU in the Pyramid Method). In any case, what the evaluation methods need as an input, is a set of summaries to serve as gold standards and a set of automatic summaries. Moreover, they all perform a quantitative evaluation with regard to different similarity metrics. To overcome these problems, we think that the quantitative evaluation might not be the only way to evaluate summaries, and a qualitative automatic evaluation would be also important.

Examples of Data Mining

Data mining has been used in many applications. Some notable examples of usage are:

Games

Since the early 1960s, with the availability of oracles for certain combinatorial games, also called tablebases (e.g. for 3x3-chess) with any beginning configuration, small-board dots-and-boxes, small-board-hex, and certain endgames in chess, dots-and-boxes, and hex; a new area for data mining has been opened. This is the extraction of human-usable strategies from these oracles. Current pattern recognition approaches do not seem to fully acquire the high level of abstraction required to be applied successfully. Instead, extensive experimentation with the tablebases – combined with an intensive study of tablebase-answers to well designed problems, and with knowledge of prior art (i.e., pre-tablebase knowledge) – is used to yield insightful patterns. Berlekamp (in dots-and-boxes, etc.) and John Nunn (in chess endgames) are notable examples of researchers doing this work, though they were not – and are not – involved in tablebase generation.

Business

In business, data mining is the analysis of historical business activities, stored as static data in data warehouse databases. The goal is to reveal hidden patterns and trends. Data mining software uses advanced pattern recognition algorithms to sift through large amounts of data to assist in discovering previously unknown strategic business information. Examples of what businesses use data mining is to include performing market analysis to identify new product bundles, finding the root cause of manufacturing problems, to prevent customer attrition and acquire new customers, cross-selling to existing customers, and profiling customers with more accuracy.

- In today's world raw data is being collected by companies at an exploding rate. For example, Walmart processes over 20 million point-of-sale transactions every day. This information is stored in a centralized database, but would be useless without some type of data mining software to analyze it. If Walmart analyzed their point-of-sale data with data mining techniques they would be able to determine sales trends, develop marketing campaigns, and more accurately predict customer loyalty.
- Categorization of the items available in the e-commerce site is a fundamental problem. A correct item categorization system is essential for user experience as it helps determine the items relevant to him for search and browsing. Item categorization can be formulated as a supervised classification problem in data mining where the categories are the target classes and the features are the words composing some textual description of the items. One of the approaches is to find groups initially which are similar and place them together in a latent group. Now given a new item, first classify into a latent group which is called coarse level classification. Then, do a second round of classification to find the category to which the item belongs to.
- Every time a credit card or a store loyalty card is being used, or a warranty card is being filled, data is being collected about the users behavior. Many people find the amount of information stored about us from companies, such as Google, Facebook, and Amazon, disturbing and are concerned about privacy. Although there is the potential for our personal data to be used in harmful, or unwanted, ways it is also being used to make our lives better. For example, Ford and Audi hope to one day collect information about customer driving patterns so they can recommend safer routes and warn drivers about dangerous road conditions.
- Data mining in customer relationship management applications can contribute significantly to the bottom line. Rather than randomly contacting a prospect or customer through a call center or sending mail, a company can concentrate its efforts on prospects that are predicted to have a high likelihood of responding to an offer. More sophisticated methods may be used to optimize resources across campaigns so that one may predict to which

channel and to which offer an individual is most likely to respond (across all potential offers). Additionally, sophisticated applications could be used to automate mailing. Once the results from data mining (potential prospect/customer and channel/offer) are determined, this “sophisticated application” can either automatically send an e-mail or a regular mail. Finally, in cases where many people will take an action without an offer, “uplift modeling” can be used to determine which people have the greatest increase in response if given an offer. Uplift modeling thereby enables marketers to focus mailings and offers on persuadable people, and not to send offers to people who will buy the product without an offer. Data clustering can also be used to automatically discover the segments or groups within a customer data set.

- Businesses employing data mining may see a return on investment, but also they recognize that the number of predictive models can quickly become very large. For example, rather than using one model to predict how many customers will churn, a business may choose to build a separate model for each region and customer type. In situations where a large number of models need to be maintained, some businesses turn to more automated data mining methodologies.

- Data mining can be helpful to human resources (HR) departments in identifying the characteristics of their most successful employees. Information obtained – such as universities attended by highly successful employees – can help HR focus recruiting efforts accordingly. Additionally, Strategic Enterprise Management applications help a company translate corporate-level goals, such as profit and margin share targets, into operational decisions, such as production plans and workforce levels.

- Market basket analysis, relates to data-mining use in retail sales. If a clothing store records the purchases of customers, a data mining system could identify those customers who favor silk shirts over cotton ones. Although some explanations of relationships may be difficult, taking advantage of it is easier. The example deals with association rules within transaction-based data. Not all data are transaction based and logical, or inexact rules may also be present within a database.

- Market basket analysis has been used to identify the purchase patterns of the Alpha Consumer. Analyzing the data collected on this type of user has allowed companies to predict future buying trends and forecast supply demands.

- Data mining is a highly effective tool in the catalog marketing industry. Catalogers have a rich database of history of their customer transactions for millions of customers dating back a number of years. Data mining tools can identify patterns among customers and help identify the most likely customers to respond to upcoming mailing campaigns.

- Data mining for business applications can be integrated into a complex modeling and decision making process. LIONsolver uses Reactive business intelligence (RBI) to advocate a “holistic” approach that integrates data mining, modeling, and interactive visualization into an end-to-end discovery and continuous innovation process powered by human and automated learning.

- In the area of decision making, the RBI approach has been used to mine knowledge that is progressively acquired from the decision maker, and then self-tune the decision method accordingly. The relation between the quality of a data mining system and the amount of investment that the decision maker is willing to make was formalized by providing an economic perspective on the value of “extracted knowledge” in terms of its payoff to the organization This decision-theoretic classification framework was applied to a real-world semiconductor wafer manufacturing line, where decision rules for effectively monitoring and controlling the semiconductor wafer fabrication line were developed.

- An example of data mining related to an integrated-circuit (IC) production line is described in the paper “Mining IC Test Data to Optimize VLSI Testing.” In this paper, the application of data mining and decision analysis to the problem of die-level functional testing is described. Experiments mentioned demonstrate the ability to apply a system of mining historical die-test data to create a probabilistic model of patterns of die failure. These patterns are then utilized to decide, in real time, which die to test next and when to stop testing. This system has been shown, based on experiments with historical test data, to have the potential to improve profits on mature IC products. Other examples of the application of data mining methodologies in semiconductor manufacturing environments suggest that data mining methodologies may be particularly useful when data is scarce, and the various physical and chemical parameters that affect the process exhibit highly complex interactions. Another implication is that on-line monitoring of the semiconductor manufacturing process using data mining may be highly effective.

Science and Engineering

In recent years, data mining has been used widely in the areas of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

- In the study of human genetics, sequence mining helps address the important goal of understanding the mapping relationship between the inter-individual variations in human DNA sequence and the variability in disease susceptibility. In simple terms, it aims to find out how the changes in an individual’s DNA sequence affects the risks of developing common diseases such as cancer, which is of great importance to improving methods of diagnosing, preventing, and treating these diseases. One data mining method that is used to perform this task is known as multifactor dimensionality reduction.
- In the area of electrical power engineering, data mining methods have been widely used for condition monitoring of high voltage electrical equipment. The purpose of condition monitoring is to obtain valuable information on, for example, the status of the insulation (or other important safety-related parameters). Data clustering techniques – such as the self-organizing map (SOM), have been applied to vibration monitoring and analysis of transformer on-load tap-changers (OLTCS). Using vibration monitoring, it can be observed that each tap change operation generates a signal that contains information about the condition of the tap changer contacts and the drive mechanisms. Obviously, different tap positions will generate different signals. However, there was considerable variability amongst normal condition signals for exactly the same tap position. SOM has been applied to detect abnormal conditions and to hypothesize about the nature of the abnormalities.
- Data mining methods have been applied to dissolved gas analysis (DGA) in power transformers. DGA, as a diagnostics for power transformers, has been available for many years.

Methods such as SOM has been applied to analyze generated data and to determine trends which are not obvious to the standard DGA ratio methods (such as Duval Triangle).

- In educational research, where data mining has been used to study the factors leading students to choose to engage in behaviors which reduce their learning, and to understand factors influencing university student retention. A similar example of social application of data mining is its use in expertise finding systems, whereby descriptors of human expertise are extracted, normalized, and classified so as to facilitate the finding of experts, particularly in scientific and technical fields. In this way, data mining can facilitate institutional memory.
- Data mining methods of biomedical data facilitated by domain ontologies, mining clinical trial data, and traffic analysis using SOM.
- In adverse drug reaction surveillance, the Uppsala Monitoring Centre has, since 1998, used data mining methods to routinely screen for reporting patterns indicative of emerging drug safety issues in the WHO global database of 4.6 million suspected adverse drug reaction incidents. Recently, similar methodology has been developed to mine large collections of electronic health records for temporal patterns associating drug prescriptions to medical diagnoses.
- Data mining has been applied to software artifacts within the realm of software engineering: Mining Software Repositories.

Human Rights

Data mining of government records – particularly records of the justice system (i.e., courts, prisons) – enables the discovery of systemic human rights violations in connection to generation and publication of invalid or fraudulent legal records by various government agencies.

Medical Data Mining

Some machine learning algorithms can be applied in medical field as second-opinion diagnostic tools and as tools for the knowledge extraction phase in the process of knowledge discovery in databases. One of these classifiers (called *Prototype exemplar learning classifier* (PEL-C) is able to discover syndromes as well as atypical clinical cases.

In 2011, the case of Sorrell v. IMS Health, Inc., decided by the Supreme Court of the United States, ruled that pharmacies may share information with outside companies. This practice was authorized under the 1st Amendment of the Constitution, protecting the “freedom of speech.” However, the passage of the Health Information Technology for Economic and Clinical Health Act (HITECH Act) helped to initiate the adoption of the electronic health record (EHR) and supporting technology in the United States. The HITECH Act was signed into law on February 17, 2009 as part of the American Recovery and Reinvestment Act (ARRA) and helped to open the door to medical data mining. Prior to the signing of this law, estimates of only 20% of United States-based physicians were utilizing electronic patient records. Søren Brunak notes that “the patient record becomes as information-rich as possible” and thereby “maximizes the data mining opportunities.” Hence, electronic patient records further expands the possibilities regarding medical data mining thereby opening the door to a vast source of medical data analysis.

Spatial Data Mining

Spatial data mining is the application of data mining methods to spatial data. The end objective of spatial data mining is to find patterns in data with respect to geography. So far, data mining and Geographic Information Systems (GIS) have existed as two separate technologies, each with its own methods, traditions, and approaches to visualization and data analysis. Particularly, most contemporary GIS have only very basic spatial analysis functionality. The immense explosion in geographically referenced data occasioned by developments in IT, digital mapping, remote sensing, and the global diffusion of GIS emphasizes the importance of developing data-driven inductive approaches to geographical analysis and modeling.

Data mining offers great potential benefits for GIS-based applied decision-making. Recently, the task of integrating these two technologies has become of critical importance, especially as various public and private sector organizations possessing huge databases with thematic and geographically referenced data begin to realize the huge potential of the information contained therein. Among those organizations are:

- offices requiring analysis or dissemination of geo-referenced statistical data
- public health services searching for explanations of disease clustering
- environmental agencies assessing the impact of changing land-use patterns on climate change
- geo-marketing companies doing customer segmentation based on spatial location.

Challenges in Spatial mining: Geospatial data repositories tend to be very large. Moreover, existing GIS datasets are often splintered into feature and attribute components that are conventionally archived in hybrid data management systems. Algorithmic requirements differ substantially for relational (attribute) data management and for topological (feature) data management. Related to this is the range and diversity of geographic data formats, which present unique challenges. The digital geographic data revolution is creating new types of data formats beyond the traditional "vector" and "raster" formats. Geographic data repositories increasingly include ill-structured data, such as imagery and geo-referenced multi-media.

There are several critical research challenges in geographic knowledge discovery and data mining. Miller and Han offer the following list of emerging research topics in the field:

- Developing and supporting geographic data warehouses (GDW's): Spatial properties are often reduced to simple aspatial attributes in mainstream data warehouses. Creating an integrated GDW requires solving issues of spatial and temporal data interoperability – including differences in semantics, referencing systems, geometry, accuracy, and position.
- Better spatio-temporal representations in geographic knowledge discovery: Current geographic knowledge discovery (GKD) methods generally use very simple representations of geographic objects and spatial relationships. Geographic data mining methods should recognize more complex geographic objects (i.e., lines and polygons) and relationships (i.e., non-Euclidean distances, direction, connectivity, and interaction through attributed geographic space such as terrain). Furthermore, the time dimension needs to be more fully

integrated into these geographic representations and relationships.

- Geographic knowledge discovery using diverse data types: GKD methods should be developed that can handle diverse data types beyond the traditional raster and vector models, including imagery and geo-referenced multimedia, as well as dynamic data types (video streams, animation).

Temporal Data Mining

Data may contain attributes generated and recorded at different times. In this case finding meaningful relationships in the data may require considering the temporal order of the attributes. A temporal relationship may indicate a causal relationship, or simply an association.

Sensor Data Mining

Wireless sensor networks can be used for facilitating the collection of data for spatial data mining for a variety of applications such as air pollution monitoring. A characteristic of such networks is that nearby sensor nodes monitoring an environmental feature typically register similar values. This kind of data redundancy due to the spatial correlation between sensor observations inspires the techniques for in-network data aggregation and mining. By measuring the spatial correlation between data sampled by different sensors, a wide class of specialized algorithms can be developed to develop more efficient spatial data mining algorithms.

Visual Data Mining

In the process of turning from analog into digital, large data sets have been generated, collected, and stored discovering statistical patterns, trends and information which is hidden in data, in order to build predictive patterns. Studies suggest visual data mining is faster and much more intuitive than is traditional data mining.

Music Data Mining

Data mining techniques, and in particular co-occurrence analysis, has been used to discover relevant similarities among music corpora (radio lists, CD databases) for purposes including classifying music into genres in a more objective manner.

Surveillance

Data mining has been used by the U.S. government. Programs include the Total Information Awareness (TIA) program, Secure Flight (formerly known as Computer-Assisted Passenger Prescreening System (CAPPS II)), Analysis, Dissemination, Visualization, Insight, Semantic Enhancement (ADVISE), and the Multi-state Anti-Terrorism Information Exchange (MATRIX). These programs have been discontinued due to controversy over whether they violate the 4th Amendment to the United States Constitution, although many programs that were formed under them continue to be funded by different organizations or under different names.

In the context of combating terrorism, two particularly plausible methods of data mining are “pattern mining” and “subject-based data mining”.

Pattern Mining

"Pattern mining" is a data mining method that involves finding existing patterns in data. In this context *patterns* often means association rules. The original motivation for searching association rules came from the desire to analyze supermarket transaction data, that is, to examine customer behavior in terms of the purchased products. For example, an association rule "beer ☐ potato chips (80%)" states that four out of five customers that bought beer also bought potato chips.

In the context of pattern mining as a tool to identify terrorist activity, the National Research Council provides the following definition: "Pattern-based data mining looks for patterns (including anomalous data patterns) that might be associated with terrorist activity — these patterns might be regarded as small signals in a large ocean of noise." Pattern Mining includes new areas such a Music Information Retrieval (MIR) where patterns seen both in the temporal and non temporal domains are imported to classical knowledge discovery search methods.

Subject-based Data Mining

"Subject-based data mining" is a data mining method involving the search for associations between individuals in data. In the context of combating terrorism, the National Research Council provides the following definition: "Subject-based data mining uses an initiating individual or other datum that is considered, based on other information, to be of high interest, and the goal is to determine what other persons or financial transactions or movements, etc., are related to that initiating datum."

Knowledge Grid

Knowledge discovery "On the Grid" generally refers to conducting knowledge discovery in an open environment using grid computing concepts, allowing users to integrate data from various online data sources, as well make use of remote resources, for executing their data mining tasks. The earliest example was the Discovery Net, developed at Imperial College London, which won the "Most Innovative Data-Intensive Application Award" at the ACM SC02 (Supercomputing 2002) conference and exhibition, based on a demonstration of a fully interactive distributed knowledge discovery application for a bioinformatics application. Other examples include work conducted by researchers at the University of Calabria, who developed a Knowledge Grid architecture for distributed knowledge discovery, based on grid computing.

References

- Battiti, Roberto; and Brunato, Mauro; Reactive Business Intelligence. From Data to Models to Insight, Reactive Search Srl, Italy, February 2011. ISBN 978-88-905795-0-9.
- Monk, Ellen; Wagner, Bret (2006). Concepts in Enterprise Resource Planning, Second Edition. Boston, MA: Thomson Course Technology. ISBN 0-619-21663-8. OCLC 224465825.
- Zhu, Xingquan; Davidson, Ian (2007). Knowledge Discovery and Data Mining: Challenges and Realities. New York, NY: Hershey. pp. 163–189. ISBN 978-1-59904-252-7.
- Haag, Stephen; Cummings, Maeve; Phillips, Amy (2006). Management Information Systems for the information age. Toronto: McGraw-Hill Ryerson. p. 28. ISBN 0-07-095569-7. OCLC 63194770.
- Good, P. I.; Hardin, J. W. (2009). Common Errors in Statistics (And How to Avoid Them) (3rd ed.). Hoboken, New Jersey: Wiley. p. 211. ISBN 978-0-470-45798-6.

- Fotheringham, A. Stewart; Brunsdon, Chris; Charlton, Martin (2002). Geographically weighted regression: the analysis of spatially varying relationships (Reprint ed.). Chichester, England: John Wiley. ISBN 978-0-471-49616-8.
- Manning, Christopher D.; Raghavan, Prabhakar; Schütze, Hinrich. Introduction to Information Retrieval. Cambridge University Press. ISBN 978-0-521-86571-5.
- Bailey, Ken (1994). "Numerical Taxonomy and Cluster Analysis". Typologies and Taxonomies. p. 34. ISBN 9780803952591.
- Hájek, Petr; Feglar, Tomas; Rauch, Jan; and Coufal, David; The GUHA method, data preprocessing and mining, Database Support for Data Mining Applications, Springer, 2004, ISBN 978-3-540-22479-2
- Tan, Pang-Ning; Michael, Steinbach; Kumar, Vipin (2005). "Chapter 6. Association Analysis: Basic Concepts and Algorithms" (PDF). Introduction to Data Mining. Addison-Wesley. ISBN 0-321-32136-7.
- Angiulli, F.; Pizzuti, C. (2002). Fast Outlier Detection in High Dimensional Spaces. Principles of Data Mining and Knowledge Discovery. Lecture Notes in Computer Science. 2431. p. 15. doi:10.1007/3-540-45681-3_2. ISBN 978-3-540-44037-6.
- Witten, Ian H.; Frank, Eibe; Hall, Mark A. (30 January 2011). Data Mining: Practical Machine Learning Tools and Techniques (3 ed.). Elsevier. ISBN 978-0-12-374856-0.
- Mena, Jesús (2011). Machine Learning Forensics for Law Enforcement, Security, and Intelligence. Boca Raton, FL: CRC Press (Taylor & Francis Group). ISBN 978-1-4398-6069-4.
- Kantardzic, Mehmed (2003). Data Mining: Concepts, Models, Methods, and Algorithms. John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

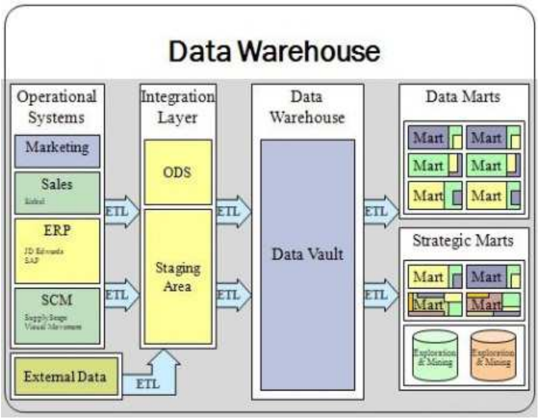
4

Understanding Data Warehousing

Data warehouse is the core of business intelligence. It is majorly used for reporting and analyzing data. Data mart, master data management, dimension, slowly changing dimension and star schema. This text elucidates the crucial theories and principles of data warehousing.

Data Warehouse

In computing, a data warehouse (DW or DWH), also known as an enterprise data warehouse (EDW), is a system used for reporting and data analysis, and is considered a core component of business intelligence. DWs are central repositories of integrated data from one or more disparate sources. They store current and historical data and are used for creating analytical reports for knowledge workers throughout the enterprise. Examples of reports could range from annual and quarterly comparisons and trends to detailed daily sales analysis.


Data Warehouse
Operational Systems
Marketing
Sales
ERP
JD Edwards
SAP
SCM
External Data
ETL
Integration Layer
ODS
Staging Area
Data Warehouse
Data Vault
Data Marts
Mart
Strategic Marts
Exploration & Mining

Data Warehouse Overview

The data stored in the warehouse is uploaded from the operational systems (such as marketing or sales). The data may pass through an operational data store for additional operations before it is used in the DW for reporting.

Types of Systems

Data Mart

> A data mart is a simple form of a data warehouse that is focused on a single subject (or functional area), hence they draw data from a limited number of sources such as sales, finance or marketing. Data marts are often built and controlled by a single department within an organization. The sources could be internal operational systems, a central data

warehouse, or external data. Denormalization is the norm for data modeling techniques in this system. Given that data marts generally cover only a subset of the data contained in a data warehouse, they are often easier and faster to implement.

Difference between data warehouse and data mart	
Data warehouse	Data mart
enterprise-wide data	department-wide data
multiple subject areas	single subject area
difficult to build	easy to build
takes more time to build	less time to build
larger memory	limited memory

Types of Data Marts

- Dependent data mart
- Independent data mart
- Hybrid data mart

Online analytical processing (OLAP)

OLAP is characterized by a relatively low volume of transactions. Queries are often very complex and involve aggregations. For OLAP systems, response time is an effectiveness measure. OLAP applications are widely used by Data Mining techniques. OLAP databases store aggregated, historical data in multi-dimensional schemas (usually star schemas). OLAP systems typically have data latency of a few hours, as opposed to data marts, where latency is expected to be closer to one day. The OLAP approach is used to analyze multidimensional data from multiple sources and perspectives. The three basic operations in OLAP are : Roll-up (Consolidation), Drill-down and Slicing & Dicing.

Online transaction processing (OLTP)

OLTP is characterized by a large number of short on-line transactions (INSERT, UPDATE, DELETE). OLTP systems emphasize very fast query processing and maintaining data integrity in multi-access environments. For OLTP systems, effectiveness is measured by the number of transactions per second. OLTP databases contain detailed and current data. The schema used to store transactional databases is the entity model (usually 3NF). Normalization is the norm for data modeling techniques in this system.

Predictive analysis

Predictive analysis is about finding and quantifying hidden patterns in the data using complex mathematical models that can be used to predict future outcomes. Predictive analysis is different from OLAP in that OLAP focuses on historical data analysis and is reactive in nature, while predictive analysis focuses on the future. These systems are also used for CRM (customer relationship management).

Software Tools

The typical extract-transform-load (ETL)-based data warehouse uses staging, data integration, and access layers to house its key functions. The staging layer or staging database stores raw data extracted from each of the disparate source data systems. The integration layer integrates the disparate data sets by transforming the data from the staging layer often storing this transformed data in an operational data store (ODS) database. The integrated data are then moved to yet another database, often called the data warehouse database, where the data is arranged into hierarchical groups often called dimensions and into facts and aggregate facts. The combination of facts and dimensions is sometimes called a star schema. The access layer helps users retrieve data.

This definition of the data warehouse focuses on data storage. The main source of the data is cleaned, transformed, cataloged and made available for use by managers and other business professionals for data mining, online analytical processing, market research and decision support. However, the means to retrieve and analyze data, to extract, transform and load data, and to manage the data dictionary are also considered essential components of a data warehousing system. Many references to data warehousing use this broader context. Thus, an expanded definition for data warehousing includes business intelligence tools, tools to extract, transform and load data into the repository, and tools to manage and retrieve metadata.

Benefits

A data warehouse maintains a copy of information from the source transaction systems. This architectural complexity provides the opportunity to:

- Integrate data from multiple sources into a single database and data model. Mere congregation of data to single database so a single query engine can be used to present data is an ODS.
- Mitigate the problem of database isolation level lock contention in transaction processing systems caused by attempts to run large, long running, analysis queries in transaction processing databases.
- Maintain data history, even if the source transaction systems do not.
- Integrate data from multiple source systems, enabling a central view across the enterprise. This benefit is always valuable, but particularly so when the organization has grown by merger.
- Improve data quality, by providing consistent codes and descriptions, flagging or even fixing bad data.
- Present the organization's information consistently.
- Provide a single common data model for all data of interest regardless of the data's source.
- Restructure the data so that it makes sense to the business users.
- Restructure the data so that it delivers excellent query performance, even for complex an-

alytic queries, without impacting the operational systems.

- Add value to operational business applications, notably customer relationship management (CRM) systems.
- Make decision–support queries easier to write.
- Optimized data warehouse architectures allow data scientists to organize and disambiguate repetitive data.

Generic Environment

The environment for data warehouses and marts includes the following:

- Source systems that provide data to the warehouse or mart;
- Data integration technology and processes that are needed to prepare the data for use;
- Different architectures for storing data in an organization’s data warehouse or data marts;
- Different tools and applications for the variety of users;
- Metadata, data quality, and governance processes must be in place to ensure that the warehouse or mart meets its purposes.

In regards to source systems listed above, Rainer states, “A common source for the data in data warehouses is the company’s operational databases, which can be relational databases”.

Regarding data integration, Rainer states, “It is necessary to extract data from source systems, transform them, and load them into a data mart or warehouse”.

Rainer discusses storing data in an organization’s data warehouse or data marts.

Metadata are data about data. “IT personnel need information about data sources; database, table, and column names; refresh schedules; and data usage measures”.

Today, the most successful companies are those that can respond quickly and flexibly to market changes and opportunities. A key to this response is the effective and efficient use of data and information by analysts and managers. A “data warehouse” is a repository of historical data that are organized by subject to support decision makers in the organization. Once data are stored in a data mart or warehouse, they can be accessed.

History

The concept of data warehousing dates back to the late 1980s when IBM researchers Barry Devlin and Paul Murphy developed the “business data warehouse”. In essence, the data warehousing concept was intended to provide an architectural model for the flow of data from operational systems to decision support environments. The concept attempted to address the various problems associated with this flow, mainly the high costs associated with it. In the absence of a data warehousing architecture, an enormous amount of redundancy was required to support multiple decision support environments. In larger corporations it was typical for multiple decision support environments to operate independently. Though each environment served different users, they often required much of the same stored data. The process of gathering, cleaning and integrating data from various sources, usually from long-term existing operational systems (usually referred to as legacy systems), was typically in part replicated for each environment. Moreover, the operational systems were frequently reexamined as new decision support requirements emerged. Often new requirements necessitated gathering, cleaning and integrating new data from "data marts" that were tailored for ready access by users.

Key developments in early years of data warehousing were:

- 1960s – General Mills and Dartmouth College, in a joint research project, develop the terms *dimensions* and *facts*.
- 1970s – ACNielsen and IRI provide dimensional data marts for retail sales.
- 1970s – Bill Inmon begins to define and discuss the term: Data Warehouse.
- 1975 – Sperry Univac introduces MAPPER (MAintain, Prepare, and Produce Executive Reports) is a database management and reporting system that includes the world's first 4GL. First platform designed for building Information Centers (a forerunner of contemporary Enterprise Data Warehousing platforms)
- 1983 – Teradata introduces a database management system specifically designed for decision support.
- 1984 – Metaphor Computer Systems, founded by David Liddle and Don Massaro, releases Data Interpretation System (DIS). DIS was a hardware/software package and GUI for business users to create a database management and analytic system.
- 1988 – Barry Devlin and Paul Murphy publish the article *An architecture for a business and information system* where they introduce the term "business data warehouse".
- 1990 – Red Brick Systems, founded by Ralph Kimball, introduces Red Brick Warehouse, a database management system specifically for data warehousing.
- 1991 – Prism Solutions, founded by Bill Inmon, introduces Prism Warehouse Manager, software for developing a data warehouse.
- 1992 – Bill Inmon publishes the book *Building the Data Warehouse*.
- 1995 – The Data Warehousing Institute, a for-profit organization that promotes data warehousing, is founded.
- 1996 – Ralph Kimball publishes the book *The Data Warehouse Toolkit*.
- 2012 – Bill Inmon developed and made public technology known as "textual disambiguation". Textual disambiguation applies context to raw text and reformats the raw text and context into a standard data base format. Once raw text is passed through textual disambiguation, it can easily and efficiently be accessed and analyzed by standard business intelligence technology. Textual disambiguation is accomplished through the execution of textual ETL. Textual disambiguation is useful wherever raw text is found, such as in documents, Hadoop, email, and so forth.

Information Storage

Facts

A fact is a value or measurement, which represents a fact about the managed entity or system.

Facts as reported by the reporting entity are said to be at raw level. E.g. in a mobile telephone system, if a BTS (base transceiver station) received 1,000 requests for traffic channel allocation, it allocates for 820 and rejects the remaining then it would report 3 facts or measurements to a management system:

- tch_req_total = 1000
- tch_req_success = 820
- tch_req_fail = 180

Facts at the raw level are further aggregated to higher levels in various dimensions to extract more service or business-relevant information from it. These are called aggregates or summaries or aggregated facts.

For instance, if there are 3 BTSs in a city, then the facts above can be aggregated from the BTS to the city level in the network dimension. For example:

- $tch\_req\_success\_city = tch\_req\_success\_bts1 + tch\_req\_success\_bts2 + tch\_req\_success\_bts3$
- $avg\_tch\_req\_success\_city = (tch\_req\_success\_bts1 + tch\_req\_success\_bts2 + tch\_req\_success\_bts3)/3$

Dimensional Versus Normalized Approach for Storage of Data

There are three or more leading approaches to storing data in a data warehouse — the most important approaches are the dimensional approach and the normalized approach.

The dimensional approach refers to Ralph Kimball's approach in which it is stated that the data warehouse should be modeled using a Dimensional Model/star schema. The normalized approach, also called the 3NF model (Third Normal Form) refers to Bill Inmon's approach in which it is stated that the data warehouse should be modeled using an E-R model/normalized model.

In a dimensional approach, transaction data are partitioned into "facts", which are generally numeric transaction data, and "dimensions", which are the reference information that gives context to the facts. For example, a sales transaction can be broken up into facts such as the number of products ordered and the total price paid for the products, and into dimensions such as order date, customer name, product number, order ship-to and bill-to locations, and salesperson responsible for receiving the order.

A key advantage of a dimensional approach is that the data warehouse is easier for the user to understand and to use. Also, the retrieval of data from the data warehouse tends to operate very quickly. Dimensional structures are easy to understand for business users, because the structure is divided into measurements/facts and context/dimensions. Facts are related to the organization's business processes and operational system whereas the dimensions surrounding them contain context about the measurement (Kimball, Ralph 2008). Another advantage offered by dimensional model is that it does not involve a relational database every time. Thus, this type of modeling technique is very useful for end-user queries in data warehouse.

The main disadvantages of the dimensional approach are the following:

1. In order to maintain the integrity of facts and dimensions, loading the data warehouse with data from different operational systems is complicated.
2. It is difficult to modify the data warehouse structure if the organization adopting the dimensional approach changes the way in which it does business.

In the normalized approach, the data in the data warehouse are stored following, to a degree, database normalization rules. Tables are grouped together by *subject areas* that reflect general data categories (e.g., data on customers, products, finance, etc.). The normalized structure divides data into entities, which creates several tables in a relational database. When applied in large enterprises the result is dozens of tables that are linked together by a web of joins. Furthermore, each of the created entities is converted into separate physical tables when the database is implemented (Kimball, Ralph 2008). The main advantage of this approach is that it is straightforward to add information into the database. Some disadvantages of this approach are that, because of the number of tables involved, it can be difficult for users to join data from different sources into meaningful information and to access the information without a precise understanding of the sources of data and of the data structure of the data warehouse.

Both normalized and dimensional models can be represented in entity-relationship diagrams as both contain joined relational tables. The difference between the two models is the degree of normalization (also known as Normal Forms). These approaches are not mutually exclusive, and there are other approaches. Dimensional approaches can involve normalizing data to a degree (Kimball, Ralph 2008).

In *Information-Driven Business*, Robert Hillard proposes an approach to comparing the two approaches based on the information needs of the business problem. The technique shows that normalized models hold far more information than their dimensional equivalents (even when the same fields are used in both models) but this extra information comes at the cost of usability. The technique measures information quantity in terms of information entropy and usability in terms of the Small Worlds data transformation measure.

Design Methods

Bottom-up Design

In the *bottom-up* approach, data marts are first created to provide reporting and analytical capabilities for specific business processes. These data marts can then be integrated to create a comprehensive data warehouse. The data warehouse bus architecture is primarily an implementation of “the bus”, a collection of conformed dimensions and conformed facts, which are dimensions that are shared (in a specific way) between facts in two or more data marts.

Top-down Design

The *top-down* approach is designed using a normalized enterprise data model. “Atomic” data, that is, data at the greatest level of detail, are stored in the data warehouse. Dimensional data marts containing data needed for specific business processes or specific departments are created from the data warehouse.

Hybrid Design

Data warehouses (DW) often resemble the hub and spokes architecture. Legacy systems feeding the warehouse often include customer relationship management and enterprise resource planning, generating large amounts of data. To consolidate these various data models, and facilitate the extract transform load process, data warehouses often make use of an operational data store, the information from which is parsed into the actual DW. To reduce data redundancy, larger systems often store the data in a normalized way. Data marts for specific reports can then be built on top of the DW.

The DW database in a hybrid solution is kept on third normal form to eliminate data redundancy. A normal relational database, however, is not efficient for business intelligence reports where dimensional modelling is prevalent. Small data marts can shop for data from the consolidated warehouse and use the filtered, specific data for the fact tables and dimensions required. The DW provides a single source of information from which the data marts can read, providing a wide range of business information. The hybrid architecture allows a DW to be replaced with a master data management solution where operational, not static information could reside.

The Data Vault Modeling components follow hub and spokes architecture. This modeling style is a hybrid design, consisting of the best practices from both third normal form and star schema. The Data Vault model is not a true third normal form, and breaks some of its rules, but it is a top-down architecture with a bottom up design. The Data Vault model is geared to be strictly a data warehouse. It is not geared to be end-user accessible, which when built, still requires the use of a data mart or star schema based release area for business purposes.

Versus Operational System

Operational systems are optimized for preservation of data integrity and speed of recording of business transactions through use of database normalization and an entity-relationship model. Operational system designers generally follow the Codd rules of database normalization in order to ensure data integrity. Codd defined five increasingly stringent rules of normalization. Fully normalized database designs (that is, those satisfying all five Codd rules) often result in information from a business transaction being stored in dozens to hundreds of tables. Relational databases are efficient at managing the relationships between these tables. The databases have very fast insert/update performance because only a small amount of data in those tables is affected each time a transaction is processed. Finally, in order to improve performance, older data are usually periodically purged from operational systems.

Data warehouses are optimized for analytic access patterns. Analytic access patterns generally involve selecting specific fields and rarely if ever 'select *' as is more common in operational databases. Because of these differences in access patterns, operational databases (loosely, OLTP) benefit from the use of a row-oriented DBMS whereas analytics databases (loosely, OLAP) benefit from the use of a column-oriented DBMS. Unlike operational systems which maintain a snapshot of the business, data warehouses generally maintain an infinite history which is implemented through ETL processes that periodically migrate data from the operational systems over to the data warehouse.

Evolution in Organization Use

These terms refer to the level of sophistication of a data warehouse:

Offline operational data warehouse

> Data warehouses in this stage of evolution are updated on a regular time cycle (usually daily, weekly or monthly) from the operational systems and the data is stored in an integrated reporting-oriented data

Offline data warehouse

> Data warehouses at this stage are updated from data in the operational systems on a regular basis and the data warehouse data are stored in a data structure designed to facilitate reporting.

On time data warehouse

> Online Integrated Data Warehousing represent the real time Data warehouses stage data in the warehouse is updated for every transaction performed on the source data

Integrated data warehouse

> These data warehouses assemble data from different areas of business, so users can look up the information they need across other systems.

Data Mart

A data mart is the access layer of the data warehouse environment that is used to get data out to the users. The data mart is a subset of the data warehouse and is usually oriented to a specific business line or team. Whereas data warehouses have an enterprise-wide depth, the information in data marts pertains to a single department. In some deployments, each department or business unit is considered the *owner* of its data mart including all the *hardware*, *software* and *data*. This enables each department to isolate the use, manipulation and development of their data. In other deployments where conformed dimensions are used, this business unit ownership will not hold true for shared dimensions like customer, product, etc.

Organizations build data warehouses and data marts because the information in the database is not organized in a way that makes it readily accessible, requiring queries that are too complicated or resource-consuming.

While transactional databases are designed to be updated, data warehouses or marts are read only. Data warehouses are designed to access large groups of related records. Data marts improve end-user response time by allowing users to have access to the specific type of data they need to view most often by providing the data in a way that supports the collective view of a group of users.

A data mart is basically a condensed and more focused version of a data warehouse that reflects the regulations and process specifications of each business unit within an organization. Each data mart is dedicated to a specific business function or region. This subset of data may span across many or all of an enterprise's functional subject areas. It is common for multiple data marts to be used in order to serve the needs of each individual business unit (different data marts can be used to obtain specific information for various enterprise departments, such as accounting, marketing, sales, etc.).

The related term spreadmart is a derogatory label describing the situation that occurs when one or more business analysts develop a system of linked spreadsheets to perform a business analysis, then grow it to a size and degree of complexity that makes it nearly impossible to maintain.

Data mart vs data warehouse

Data warehouse:

- Holds multiple subject areas
- Holds very detailed information
- Works to integrate all data sources
- Does not necessarily use a dimensional model but feeds dimensional models.

Data mart:

- Often holds only one subject area- for example, Finance, or Sales
- May hold more summarized data (although many hold full detail)
- Concentrates on integrating information from a given subject area or set of source systems
- Is built focused on a dimensional model using a star schema.

Design Schemas

- Star schema - fairly popular design choice; enables a relational database to emulate the analytical functionality of a multidimensional database
- Snowflake schema

Reasons for Creating a Data Mart

- Easy access to frequently needed data
- Creates collective view by a group of users
- Improves end-user response time
- Ease of creation
- Lower cost than implementing a full data warehouse
- Potential users are more clearly defined than in a full data warehouse
- Contains only business essential data and is less cluttered.

Dependent Data Mart

According to the Inmon school of data warehousing, a dependent data mart is a logical subset (view) or a physical subset (extract) of a larger data warehouse, isolated for one of the following reasons:

- A need refreshment for a special data model or schema: e.g., to restructure for OLAP
- Performance: to offload the data mart to a separate computer for greater efficiency or to eliminate the need to manage that workload on the centralized data warehouse.
- Security: to separate an authorized data subset selectively
- Expediency: to bypass the data governance and authorizations required to incorporate a new application on the Enterprise Data Warehouse
- Proving Ground: to demonstrate the viability and ROI (return on investment) potential of an application prior to migrating it to the Enterprise Data Warehouse
- Politics: a coping strategy for IT (Information Technology) in situations where a user group has more influence than funding or is not a good citizen on the centralized data warehouse.
- Politics: a coping strategy for consumers of data in situations where a data warehouse team is unable to create a usable data warehouse.

According to the Inmon school of data warehousing, tradeoffs inherent with data marts include limited scalability, duplication of data, data inconsistency with other silos of information, and inability to leverage enterprise sources of data.

The alternative school of data warehousing is that of Ralph Kimball. In his view, a data warehouse is nothing more than the union of all the data marts. This view helps to reduce costs and provides fast development, but can create an inconsistent data warehouse, especially in large organizations. Therefore, Kimball's approach is more suitable for small-to-medium corporations.

Master Data Management

In business, master data management (MDM) comprises the processes, governance, policies, standards and tools that consistently define and manage the critical data of an organization to provide a single point of reference.

The data that is mastered may include:

- reference data – the business objects for transactions, and the dimensions for analysis
- analytical data – supports decision making

In computing, a master data management tool can be used to support master data management by removing duplicates, standardizing data (mass maintaining), and incorporating rules to eliminate incorrect data from entering the system in order to create an authoritative source of master data. Master data are the products, accounts and parties for which the business transactions are completed. The root cause problem stems from business unit and product line segmentation, in which the same customer will be serviced by different product lines, with redundant data being entered about the customer (aka party in the role of customer) and account in order to process the transaction. The redundancy of party and account data is compounded in the front to back office life cycle, where the authoritative single source for the party, account and product data is needed but is often once again redundantly entered or augmented.

Master data management has the objective of providing processes for collecting, aggregating, matching, consolidating, quality-assuring, persisting and distributing such data throughout an organization to ensure consistency and control in the ongoing maintenance and application use of this information.

The term recalls the concept of a *master file* from an earlier computing era.

Definition

Master data management (MDM) is a comprehensive method of enabling an enterprise to link all of its critical data to one file, called a master file, that provides a common point of reference. When properly done, master data management streamlines data sharing among personnel and departments. In addition, master data management can facilitate computing in multiple system architectures, platforms and applications.

At its core Master Data Management (MDM) can be viewed as a “discipline for specialized quality improvement” defined by the policies and procedures put in place by a data governance organization. The ultimate goal being to provide the end user community with a “trusted single version of the truth” from which to base decisions.

Issues

At a basic level, master data management seeks to ensure that an organization does not use multiple (potentially inconsistent) versions of the same master data in different parts of its operations, which can occur in large organizations. A typical example of poor master data management is the scenario of a bank at which a customer has taken out a mortgage and the bank begins to send mortgage solicitations to that customer, ignoring the fact that the person already has a mortgage account relationship with the bank. This happens because the customer information used by the marketing section within the bank lacks integration with the customer information used by the customer services section of the bank. Thus the two groups remain unaware that an existing customer is also considered a sales lead. The process of record linkage is used to associate different records that correspond to the same entity, in this case the same person.

Other problems include (for example) issues with the quality of data, consistent classification and identification of data, and data-reconciliation issues. Master data management of disparate data systems requires data transformations as the data extracted from the disparate source data system is transformed and loaded into the master data management hub. To synchronize the disparate source master data, the managed master data extracted from the master data management hub is again transformed and loaded into the disparate source data system as the master data is updated. As with other Extract, Transform, Load-based data movement, these processes are expensive and inefficient to develop and to maintain which greatly reduces the return on investment for the master data management product.

One of the most common reasons some large corporations experience massive issues with master data management is growth through mergers or acquisitions. Any organizations which merge will typically create an entity with duplicate master data (since each likely had at least one master database of its own prior to the merger). Ideally, database administrators resolve this problem through deduplication of the master data as part of the merger. In practice, however, reconciling several master data systems can present difficulties because of the dependencies that existing applications have on the master databases. As a result, more often than not the two systems do not fully merge, but remain separate, with a special reconciliation process defined that ensures consistency between the data stored in the two systems. Over time, however, as further mergers and acquisitions occur, the problem multiplies, more and more master databases appear, and data-reconciliation processes become extremely complex, and consequently unmanageable and unreliable. Because of this trend, one can find organizations with 10, 15, or even as many as 100 separate, poorly integrated master databases, which can cause serious operational problems in the areas of customer satisfaction, operational efficiency, decision support, and regulatory compliance.

Solutions

Processes commonly seen in master data management include source identification, data collection, data transformation, normalization, rule administration, error detection and correction, data consolidation, data storage, data distribution, data classification, taxonomy services, item master creation, schema mapping, product codification, data enrichment and data governance.

The selection of entities considered for master data management depends somewhat on the nature of an organization. In the common case of commercial enterprises, master data management may apply to such entities as customer (customer data integration), product (product information management), employee, and vendor. Master data management processes identify the sources from which to collect descriptions of these entities. In the course of transformation and normalization, administrators adapt descriptions to conform to standard formats and data domains, making it possible to remove duplicate instances of any entity. Such processes generally result in an organizational master data management repository, from which all requests for a certain entity instance produce the same description, irrespective of the originating sources and the requesting destination.

The tools include data networks, file systems, a data warehouse, data marts, an operational data store, data mining, data analysis, data visualization, data federation and data virtualization. One of the newest tools, virtual master data management utilizes data virtualization and a persistent metadata server to implement a multi-level automated master data management hierarchy.

Transmission of Master Data

There are several ways in which master data may be collated and distributed to other systems. This includes:

- Data consolidation – The process of capturing master data from multiple sources and integrating into a single hub (operational data store) for replication to other destination systems.

- Data federation – The process of providing a single virtual view of master data from one or more sources to one or more destination systems.
- Data propagation – The process of copying master data from one system to another, typically through point-to-point interfaces in legacy systems.

Dimension (Data Warehouse)

A dimension is a structure that categorizes facts and measures in order to enable users to answer business questions. Commonly used dimensions are people, products, place and time.

In a data warehouse, dimensions provide structured labeling information to otherwise unordered numeric measures. The dimension is a data set composed of individual, non-overlapping data elements. The primary functions of dimensions are threefold: to provide filtering, grouping and labelling.

These functions are often described as “slice and dice”. Slicing refers to filtering data. Dicing refers to grouping data. A common data warehouse example involves sales as the measure, with customer and product as dimensions. In each sale a customer buys a product. The data can be sliced by removing all customers except for a group under study, and then diced by grouping by product.

A dimensional data element is similar to a categorical variable in statistics.

Typically dimensions in a data warehouse are organized internally into one or more hierarchies. “Date” is a common dimension, with several possible hierarchies:

- “Days (are grouped into) Months (which are grouped into) Years”,
- “Days (are grouped into) Weeks (which are grouped into) Years”
- “Days (are grouped into) Months (which are grouped into) Quarters (which are grouped into) Years”
- etc.

Types

Conformed Dimension

A conformed dimension is a set of data attributes that have been physically referenced in multiple database tables using the same key value to refer to the same structure, attributes, domain values, definitions and concepts. A conformed dimension cuts across many facts.

Dimensions are conformed when they are either exactly the same (including keys) or one is a perfect subset of the other. Most important, the row headers produced in two different answer sets from the same conformed dimension(s) must be able to match perfectly.

Conformed dimensions are either identical or strict mathematical subsets of the most granular, detailed dimension. Dimension tables are not conformed if the attributes are labeled differently or contain different values. Conformed dimensions come in several different flavors. At the most basic level, conformed dimensions mean exactly the same thing with every possible fact table to which they are joined. The date dimension table connected to the sales facts is identical to the date dimension connected to the inventory facts.

Junk Dimension

A junk dimension is a convenient grouping of typically low-cardinality flags and indicators. By creating an abstract dimension, these flags and indicators are removed from the fact table while placing them into a useful dimensional framework. A Junk Dimension is a dimension table consisting of attributes that do not belong in the fact table or in any of the existing dimension tables. The nature of these attributes is usually text or various flags, e.g. non-generic comments or just simple yes/no or true/false indicators. These kinds of attributes are typically remaining when all the obvious dimensions in the business process have been identified and thus the designer is faced with the challenge of where to put these attributes that do not belong in the other dimensions.

One solution is to create a new dimension for each of the remaining attributes, but due to their nature, it could be necessary to create a vast number of new dimensions resulting in a fact table with a very large number of foreign keys. The designer could also decide to leave the remaining attributes in the fact table but this could make the row length of the table unnecessarily large if, for example, the attributes is a long text string.

The solution to this challenge is to identify all the attributes and then put them into one or several Junk Dimensions. One Junk Dimension can hold several true/false or yes/no indicators that have no correlation with each other, so it would be convenient to convert the indicators into a more describing attribute. An example would be an indicator about whether a package had arrived, instead of indicating this as “yes” or “no”, it would be converted into “arrived” or “pending” in the junk dimension. The designer can choose to build the dimension table so it ends up holding all the indicators occurring with every other indicator so that all combinations are covered. This sets up a fixed size for the table itself which would be 2^x rows, where x is the number of indicators. This solution is appropriate in situations where the designer would expect to encounter a lot of different combinations and where the possible combinations are limited to an acceptable level. In a situation where the number of indicators are large, thus creating a very big table or where the designer only expect to encounter a few of the possible combinations, it would be more appropriate to build each row in the junk dimension as new combinations are encountered. To limit the size of the tables, multiple junk dimensions might be appropriate in other situations depending on the correlation between various indicators.

Junk dimensions are also appropriate for placing attributes like non-generic comments from the fact table. Such attributes might consist of data from an optional comment field when a customer places an order and as a result will probably be blank in many cases. Therefore, the junk dimension should contain a single row representing the blanks as a surrogate key that will be used in the fact table for every row returned with a blank comment field

Degenerate Dimension

A degenerate dimension is a key, such as a transaction number, invoice number, ticket number, or bill-of-lading number, that has no attributes and hence does not join to an actual dimension table. Degenerate dimensions are very common when the grain of a fact table represents a single transaction item or line item because the degenerate dimension represents the unique identifier of the parent. Degenerate dimensions often play an integral role in the fact table's primary key.

Role-playing Dimension

Dimensions are often recycled for multiple applications within the same database. For instance, a "Date" dimension can be used for "Date of Sale", as well as "Date of Delivery", or "Date of Hire". This is often referred to as a "role-playing dimension".

Use of ISO Representation Terms

When referencing data from a metadata registry such as ISO/IEC 11179, representation terms such as Indicator (a boolean true/false value), Code (a set of non-overlapping enumerated values) are typically used as dimensions. For example, using the National Information Exchange Model (NIEM) the data element name would be PersonGenderCode and the enumerated values would be male, female and unknown.

Dimension Table

In data warehousing, a dimension table is one of the set of companion tables to a fact table.

The fact table contains business facts (or *measures*), and foreign keys which refer to candidate keys (normally primary keys) in the dimension tables.

Contrary to *fact* tables, *dimension* tables contain descriptive attributes (or fields) that are typically textual fields (or discrete numbers that behave like text). These attributes are designed to serve two critical purposes: query constraining and/or filtering, and query result set labeling.

Dimension attributes should be:

- Verbose (labels consisting of full words)
- Descriptive
- Complete (having no missing values)
- Discretely valued (having only one value per dimension table row)
- Quality assured (having no misspellings or impossible values)

Dimension table rows are uniquely identified by a single key field. It is recommended that the key field be a simple integer because a key value is meaningless, used only for joining fields between the fact and dimension tables. Dimension tables often use primary keys that are also surrogate keys. Surrogate keys are often auto-generated (e.g. a Sybase or SQL Server "identity column", a PostgreSQL or Informix serial, an Oracle SEQUENCE or a column defined with AUTO_INCREMENT in MySQL).

The use of surrogate dimension keys brings several advantages, including:

- Performance. Join processing is made much more efficient by using a single field (the surrogate key)
- Buffering from operational key management practices. This prevents situations where removed data rows might reappear when their natural keys get reused or reassigned after a long period of dormancy
- Mapping to integrate disparate sources
- Handling unknown or not-applicable connections
- Tracking changes in dimension attribute values

Although surrogate key use places a burden put on the ETL system, pipeline processing can be improved, and ETL tools have built-in improved surrogate key processing.

The goal of a dimension table is to create standardized, conformed dimensions that can be shared across the enterprise's data warehouse environment, and enable joining to multiple fact tables representing various business processes.

Conformed dimensions are important to the enterprise nature of DW/BI systems because they promote:

- Consistency. Every fact table is filtered consistently, so that query answers are labeled consistently.
- Integration. Queries can drill into different process fact tables separately for each individual fact table, then join the results on common dimension attributes.
- Reduced development time to market. The common dimensions are available without recreating them.

Over time, the attributes of a given row in a dimension table may change. For example, the shipping address for a company may change. Kimball refers to this phenomenon as Slowly Changing Dimensions. Strategies for dealing with this kind of change are divided into three categories:

- Type One. Simply overwrite the old value(s).
- Type Two. Add a new row containing the new value(s), and distinguish between the rows using Tuple-versioning techniques.
- Type Three. Add a new attribute to the existing row.

Common Patterns

Date and time

Since many fact tables in a data warehouse are time series of observations, one or more date dimensions are often needed. One of the reasons to have date dimensions is to place calendar knowledge in the data warehouse instead of hard coded in an application. While a simple SQL date/timestamp is useful for providing accurate information about the time a fact was recorded, it can not give information about holidays, fiscal periods, etc. An SQL date/timestamp can still be useful to store in the fact table, as it allows for precise calculations.

Having both the date and time of day in the same dimension, may easily result in a huge dimension with millions of rows. If a high amount of detail is needed it is usually a good idea to split date and time into two or more separate dimensions. A time dimension with a grain of seconds in a day will only have 86400 rows. A more or less detailed grain for date/time dimensions can be chosen depending on needs. As examples, date dimensions can be accurate to year, quarter, month or day and time dimensions can be accurate to hours, minutes or seconds.

As a rule of thumb, time of day dimension should only be created if hierarchical groupings are needed or if there are meaningful textual descriptions for periods of time within the day (ex. "evening rush" or "first shift").

If the rows in a fact table are coming from several timezones, it might be useful to store date and time in both local time and a standard time. This can be done by having two dimensions for each date/time dimension needed – one for local time, and one for standard time. Storing date/time in both local and standard time, will allow for analysis on when facts are created in a local setting and in a global setting as well. The standard time chosen can be a global standard time (ex. UTC), it can be the local time of the business' headquarter, or any other time zone that would make sense to use.

Slowly Changing Dimension

Dimensions in data management and data warehousing contain relatively static data about such entities as geographical locations, customers, or products. Data captured by Slowly Changing Dimensions (SCDs) change slowly but unpredictably, rather than according to a regular schedule.

Some scenarios can cause Referential integrity problems.

For example, a database may contain a fact table that stores sales records. This fact table would be linked to dimensions by means of foreign keys. One of these dimensions may contain data about the company's salespeople: e.g., the regional offices in which they work. However, the salespeople are sometimes transferred from one regional office to another. For historical sales reporting purposes it may be necessary to keep a record of the fact that a particular sales person had been assigned to a particular regional office at an earlier date, whereas that sales person is now assigned to a different regional office.

Dealing with these issues involves SCD management methodologies referred to as Type 0 through 6. Type 6 SCDs are also sometimes called Hybrid SCDs.

Type 0: Retain Original

The Type 0 method is passive. It manages dimensional changes and no action is performed. Values remain as they were at the time the dimension record was first inserted. In certain circumstances history is preserved with a Type 0. Higher order types are employed to guarantee the preservation of history whereas Type 0 provides the least or no control. This is rarely used.

Type 1: Overwrite

This methodology overwrites old with new data, and therefore does not track historical data.

Example of a supplier table:

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	CA

In the above example, Supplier_Code is the natural key and Supplier_Key is a surrogate key. Technically, the surrogate key is not necessary, since the row will be unique by the natural key (Supplier_Code). However, to optimize performance on joins use integer rather than character keys (unless the number of bytes in the character key is less than the number of bytes in the integer key).

If the supplier relocates the headquarters to Illinois the record would be overwritten:

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State
123	ABC	Acme Supply Co	IL

The disadvantage of the Type 1 method is that there is no history in the data warehouse. It has the advantage however that it's easy to maintain.

If one has calculated an aggregate table summarizing facts by state, it will need to be recalculated when the Supplier_State is changed.

Type 2: Add New Row

This method tracks historical data by creating multiple records for a given natural key in the dimensional tables with separate surrogate keys and/or different version numbers. Unlimited history is preserved for each insert.

For example, if the supplier relocates to Illinois the version numbers will be incremented sequentially:

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Version.
123	ABC	Acme Supply Co	CA	0
124	ABC	Acme Supply Co	IL	1

Another method is to add 'effective date' columns.

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date
123	ABC	Acme Supply Co	CA	01-Jan-2000	21-Dec-2004
124	ABC	Acme Supply Co	IL	22-Dec-2004	NULL

The null End_Date in row two indicates the current tuple version. In some cases, a standardized surrogate high date (e.g. 9999-12-31) may be used as an end date, so that the field can be included in an index, and so that null-value substitution is not required when querying.

Transactions that reference a particular surrogate key (Supplier_Key) are then permanently bound to the time slices defined by that row of the slowly changing dimension table. An aggregate table summarizing facts by state continues to reflect the historical state, i.e. the state the supplier was in at the time of the transaction; no update is needed. To reference the entity via the natural key, it is necessary to remove the unique constraint making Referential integrity by DBMS impossible.

If there are retroactive changes made to the contents of the dimension, or if new attributes are added to the dimension (for example a Sales_Rep column) which have different effective dates from those already defined, then this can result in the existing transactions needing to be updated to reflect the new situation. This can be an expensive database operation, so Type 2 SCDs are not a good choice if the dimensional model is subject to change.

Type 3: Add New Attribute

This method tracks changes using separate columns and preserves limited history. The Type 3 preserves limited history as it is limited to the number of columns designated for storing historical data. The original table structure in Type 1 and Type 2 is the same but Type 3 adds additional columns. In the following example, an additional column has been added to the table to record the supplier's original state - only the previous history is stored.

Supplier_ Key	Supplier_ Code	Supplier_ Name	Original_Supplier_State	Effective_ Date	Current_Supplier_ State
123	ABC	Acme Supply Co	CA	22-Dec-2004	IL

This record contains a column for the original state and current state—cannot track the changes if the supplier relocates a second time.

One variation of this is to create the field Previous_Supplier_State instead of Original_Supplier_State which would track only the most recent historical change.

Type 4: Add History Table

The Type 4 method is usually referred to as using "history tables", where one table keeps the current data, and an additional table is used to keep a record of some or all changes. Both the surrogate keys are referenced in the Fact table to enhance query performance.

For the above example the original table name is Supplier and the history table is Supplier_History.

Supplier			
Supplier_key	Supplier_Code	Supplier_Name	Supplier_State
124	ABC	Acme & Johnson Supply Co	IL

Supplier_History				
Supplier_key	Supplier_Code	Supplier_Name	Supplier_State	Create_Date
123	ABC	Acme Supply Co	CA	14-June-2003
124	ABC	Acme & Johnson Supply Co	IL	22-Dec-2004

This method resembles how database audit tables and change data capture techniques function.

Type 6: Hybrid

The Type 6 method combines the approaches of types 1, 2 and 3 (1 + 2 + 3 = 6). One possible explanation of the origin of the term was that it was coined by Ralph Kimball during a conversation with Stephen Pace from Kalido. Ralph Kimball calls this method "Unpredictable Changes with Single-Version Overlay" in *The Data Warehouse Toolkit*.

The Supplier table starts out with one record for our example supplier:

Supplier_ Key	Row_ Key	Supplier_ Code	Supplier_ Name	Current_ State	Historical_ State	Start_ Date	End_ Date	Current_ Flag
123	1	ABC	Acme Supply Co	CA	CA	01-Jan-2000	31-Dec-2009	Y

The Current_State and the Historical_State are the same. The optional Current_Flag attribute indicates that this is the current or most recent record for this supplier.

When Acme Supply Company moves to Illinois, we add a new record, as in Type 2 processing, however a row key is included to ensure we have a unique key for each row:

Supplier_ Key	Row_ Key	Supplier_ Code	Supplier_ Name	Current_ State	Historical_ State	Start_ Date	End_ Date	Current_ Flag
123	1	ABC	Acme Supply Co	IL	CA	01-Jan-2000	21-Dec-2004	N
123	2	ABC	Acme Supply Co	IL	IL	22-Dec-2004	31-Dec-2009	Y

We overwrite the Current_Flag information in the first record (Row_Key = 1) with the new information, as in Type 1 processing. We create a new record to track the changes, as in Type 2 processing. And we store the history in a second State column (Historical_State), which incorporates Type 3 processing.

For example, if the supplier were to relocate again, we would add another record to the Supplier dimension, and we would overwrite the contents of the Current_State column:

Supplier_ Key	Row_ Key	Supplier_ Code	Supplier_ Name	Current_ State	Historical_ State	Start_ Date	End_ Date	Current_ Flag
123	1	ABC	Acme Supply Co	NY	CA	01-Jan-2000	21-Dec-2004	N
123	2	ABC	Acme Supply Co	NY	IL	22-Dec-2004	03-Feb-2008	N
123	3	ABC	Acme Supply Co	NY	NY	04-Feb-2008	31-Dec-2009	Y

Note that, for the current record (Current_Flag = 'Y'), the Current_State and the Historical_State are always the same.

Type 2 / type 6 Fact Implementation

Type 2 Surrogate Key with Type 3 Attribute

In many Type 2 and Type 6 SCD implementations, the surrogate key from the dimension is put into the fact table in place of the natural key when the fact data is loaded into the data repository. The surrogate key is selected for a given fact record based on its effective date and the Start_Date and End_Date from the dimension table. This allows the fact data to be easily joined to the correct dimension data for the corresponding effective date.

Here is the Supplier table as we created it above using Type 6 Hybrid methodology:

Supplier_ Key	Supplier_ Code	Supplier_ Name	Current_ State	Historical_ State	Start_ Date	End_ Date	Current_ Flag
123	ABC	Acme Supply Co	NY	CA	01-Jan-2000	21-Dec-2004	N
124	ABC	Acme Supply Co	NY	IL	22-Dec-2004	03-Feb-2008	N
125	ABC	Acme Supply Co	NY	NY	04-Feb-2008	31-Dec-9999	Y

Once the Delivery table contains the correct Supplier_Key, it can easily be joined to the Supplier table using that key. The following SQL retrieves, for each fact record, the current supplier state and the state the supplier was located in at the time of the delivery:

```
SELECT
  delivery.delivery_cost,
  supplier.supplier_name,
  supplier.historical_state,
  supplier.current_state
FROM delivery
INNER JOIN supplier
  ON delivery.supplier_key = supplier.supplier_key
```

Pure type 6 Implementation

Having a Type 2 surrogate key for each time slice can cause problems if the dimension is subject to change.

A pure Type 6 implementation does not use this, but uses a Surrogate Key for each master data item (e.g. each unique supplier has a single surrogate key).

This avoids any changes in the master data having an impact on the existing transaction data.

It also allows more options when querying the transactions.

Here is the Supplier table using the pure Type 6 methodology:

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date
456	ABC	Acme Supply Co	CA	01-Jan-2000	21-Dec-2004
456	ABC	Acme Supply Co	IL	22-Dec-2004	03-Feb-2008
456	ABC	Acme Supply Co	NY	04-Feb-2008	31-Dec-9999

The following example shows how the query must be extended to ensure a single supplier record is retrieved for each transaction.

```
SELECT
  supplier.supplier_code,
  supplier.supplier_state
FROM supplier
INNER JOIN delivery
  ON supplier.supplier_key = delivery.supplier_key
 AND delivery.delivery_date BETWEEN supplier.start_date AND supplier.end_date
```

A fact record with an effective date (Delivery_Date) of August 9, 2001 will be linked to Supplier_Code of ABC, with a Supplier_State of 'CA'. A fact record with an effective date of October 11, 2007 will also be linked to the same Supplier_Code ABC, but with a Supplier_State of 'IL'.

Whilst more complex, there are a number of advantages of this approach, including:

1. Referential integrity by DBMS is now possible, but one cannot use Supplier_Code as foreign key on Product table and using Supplier_Key as foreign key each product is tied on specific time slice.
2. If there is more than one date on the fact (e.g. Order Date, Delivery Date, Invoice Payment Date) one can choose which date to use for a query.
3. You can do "as at now", "as at transaction time" or "as at a point in time" queries by changing the date filter logic.
4. You don't need to reprocess the Fact table if there is a change in the dimension table (e.g. adding additional fields retrospectively which change the time slices, or if one makes a mistake in the dates on the dimension table one can correct them easily).
5. You can introduce bi-temporal dates in the dimension table.
6. You can join the fact to the multiple versions of the dimension table to allow reporting of the same information with different effective dates, in the same query.

The following example shows how a specific date such as '2012-01-01 00:00:00' (which could be the current datetime) can be used.

```
SELECT
  supplier.supplier_code,
  supplier.supplier_state
FROM supplier
INNER JOIN delivery
  ON supplier.supplier_key = delivery.supplier_key
 AND '2012-01-01 00:00:00' BETWEEN supplier.start_date AND supplier.end_date
```

Both Surrogate and Natural Key

An alternative implementation is to place *both* the surrogate key and the natural key into the fact table. This allows the user to select the appropriate dimension records based on:

- the primary effective date on the fact record (above),
- the most recent or current information,
- any other date associated with the fact record.

This method allows more flexible links to the dimension, even if one has used the Type 2 approach instead of Type 6.

Here is the Supplier table as we might have created it using Type 2 methodology:

Supplier_Key	Supplier_Code	Supplier_Name	Supplier_State	Start_Date	End_Date	Current_Flag
123	ABC	Acme Supply Co	CA	01-Jan-2000	21-Dec-2004	N
124	ABC	Acme Supply Co	IL	22-Dec-2004	03-Feb-2008	N
125	ABC	Acme Supply Co	NY	04-Feb-2008	31-Dec-9999	Y

The following SQL retrieves the most current Supplier_Name and Supplier_State for each fact record:

```
SELECT
  delivery.delivery_cost,
  supplier.supplier_name,
  supplier.supplier_state
FROM delivery
```

```
INNER JOIN supplier
  ON delivery.supplier_code = supplier.supplier_code
WHERE supplier.current_flag = 'Y'
```

If there are multiple dates on the fact record, the fact can be joined to the dimension using another date instead of the primary effective date. For instance, the Delivery table might have a primary effective date of Delivery_Date, but might also have an Order_Date associated with each record.

The following SQL retrieves the correct Supplier_Name and Supplier_State for each fact record based on the Order_Date:

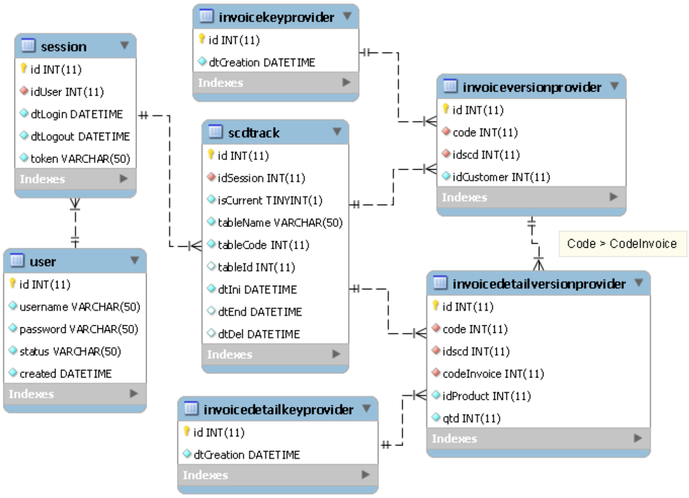
```
SELECT
  delivery.delivery_cost,
  supplier.supplier_name,
  supplier.supplier_state
FROM delivery
INNER JOIN supplier
  ON delivery.supplier_code = supplier.supplier_code
 AND delivery.order_date BETWEEN supplier.start_date AND supplier.end_date
```

Some cautions:

- Referential integrity by DBMS is not possible since there is not a unique to create the relationship.
- If relationship is made with surrogate to solve problem above then one ends with entity tied to a specific time slice.
- If the join query is not written correctly, it may return duplicate rows and/or give incorrect answers.
- The date comparison might not perform well.
- Some Business Intelligence tools do not handle generating complex joins well.
- The ETL processes needed to create the dimension table needs to be carefully designed to ensure that there are no overlaps in the time periods for each distinct item of reference data.
- Many of problems above can be solved using the mixed diagram of an scd model below.

Combining Types

Different SCD Types can be applied to different columns of a table. For example, we can apply Type 1 to the Supplier_Name column and Type 2 to the Supplier_State column of the same table.


invoicekeyprovider
id INT(11)
dtCreation DATETIME
Indexes
session
id INT(11)
idUser INT(11)
dtLogin DATETIME
dtLogout DATETIME
token VARCHAR(50)
Indexes
invoiceversionprovider
id INT(11)
code INT(11)
idscd INT(11)
idCustomer INT(11)
Indexes
scdtrack
id INT(11)
idSession INT(11)
isCurrent TINYINT(1)
tableName VARCHAR(50)
tableCode INT(11)
tableId INT(11)
dtIni DATETIME
dtEnd DATETIME
dtDel DATETIME
Indexes
user
id INT(11)
username VARCHAR(50)
password VARCHAR(50)
status VARCHAR(50)
created DATETIME
Indexes
Code > CodeInvoice
invoicedetailversionprovider
id INT(11)
code INT(11)
idscd INT(11)
codeInvoice INT(11)
idProduct INT(11)
qtd INT(11)
Indexes
invoicedetailkeyprovider
id INT(11)
dtCreation DATETIME
Indexes

Scd model

Data Vault Modeling

Data vault modeling is a database modeling method that is designed to provide long-term historical storage of data coming in from multiple operational systems. It is also a method of looking at historical data that deals with issues such as auditing, tracing of data, loading speed and resilience to change as well as emphasizing the need to trace where all the data in the database came from. This means that every row in a data vault must be accompanied by record source and load date attributes, enabling an auditor to trace values back to the source.

Data vault modeling makes no distinction between good and bad data ("bad" meaning not conforming to business rules). This is summarized in the statement that a data vault stores "a single version of the facts" (also expressed by Dan Linstedt as "all the data, all of the time") as opposed to the practice in other data warehouse methods of storing "a single version of the truth" where data that does not conform to the definitions is removed or "cleansed".

The modeling method is designed to be resilient to change in the business environment where the data being stored is coming from, by explicitly separating structural information from descriptive attributes. Data vault is designed to enable parallel loading as much as possible, so that very large implementations can scale out without the need for major redesign.

History and Philosophy

In data warehouse modeling there are two well-known competing options for modeling the layer where the data are stored. Either you model according to Ralph Kimball, with conformed dimensions and an enterprise data bus, or you model according to Bill Inmon with the database normalized. Both techniques have issues when dealing with changes in the systems feeding the data warehouse. For conformed dimensions you also have to cleanse data (to conform it) and this is undesirable in a number of cases since this inevitably will lose information. Data vault is designed to avoid or minimize the impact of those issues, by moving them to areas of the data warehouse that are outside the historical storage area (cleansing is done in the data marts) and by separating the structural items (business keys and the associations between the business keys) from the descriptive attributes.

Dan Linstedt, the creator of the method, describes the resulting database as follows:

The Data Vault Model is a detail oriented, historical tracking and uniquely linked set of normalized tables that support one or more functional areas of business. It is a hybrid approach encompassing the best of breed between 3rd normal form (3NF) and star schema. The design is flexible, scalable, consistent and adaptable to the needs of the enterprise

Data vault's philosophy is that all data are relevant data, even if it is not in line with established definitions and business rules. If data are not conforming to these definitions and rules then that is a problem for the business, not the data warehouse. The determination of data being "wrong" is an interpretation of the data that stems from a particular point of view that may not be valid for everyone, or at every point in time. Therefore the data vault must capture all data and only when reporting or extracting data from the data vault is the data being interpreted.

Another issue to which data vault is a response is that more and more there is a need for complete auditability and traceability of all the data in the data warehouse. Due to Sarbanes-Oxley requirements in the USA and similar measures in Europe this is a relevant topic for many business intelligence implementations, hence the focus of any data vault implementation is complete traceability and auditability of all information.

Data Vault 2.0 is the new specification, it is an open standard. The new specification contains components which define the implementation best practices, the methodology (SEI/CMMI, Six Sigma, SDLC, etc..), the architecture, and the model. *Data Vault 2.0* has a focus on including new components such as Big Data, NoSQL - and also focuses on performance of the existing model. The old specification (documented here for the most part) is highly focused on data vault modeling. It is documented in the book: Building a Scalable Data Warehouse with Data Vault 2.0.

It is necessary to evolve the specification to include the new components, along with the best practices in order to keep the EDW and BI systems current with the needs and desires of today's businesses.

History

Data vault modeling was originally conceived by Dan Linstedt in 1990 and was released in 2000 as a public domain modeling method. In a series of five articles on The Data Administration Newsletter the basic rules of the Data Vault method are expanded and explained. These contain a general overview, an overview of the components, a discussion about end dates and joins, link tables, and an article on loading practices.

An alternative (and seldom used) name for the method is "Common Foundational Integration Modelling Architecture."

Data Vault 2.0 has arrived on the scene as of 2013 and brings to the table Big Data, NoSQL, unstructured, semi-structured seamless integration, along with methodology, architecture, and implementation best practices.

Alternative Interpretations

According to Dan Linstedt, the Data Model is inspired by (or patterned off) a simplistic view of neurons, dendrites, and synapses – where neurons are associated with Hubs and Hub Satellites, Links are dendrites (vectors of information), and other Links are synapses (vectors in the opposite direction). By using a data mining set of algorithms, links can be scored with confidence and strength ratings. They can be created and dropped on the fly in accordance with learning about relationships that currently don't exist. The model can be automatically morphed, adapted, and adjusted as it is used and fed new structures.

Another view is that a data vault model provides an ontology of the Enterprise in the sense that it describes the terms in the domain of the enterprise (Hubs) and the relationships among them (Links), adding descriptive attributes (Satellites) where necessary.

Another way to think of a data vault model is as a graph model. The data vault model actually provides a "graph based" model with hubs and relationships in a relational database world. In this manner, the developer can use SQL to get at graph based relationships with sub-second responses.

Basic Notions

Data vault attempts to solve the problem of dealing with change in the environment by separating the business keys (that do not mutate as often, because they uniquely identify a business entity) and the associations between those business keys, from the descriptive attributes of those keys.

The business keys and their associations are structural attributes, forming the skeleton of the data model. The data vault method has as one of its main axioms that real business keys only change when the business changes and are therefore the most stable elements from which to derive the structure of a historical database. If you use these keys as the backbone of a data warehouse, you can organize the rest of the data around them. This means that choosing the correct keys for the hubs is of prime importance for the stability of your model. The keys are stored in tables with a few constraints on the structure. These key-tables are called hubs.

Hubs

Hubs contain a list of unique business keys with low propensity to change. Hubs also contain a surrogate key for each Hub item and metadata describing the origin of the business key. The descriptive attributes for the information on the Hub (such as the description for the key, possibly in multiple languages) are stored in structures called Satellite tables which will be discussed below.

The Hub contains at least the following fields:

- a surrogate key, used to connect the other structures to this table.
- a business key, the driver for this hub. The business key can consist of multiple fields.
- the record source, which can be used to see what system loaded each business key first.

optionally, you can also have metadata fields with information about manual updates (user/time) and the extraction date.

A hub is not allowed to contain multiple business keys, except when two systems deliver the same business key but with collisions that have different meanings.

Hubs should normally have at least one satellite.

Hub Example

This is an example for a hub-table containing cars, called "Car" (H_CAR). The driving key is vehicle identification number.

Fieldname	Description	Mandatory?	Comment
H_CAR_ID	Sequence ID and surrogate key for the hub	No	Recommended but optional
VEHICLE_ID_NR	The business key that drives this hub. Can be more than one field for a composite business key	Yes	
H_RSRC	The recordsource of this key when first loaded	Yes	
LOAD_AUDIT_ID	An ID into a table with audit information, such as load time, duration of load, number of lines, etc.	No	

Links

Associations or transactions between business keys (relating for instance the hubs for customer and product with each other through the purchase transaction) are modeled using link tables. These tables are basically many-to-many join tables, with some metadata.

Links can link to other links, to deal with changes in granularity (for instance, adding a new key to a database table would change the grain of the database table). For instance, if you have an association between customer and address, you could add a reference to a link between the hubs for product and transport company. This could be a link called "Delivery". Referencing a link in another link is considered a bad practice, because it introduces dependencies between links that make parallel loading more difficult. Since a link to another link is the same as a new link with the hubs from the other link, in these cases creating the links without referencing other links is the preferred solution.

Links sometimes link hubs to information that is not by itself enough to construct a hub. This occurs when one of the business keys associated by the link is not a real business key. As an example, take an order form with "order number" as key, and order lines that are keyed with a semi-random number to make them unique. Let's say, "unique number". The latter key is not a real business key, so it is no hub. However, we do need to use it in order to guarantee the correct granularity for the link. In this case, we do not use a hub with surrogate key, but add the business key "unique number" itself to the link. This is done only when there is no possibility of ever using the business key for another link or as key for attributes in a satellite. This construct has been called a 'peg-legged link' by Dan Linstedt on his (now defunct) forum.

Links contain the surrogate keys for the hubs that are linked, their own surrogate key for the link and metadata describing the origin of the association. The descriptive attributes for the information on the association (such as the time, price or amount) are stored in structures called *satellite tables* which are discussed below.

Link Example

This is an example for a link-table between two hubs for cars (H_CAR) and persons (H_PERSON). The link is called "Driver" (L_DRIVER).

Fieldname	Description	Mandatory?	Comment
L_DRIVER_ID	Sequence ID and surrogate key for the Link	No	Recommended but optional
H_CAR_ID	surrogate key for the car hub, the first anchor of the link	Yes	
H_PERSON_ID	surrogate key for the person hub, the second anchor of the link	Yes	
L_RSRC	The recordsource of this association when first loaded	Yes	
LOAD_AUDIT_ID	An ID into a table with audit information, such as load time, duration of load, number of lines, etc.	No	

Satellites

The hubs and links form the structure of the model, but have no temporal attributes and hold no descriptive attributes. These are stored in separate tables called *satellites*. These consist of metadata linking them to their parent hub or link, metadata describing the origin of the association and attributes, as well as a timeline with start and end dates for the attribute. Where the hubs and links provide the structure of the model, the satellites provide the "meat" of the model, the context for the business processes that are captured in hubs and links. These attributes are stored both with regards to the details of the matter as well as the timeline and can range from quite complex (all of the fields describing a clients complete profile) to quite simple (a satellite on a link with only a valid-indicator and a timeline).

Usually the attributes are grouped in satellites by source system. However, descriptive attributes such as size, cost, speed, amount or color can change at different rates, so you can also split these attributes up in different satellites based on their rate of change.

All the tables contain metadata, minimally describing at least the source system and the date on which this entry became valid, giving a complete historical view of the data as it enters the data warehouse.

Satellite Example

This is an example for a satellite on the drivers-link between the hubs for cars and persons, called "Driver insurance" (S_DRIVER_INSURANCE). This satellite contains attributes that are specific to the insurance of the relationship between the car and the person driving it, for instance an indicator whether this is the primary driver, the name of the insurance company for this car and person (could also be a separate hub) and a summary of the number of accidents involving this combination of vehicle and driver. Also included is a reference to a lookup- or reference table called R_RISK_CATEGORY containing the codes for the risk category in which this relationship is deemed to fall.

Fieldname	Description	Mandatory?	Comment
S_DRIVER_INSURANCE_ID	Sequence ID and surrogate key for the satellite on the link	No	Recommended but optional
L_DRIVER_ID	(surrogate) primary key for the driver link, the parent of the satellite	Yes	
S_SEQ_NR	Ordering or sequence number, to enforce uniqueness if there are several valid satellites for one parent key	No(**)	This can happen if, for instance, you have a hub COURSE and the name of the course is an attribute but in several different languages.
S_LDTS	Load Date (startdate) for the validity of this combination of attribute values for parent key L_DRIVER_ID	Yes	
S_LEDTS	Load End Date (enddate) for the validity of this combination of attribute values for parent key L_DRIVER_ID	No	
IND_PRIMARY_DRIVER	Indicator whether the driver is the primary driver for this car	No (*)	
INSURANCE_COMPANY	The name of the insurance company for this vehicle and this driver	No (*)	
NR_OF_ACCIDENTS	The number of accidents by this driver in this vehicle	No (*)	
R_RISK_CATEGORY_CD	The risk category for the driver. This is a reference to R_RISK_CATEGORY	No (*)	
S_RSRC	The recordsource of the information in this satellite when first loaded	Yes	
LOAD_AUDIT_ID	An ID into a table with audit information, such as load time, duration of load, number of lines, etc.	No	

(*) at least one attribute is mandatory. (**) sequence number becomes mandatory if it is needed to enforce uniqueness for multiple valid satellites on the same hub or link.

Reference Tables

Reference tables are a normal part of a healthy data vault model. They are there to prevent redundant storage of simple reference data that is referenced a lot. More formally, Dan Linstedt defines reference data as follows:

Any information deemed necessary to resolve descriptions from codes, or to translate keys in to (sic) a consistent manner. Many of these fields are "descriptive" in nature and ***describe*** *a specific state of the other more important information. As such, reference data lives in separate tables from the raw Data Vault tables.*

Reference tables are referenced from Satellites, but never bound with physical foreign keys. There is no prescribed structure for reference tables: use what works best in your specific case, ranging from simple lookup tables to small data vaults or even stars. They can be historical or have no history, but it is recommended that you stick to the natural keys and not create surrogate keys in that case. Normally, data vaults have a lot of reference tables, just like any other Data Warehouse.

Reference Example

This is an example of a reference table with risk categories for drivers of vehicles. It can be referenced from any satellite in the data vault. For now we reference it from satellite S_DRIVER_INSURANCE. The reference table is R_RISK_CATEGORY.

Fieldname	Description	Mandatory?
R_RISK_CATEGORY_CD	The code for the risk category	Yes
RISK_CATEGORY_DESC	A description of the risk category	No (*)

(*) at least one attribute is mandatory.

Loading Practices

The ETL for updating a data vault model is fairly straightforward. First you have to load all the hubs, creating surrogate IDs for any new business keys. Having done that, you can now resolve all business keys to surrogate ID's if you query the hub. The second step is to resolve the links between hubs and create surrogate IDs for any new associations. At the same time, you can also create all satellites that are attached to hubs, since you can resolve the key to a surrogate ID. Once you have created all the new links with their surrogate keys, you can add the satellites to all the links.

Since the hubs are not joined to each other except through links, you can load all the hubs in parallel. Since links are not attached directly to each other, you can load all the links in parallel as well. Since satellites can be attached only to hubs and links, you can also load these in parallel.

The ETL is quite straightforward and lends itself to easy automation or templating. Problems occur only with links relating to other links, because resolving the business keys in the link only leads to another link that has to be resolved as well. Due to the equivalence of this situation with a link to multiple hubs, this difficulty can be avoided by remodeling such cases and this is in fact the recommended practice.

Data are never deleted from the data vault, unless you have a technical error while loading data.

Data Vault and Dimensional Modelling

The data vault modelled layer is normally used to store data. It is not optimized for query performance, nor is it easy to query by the well-known query-tools such as Cognos, SAP Business Objects, Pentaho et al. Since these end-user computing tools expect or prefer their data to be contained in a dimensional model, a conversion is usually necessary.

For this purpose, the hubs and related satellites on those hubs can be considered as dimensions and the links and related satellites on those links can be viewed as fact tables in a dimensional model. This enables you to quickly prototype a dimensional model out of a data vault model using views. For performance reasons the dimensional model will usually be implemented in relational tables, after approval.

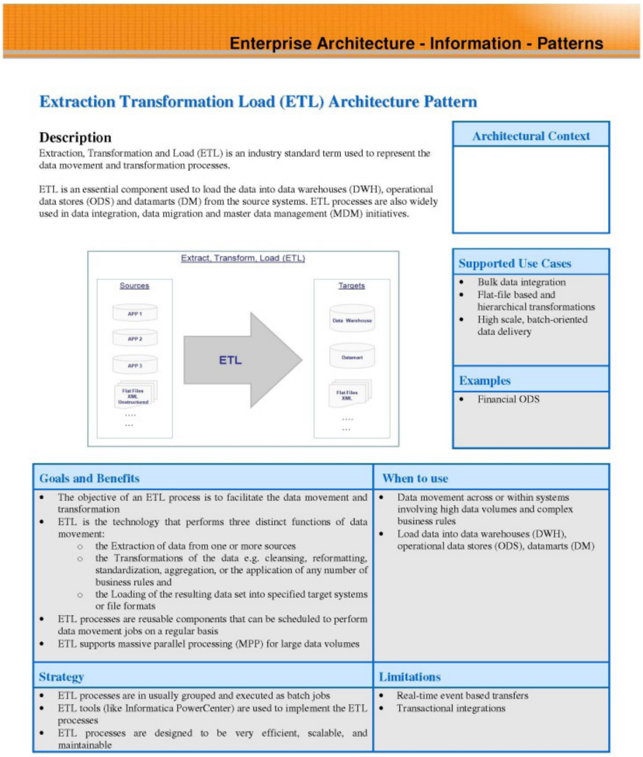
Note that while it is relatively straightforward to move data from a data vault model to a (cleansed) dimensional model, the reverse is not as easy.

Data Vault Methodology

The data vault methodology is based on SEI/CMMI Level 5 best practices. It includes multiple components of CMMI Level 5, and combines them with best practices from Six Sigma, TQM, and SDLC. Particularly, it is focused on Scott Ambler's agile methodology for build out and deployment. Data vault projects have a short, scope-controlled release cycle and should consist of a production release every 2 to 3 weeks.

Teams using the data vault methodology will automatically adopt to the repeatable, consistent, and measurable projects that are expected at CMMI Level 5. Data that flow through the EDW data vault system will begin to follow the TQM (total quality management) life-cycle that has long been missing from BI (business intelligence) projects.

Extract, Transform, Load


Extract, Transform, Load (ETL)
Sources
APP 1
APP 2
APP 3
Flat Files XML Unstructured
....
...
ETL
Targets
Data Warehouse
Datamart
Flat Files XML
....
...

Enterprise Architecture - Information - Patterns

Extraction Transformation Load (ETL) Architecture Pattern

Description

Extraction, Transformation and Load (ETL) is an industry standard term used to represent the data movement and transformation processes.

ETL is an essential component used to load the data into data warehouses (DWH), operational data stores (ODS) and datamarts (DM) from the source systems. ETL processes are also widely used in data integration, data migration and master data management (MDM) initiatives.

Architectural Context

Supported Use Cases
- Bulk data integration
- Flat-file based and hierarchical transformations
- High scale, batch-oriented data delivery

Examples
- Financial ODS

Goals and Benefits
- The objective of an ETL process is to facilitate the data movement and transformation
- ETL is the technology that performs three distinct functions of data movement:
  - the Extraction of data from one or more sources
  - the Transformations of the data e.g. cleansing, reformatting, standardization, aggregation, or the application of any number of business rules and
  - the Loading of the resulting data set into specified target systems or file formats
- ETL processes are reusable components that can be scheduled to perform data movement jobs on a regular basis
- ETL supports massive parallel processing (MPP) for large data volumes

When to use
- Data movement across or within systems involving high data volumes and complex business rules
- Load data into data warehouses (DWH), operational data stores (ODS), datamarts (DM)

Strategy
- ETL processes are in usually grouped and executed as batch jobs
- ETL tools (like Informatica PowerCenter) are used to implement the ETL processes
- ETL processes are designed to be very efficient, scalable, and maintainable

Limitations
- Real-time event based transfers
- Transactional integrations

ETL Architecture Pattern

In computing, Extract, Transform, Load (ETL) refers to a process in database usage and especially in data warehousing. Data extraction is where data is extracted from homogeneous or heterogeneous data sources; data transformation where the data is transformed for storing in the proper format or structure for the purposes of querying and analysis; data loading where the data is loaded into the final target database, more specifically, an operational data store, data mart, or data warehouse.

Since the data extraction takes time, it is common to execute the three phases in parallel. While the data is being extracted, another transformation process executes. It processes the already received data and prepares it for loading. As soon as there is some data ready to be loaded into the target, the data loading kicks off without waiting for the completion of the previous phases.

ETL systems commonly integrate data from multiple applications (systems), typically developed and supported by different vendors or hosted on separate computer hardware. The disparate systems containing the original data are frequently managed and operated by different employees. For example, a cost accounting system may combine data from payroll, sales, and purchasing.

Extract

The first part of an ETL process involves extracting the data from the source system(s). In many cases this represents the most important aspect of ETL, since extracting data correctly sets the stage for the success of subsequent processes. Most data-warehousing projects combine data from different source systems. Each separate system may also use a different data organization and/or format. Common data-source formats include relational databases, XML and flat files, but may also include non-relational database structures such as Information Management System (IMS) or other data structures such as Virtual Storage Access Method (VSAM) or Indexed Sequential Access Method (ISAM), or even formats fetched from outside sources by means such as web spidering or screen-scraping. The streaming of the extracted data source and loading on-the-fly to the destination database is another way of performing ETL when no intermediate data storage is required. In general, the extraction phase aims to convert the data into a single format appropriate for transformation processing.

An intrinsic part of the extraction involves data validation to confirm whether the data pulled from the sources has the correct/expected values in a given domain (such as a pattern/default or list of values). If the data fails the validation rules it is rejected entirely or in part. The rejected data is ideally reported back to the source system for further analysis to identify and to rectify the incorrect records. In some cases the extraction process itself may have to do a data-validation rule in order to accept the data and flow to the next phase.

Transform

In the data transformation stage, a series of rules or functions are applied to the extracted data in order to prepare it for loading into the end target. Some data does not require any transformation at all; such data is known as “direct move” or “pass through” data.

An important function of transformation is the cleaning of data, which aims to pass only “proper” data to the target. The challenge when different systems interact is in the relevant systems’ interfacing and communicating. Character sets that may be available in one system may not be so in others.

In other cases, one or more of the following transformation types may be required to meet the business and technical needs of the server or data warehouse:

- Selecting only certain columns to load: (or selecting null columns not to load). For example, if the source data has three columns (aka "attributes"), roll_no, age, and salary, then the selection may take only roll_no and salary. Or, the selection mechanism may ignore all those records where salary is not present (salary = null).
- Translating coded values: (*e.g.*, if the source system codes male as "1" and female as "2", but the warehouse codes male as "M" and female as "F")
- Encoding free-form values: (*e.g.*, mapping "Male" to "M")
- Deriving a new calculated value: (*e.g.*, sale_amount = qty * unit_price)
- Sorting or ordering the data based on a list of columns to improve search performance
- Joining data from multiple sources (*e.g.*, lookup, merge) and deduplicating the data
- Aggregating (for example, rollup — summarizing multiple rows of data — total sales for each store, and for each region, etc.)
- Generating surrogate-key values
- Transposing or pivoting (turning multiple columns into multiple rows or vice versa)
- Splitting a column into multiple columns (*e.g.*, converting a comma-separated list, specified as a string in one column, into individual values in different columns)
- Disaggregating repeating columns
- Looking up and validating the relevant data from tables or referential files
- Applying any form of data validation; failed validation may result in a full rejection of the data, partial rejection, or no rejection at all, and thus none, some, or all of the data is handed over to the next step depending on the rule design and exception handling; many of the above transformations may result in exceptions, e.g., when a code translation parses an unknown code in the extracted data

Load

The load phase loads the data into the end target that may be a simple delimited flat file or a data warehouse. Depending on the requirements of the organization, this process varies widely. Some data warehouses may overwrite existing information with cumulative information; updating extracted data is frequently done on a daily, weekly, or monthly basis. Other data warehouses (or even other parts of the same data warehouse) may add new data in a historical form at regular intervals—for example, hourly. To understand this, consider a data warehouse that is required to maintain sales records of the last year. This data warehouse overwrites any data older than a year with newer data. However, the entry of data for any one year window is made in a historical manner. The timing and scope to replace or append are strategic design choices dependent on the time available and the business needs. More complex systems can maintain a history and audit trail of all changes to the data loaded in the data warehouse.

As the load phase interacts with a database, the constraints defined in the database schema — as well as in triggers activated upon data load — apply (for example, uniqueness, referential integrity, mandatory fields), which also contribute to the overall data quality performance of the ETL process.

- For example, a financial institution might have information on a customer in several departments and each department might have that customer's information listed in a different way. The membership department might list the customer by name, whereas the accounting department might list the customer by number. ETL can bundle all of these data elements and consolidate them into a uniform presentation, such as for storing in a database or data warehouse.
- Another way that companies use ETL is to move information to another application permanently. For instance, the new application might use another database vendor and most likely a very different database schema. ETL can be used to transform the data into a format suitable for the new application to use.
- An example would be an Expense and Cost Recovery System (ECRS) such as used by accountancies, consultancies, and legal firms. The data usually ends up in the time and billing system, although some businesses may also utilize the raw data for employee productivity reports to Human Resources (personnel dept.) or equipment usage reports to Facilities Management.

Real-life ETL Cycle

The typical real-life ETL cycle consists of the following execution steps:

1. Cycle initiation
2. Build reference data
3. Extract (from sources)
4. Validate
5. Transform (clean, apply business rules, check for data integrity, create aggregates or disaggregates)
6. Stage (load into staging tables, if used)
7. Audit reports (for example, on compliance with business rules. Also, in case of failure, helps to diagnose/repair)
8. Publish (to target tables)
9. Archive

Challenges

ETL processes can involve considerable complexity, and significant operational problems can occur with improperly designed ETL systems.

The range of data values or data quality in an operational system may exceed the expectations of designers at the time validation and transformation rules are specified. Data profiling of a source during data analysis can identify the data conditions that must be managed by transform rules specifications, leading to an amendment of validation rules explicitly and implicitly implemented in the ETL process.

Data warehouses are typically assembled from a variety of data sources with different formats and purposes. As such, ETL is a key process to bring all the data together in a standard, homogeneous environment.

Design analysts should establish the scalability of an ETL system across the lifetime of its usage---including understanding the volumes of data that must be processed within service level agreements. The time available to extract from source systems may change, which may mean the same amount of data may have to be processed in less time. Some ETL systems have to scale to process terabytes of data to update data warehouses with tens of terabytes of data. Increasing volumes of data may require designs that can scale from daily batch to multiple-day micro batch to integration with message queues or real-time change-data capture for continuous transformation and update.

Performance

ETL vendors benchmark their record-systems at multiple TB (terabytes) per hour (or ~1 GB per second) using powerful servers with multiple CPUs, multiple hard drives, multiple gigabit-network connections, and lots of memory. The fastest ETL record is currently held by Syncsort, Vertica, and HP at 5.4TB in under an hour, which is more than twice as fast as the earlier record held by Microsoft and Unisys.

In real life, the slowest part of an ETL process usually occurs in the database load phase. Databases may perform slowly because they have to take care of concurrency, integrity maintenance, and indices. Thus, for better performance, it may make sense to employ:

- Direct Path Extract method or bulk unload whenever is possible (instead of querying the database) to reduce the load on source system while getting high speed extract
- Most of the transformation processing outside of the database
- Bulk load operations whenever possible

Still, even using bulk operations, database access is usually the bottleneck in the ETL process. Some common methods used to increase performance are:

- Partition tables (and indices): try to keep partitions similar in size (watch for null values that can skew the partitioning)
- Do all validation in the ETL layer before the load: disable integrity checking (disable constraint ...) in the target database tables during the load
- Disable triggers (disable trigger ...) in the target database tables during the load: simulate their effect as a separate step

- Generate IDs in the ETL layer (not in the database)
- Drop the indices (on a table or partition) before the load - and recreate them after the load (SQL: drop index ...; create index ...)
- Use parallel bulk load when possible — works well when the table is partitioned or there are no indices (Note: attempt to do parallel loads into the same table (partition) usually causes locks — if not on the data rows, then on indices)
- If a requirement exists to do insertions, updates, or deletions, find out which rows should be processed in which way in the ETL layer, and then process these three operations in the database separately; you often can do bulk load for inserts, but updates and deletes commonly go through an API (using SQL)

Whether to do certain operations in the database or outside may involve a trade-off. For example, removing duplicates using distinct may be slow in the database; thus, it makes sense to do it outside. On the other side, if using distinct significantly (x100) decreases the number of rows to be extracted, then it makes sense to remove duplications as early as possible in the database before unloading data.

A common source of problems in ETL is a big number of dependencies among ETL jobs. For example, job “B” cannot start while job “A” is not finished. One can usually achieve better performance by visualizing all processes on a graph, and trying to reduce the graph making maximum use of parallelism, and making “chains” of consecutive processing as short as possible. Again, partitioning of big tables and their indices can really help.

Another common issue occurs when the data are spread among several databases, and processing is done in those databases sequentially. Sometimes database replication may be involved as a method of copying data between databases - it can significantly slow down the whole process. The common solution is to reduce the processing graph to only three layers:

- Sources
- Central ETL layer
- Targets

This approach allows processing to take maximum advantage of parallelism. For example, if you need to load data into two databases, you can run the loads in parallel (instead of loading into the first - and then replicating into the second).

Sometimes processing must take place sequentially. For example, dimensional (reference) data are needed before one can get and validate the rows for main “fact” tables.

Parallel Processing

A recent development in ETL software is the implementation of parallel processing. It has enabled a number of methods to improve overall performance of ETL when dealing with large volumes of data.

ETL applications implement three main types of parallelism:

- Data: By splitting a single sequential file into smaller data files to provide parallel access
- Pipeline: allowing the simultaneous running of several components on the same data stream, e.g. looking up a value on record 1 at the same time as adding two fields on record 2
- Component: The simultaneous running of multiple processes on different data streams in the same job, e.g. sorting one input file while removing duplicates on another file

All three types of parallelism usually operate combined in a single job.

An additional difficulty comes with making sure that the data being uploaded is relatively consistent. Because multiple source databases may have different update cycles (some may be updated every few minutes, while others may take days or weeks), an ETL system may be required to hold back certain data until all sources are synchronized. Likewise, where a warehouse may have to be reconciled to the contents in a source system or with the general ledger, establishing synchronization and reconciliation points becomes necessary.

Rerunnability, Recoverability

Data warehousing procedures usually subdivide a big ETL process into smaller pieces running sequentially or in parallel. To keep track of data flows, it makes sense to tag each data row with "row_id", and tag each piece of the process with "run_id". In case of a failure, having these IDs help to roll back and rerun the failed piece.

Best practice also calls for *checkpoints*, which are states when certain phases of the process are completed. Once at a checkpoint, it is a good idea to write everything to disk, clean out some temporary files, log the state, and so on.

Virtual ETL

As of 2010 data virtualization had begun to advance ETL processing. The application of data virtualization to ETL allowed solving the most common ETL tasks of data migration and application integration for multiple dispersed data sources. Virtual ETL operates with the abstracted representation of the objects or entities gathered from the variety of relational, semi-structured, and unstructured data sources. ETL tools can leverage object-oriented modeling and work with entities' representations persistently stored in a centrally located hub-and-spoke architecture. Such a collection that contains representations of the entities or objects gathered from the data sources for ETL processing is called a metadata repository and it can reside in memory or be made persistent. By using a persistent metadata repository, ETL tools can transition from one-time projects to persistent middleware, performing data harmonization and data profiling consistently and in near-real time.

Dealing with Keys

Keys play an important part in all relational databases, as they tie everything together. A primary key is a column that identifies a given entity, whereas a foreign key is a column in another table that refers to a primary key. Keys can comprise several columns, in which case they are composite keys. In many cases the primary key is an auto-generated integer that has no meaning for the business entity being represented, but solely exists for the purpose of the relational database - commonly referred to as a surrogate key.

As there is usually more than one data source getting loaded into the warehouse, the keys are an important concern to be addressed. For example: customers might be represented in several data sources, with their Social Security Number as the primary key in one source, their phone number in another, and a surrogate in the third. Yet a data warehouse may require the consolidation of all the customer information into one dimension table.

A recommended way to deal with the concern involves adding a warehouse surrogate key, which is used as a foreign key from the fact table.

Usually updates occur to a dimension's source data, which obviously must be reflected in the data warehouse.

If the primary key of the source data is required for reporting, the dimension already contains that piece of information for each row. If the source data uses a surrogate key, the warehouse must keep track of it even though it is never used in queries or reports; it is done by creating a lookup table that contains the warehouse surrogate key and the originating key. This way, the dimension is not polluted with surrogates from various source systems, while the ability to update is preserved.

The lookup table is used in different ways depending on the nature of the source data. There are 5 types to consider; three are included here:

Type 1

> The dimension row is simply updated to match the current state of the source system; the warehouse does not capture history; the lookup table is used to identify the dimension row to update or overwrite

Type 2

> A new dimension row is added with the new state of the source system; a new surrogate key is assigned; source key is no longer unique in the lookup table

Fully logged

> A new dimension row is added with the new state of the source system, while the previous dimension row is updated to reflect it is no longer active and time of deactivation.

Tools

By using an established ETL framework, one may increase one's chances of ending up with better connectivity and scalability. A good ETL tool must be able to communicate with the many different relational databases and read the various file formats used throughout an organization. ETL tools have started to migrate into Enterprise Application Integration, or even Enterprise Service Bus, systems that now cover much more than just the extraction, transformation, and loading of data. Many ETL vendors now have data profiling, data quality, and metadata capabilities. A common use case for ETL tools include converting CSV files to formats readable by relational databases. A typical translation of millions of records is facilitated by ETL tools that enable users to input csv-like data feeds/files and import it into a database with as little code as possible.

ETL tools are typically used by a broad range of professionals - from students in computer science looking to quickly import large data sets to database architects in charge of company account management, ETL tools have become a convenient tool that can be relied on to get maximum performance. ETL tools in most cases contain a GUI that helps users conveniently transform data, using a visual data mapper, as opposed to writing large programs to parse files and modify data types.

While ETL Tools have traditionally been for developers and I.T. staff, the new trend is to provide these capabilities to business users so they can themselves create connections and data integrations when needed, rather than going to the I.T. staff. Gartner refers to these non-technical users as Citizen Integrators.

Star Schema

In computing, the star schema is the simplest style of data mart schema and is the approach most widely used to develop data warehouses and dimensional data marts. The star schema consists of one or more fact tables referencing any number of dimension tables. The star schema is an important special case of the snowflake schema, and is more effective for handling simpler queries.

The star schema gets its name from the physical model's resemblance to a star shape with a fact table at its center and the dimension tables surrounding it representing the star's points.

Model

The star schema separates business process data into facts, which hold the measurable, quantitative data about a business, and dimensions which are descriptive attributes related to fact data. Examples of fact data include sales price, sale quantity, and time, distance, speed, and weight measurements. Related dimension attribute examples include product models, product colors, product sizes, geographic locations, and salesperson names.

A star schema that has many dimensions is sometimes called a *centipede schema*. Having dimensions of only a few attributes, while simpler to maintain, results in queries with many table joins and makes the star schema less easy to use.

Fact Tables

Fact tables record measurements or metrics for a specific event. Fact tables generally consist of numeric values, and foreign keys to dimensional data where descriptive information is kept. Fact tables are designed to a low level of uniform detail (referred to as "granularity" or "grain"), meaning facts can record events at a very atomic level. This can result in the accumulation of a large number of records in a fact table over time. Fact tables are defined as one of three types:

- Transaction fact tables record facts about a specific event (e.g., sales events)
- Snapshot fact tables record facts at a given point in time (e.g., account details at month end)

- Accumulating snapshot tables record aggregate facts at a given point in time (e.g., total month-to-date sales for a product)

Fact tables are generally assigned a surrogate key to ensure each row can be uniquely identified. This key is a simple primary key.

Dimension Tables

Dimension tables usually have a relatively small number of records compared to fact tables, but each record may have a very large number of attributes to describe the fact data. Dimensions can define a wide variety of characteristics, but some of the most common attributes defined by dimension tables include:

- Time dimension tables describe time at the lowest level of time granularity for which events are recorded in the star schema
- Geography dimension tables describe location data, such as country, state, or city
- Product dimension tables describe products
- Employee dimension tables describe employees, such as sales people
- Range dimension tables describe ranges of time, dollar values, or other measurable quantities to simplify reporting

Dimension tables are generally assigned a surrogate primary key, usually a single-column integer data type, mapped to the combination of dimension attributes that form the natural key.

Benefits

Star schemas are denormalized, meaning the normal rules of normalization applied to transactional relational databases are relaxed during star schema design and implementation. The benefits of star schema denormalization are:

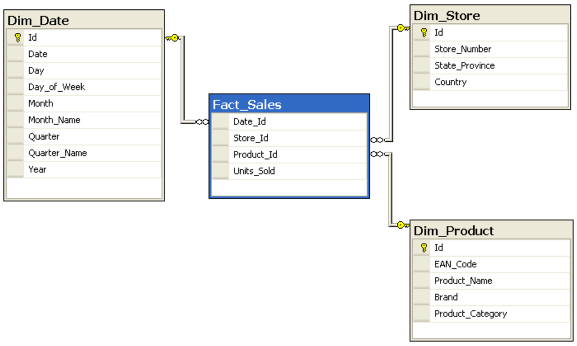
- Simpler queries - star schema join logic is generally simpler than the join logic required to retrieve data from a highly normalized transactional schema.
- Simplified business reporting logic - when compared to highly normalized schemas, the star schema simplifies common business reporting logic, such as period-over-period and as-of reporting.
- Query performance gains - star schemas can provide performance enhancements for read-only reporting applications when compared to highly normalized schemas.
- Fast aggregations - the simpler queries against a star schema can result in improved performance for aggregation operations.
- Feeding cubes - star schemas are used by all OLAP systems to build proprietary OLAP cubes efficiently; in fact, most major OLAP systems provide a ROLAP mode of operation which can use a star schema directly as a source without building a proprietary cube structure.

Disadvantages

The main disadvantage of the star schema is that data integrity is not enforced as well as it is in a highly normalized database. One-off inserts and updates can result in data anomalies which normalized schemas are designed to avoid. Generally speaking, star schemas are loaded in a highly controlled fashion via batch processing or near-real time "trickle feeds", to compensate for the lack of protection afforded by normalization.

Star schema is also not as flexible in terms of analytical needs as a normalized data model. Normalized models allow any kind of analytical queries to be executed as long as they follow the business logic defined in the model. Star schemas tend to be more purpose-built for a particular view of the data, thus not really allowing more complex analytics. Star schemas don't support many-to-many relationships between business entities - at least not very naturally. Typically these relationships are simplified in star schema to conform to the simple dimensional model.

Example


Dim_Date
Id
Date
Day
Day_of_Week
Month
Month_Name
Quarter
Quarter_Name
Year
Fact_Sales
Date_Id
Store_Id
Product_Id
Units_Sold
Dim_Store
Id
Store_Number
State_Province
Country
Dim_Product
Id
EAN_Code
Product_Name
Brand
Product_Category

Star schema used by example query.

Consider a database of sales, perhaps from a store chain, classified by date, store and product. The image of the schema to the right is a star schema version of the sample schema provided in the snowflake schema article.

Fact_Sales is the fact table and there are three dimension tables Dim_Date, Dim_Store and Dim_Product.

Each dimension table has a primary key on its Id column, relating to one of the columns (viewed as rows in the example schema) of the Fact_Sales table's three-column (compound) primary key (Date_Id, Store_Id, Product_Id). The non-primary key Units_Sold column of the fact table in this example represents a measure or metric that can be used in calculations and analysis. The non-primary key columns of the dimension tables represent additional attributes of the dimensions (such as the Year of the Dim_Date dimension).

For example, the following query answers how many TV sets have been sold, for each brand and country, in 1997:

```
SELECT

        P.Brand,
        S.Country AS Countries,
        SUM(F.Units_Sold)
FROM Fact_Sales F
INNER JOIN Dim_Date D    ON (F.Date_Id = D.Id)
INNER JOIN Dim_Store S   ON (F.Store_Id = S.Id)
INNER JOIN Dim_Product P ON (F.Product_Id = P.Id)
WHERE D.Year = 1997 AND  P.Product_Category = ‘tv’
GROUP BY
        P.Brand,
        S.Country
```

References

- Ralph Kimball, Margy Ross, The Data Warehouse Toolkit: The Complete Guide to Dimensional Modeling, Second Edition, Wiley Computer Publishing, 2002. ISBN 0471-20024-7.
- Ralph Kimball, The Data Warehouse Toolkit, Second Edition, Wiley Publishing, Inc., 2008. ISBN 978-0-470-14977-5, Pages 253-256
- Thomas C. Hammergren; Alan R. Simon (February 2009). Data Warehousing for Dummies, 2nd edition. John Wiley & Sons. ISBN 978-0-470-40747-9.
- “Information Theory & Business Intelligence Strategy - Small Worlds Data Transformation Measure - MIKE2.0, the open source methodology for Information Development”. Mike2.openmethodology.org. Retrieved 2013-06-14.
- Linstedt, Dan. “Data Vault Series 1 – Data Vault Overview”. Data Vault Series. The Data Administration Newsletter. Retrieved 12 September 2011.

5

Market Research: An Integrated Study

The effort put in to gather information related to customers or markets is known as market research. Market research is an important part of business strategy. Market segmentation, market trend, SWOT analysis and market research are some of the topics elucidated in this chapter.

Market Research

Market research is any organized effort to gather information about target markets or customers. It is a very important component of business strategy. The term is commonly interchanged with marketing research; however, expert practitioners may wish to draw a distinction, in that *marketing* research is concerned specifically about marketing processes, while *market* research is concerned specifically with markets.

Market research is one of the key factors used in maintaining competitiveness over competitors. Market research provides important information to identify and analyze the market need, market size and competition. Market-research techniques encompass both qualitative techniques such as focus groups, in-depth interviews, and ethnography, as well as quantitative techniques such as customer surveys, and analysis of secondary data.

Market research, which includes social and opinion research, is the systematic gathering and interpretation of information about individuals or organizations using statistical and analytical methods and techniques of the applied social sciences to gain insight or support decision making.

History

Market research began to be conceptualized and put into formal practice during the 1920s, as an offshoot of the advertising boom of the Golden Age of radio in the United States. Advertisers began to realize the significance of demographics revealed by sponsorship of different radio programs.

Market Research for Business/Planning

Market research is a way of getting an overview of consumers' wants, needs and beliefs. It can also involve discovering how they act. The research can be used to determine how a product could be marketed. Peter Drucker believed market research to be the quintessence of marketing.

There are two major types of market research. Primary Research sub-divided into Quantitative and Qualitative research and Secondary research.

Factors that can be investigated through market research include

Market information

Through Market information one can know the prices of different commodities in the market, as well as the supply and demand situation. Market researchers have a wider role than previously recognized by helping their clients to understand social, technical, and even legal aspects of markets.

Market segmentation

Market segmentation is the division of the market or population into subgroups with similar motivations. It is widely used for segmenting on geographic differences, personality differences, demographic differences, technographic differences, use of product differences, psychographic differences and gender differences. For B2B segmentation firmographics is commonly used.

Market trends

Market trends are the upward or downward movement of a market, during a period of time. Determining the market size may be more difficult if one is starting with a new innovation. In this case, you will have to derive the figures from the number of potential customers, or customer segments. [Ilar 1998]

SWOT analysis

SWOT is a written analysis of the Strengths, Weaknesses, Opportunities and Threats to a business entity. Not only should a SWOT be used in the creation stage of the company but could also be used throughout the life of the company. A SWOT may also be written up for the competition to understand how to develop the marketing and product mixes.

Another factor that can be measured is marketing effectiveness. This includes

- Customer analysis
- Choice modelling
- Competitor analysis
- Risk analysis
- Product research
- Advertising the research
- Marketing mix modeling
- Simulated Test Marketing

Market Research - Benefits

Market Research is an essential tool which assists in making strategic decisions. It reduces the risks involved in making decisions as well as strategies. Companies either do this research in house or outsource this process to business experts or organizations who have dedicated and trained resources to perform this.In the recent years an increasing trend of such market research companies assisting business strategists have come up. Some major benefits are -

i. Marketing research assists in providing accurate and latest trends related to demand, consumer behavior, sales, growth opportunities etc.

ii. It helps in better understanding of the market, thus helps in product design, features and demand forecasts

iii. It assists in studying and understanding the competitors, thus identifying unique selling propositions for a business.

Market research reaps many such benefits which are Industry and business specific and have high ROI's.

Market Research for The Film Industry

It is important to test marketing material for films to see how an audience will receive it. There are several market research practices that may be used: (1) concept testing, which evaluates reactions to a film idea and is fairly rare; (2) positioning studios, which analyze a script for marketing opportunities; (3) focus groups, which probe viewers' opinions about a film in small groups prior to release; (4) test screenings, which involve the previewing of films prior to theatrical release; (5) tracking studies, which gauge (often by telephone polling) an audience's awareness of a film on a weekly basis prior to and during theatrical release; (6) advertising testing, which measures responses to marketing materials such as trailers and television advertisements; and finally (7) exit surveys, that measure audience reactions after seeing the film in the cinema.

Influence from The Internet

The availability of research by way of the Internet has influenced a vast number of consumers using this media; for gaining knowledge relating to virtually every type of available product and service. It has been added to by the growth factor of emerging global markets, such as China, Indonesia and Russia, which is significantly exceeding that of the established and more advanced B2B e-commerce markets. Various statistics show that the increasing demands of consumers are reflected not only in the wide and varied range of general Internet researching applications, but in online shopping research penetration.

This is stimulated by product-enhancing websites, graphics, and content designed to attract casual "surfing" shoppers, researching for their particular needs, competitive prices and quality. According to the Small Business Administration (SBA), a successful business is significantly contributed to by gaining knowledge about customers, competitors, and the associated industry. Market research creates not only this understanding, but is the process of data analysis regarding which products and services are in demand.

The convenience and easy accessibility of the Internet has created a global B2C E-commerce research facility, for a vast online shopping network that has motivated retail markets in developed countries. In 2010, between $400 billion and $600 billion in revenue was generated by this medium also, it is anticipated that in 2015, this online market will generate revenue between $700 billion and $950 billion. The influence of market research, irrespective of what form it takes, is an extremely powerful incentive for any type of consumer and their providers!

Beyond online web-based market research activities, the Internet has also influenced high-street modes of data collection by, for example, replacing the traditional paper clipboard with online survey providers. Over the last 5 years, mobile surveys have become increasingly popular. Mobile has opened the door to innovative new methods of engaging respondents, such as social voting communities.

Research and Social Media Applications

The UK Market Research Society (MRS) reports research has shown that on average, the three social media platforms primarily used by Millennials are LinkedIn, Facebook and YouTube. Social Media applications, according to T-Systems, help generate the B2B E-commerce market and develop electronic business process efficiency. This application is a highly effective vehicle for market research, which combined with E-commerce, is now regarded as a separate, extremely profitable field of global business. While many B2B business models are being updated, the various advantages and benefits offered by Social Media platforms are being integrated within them.

Business intelligence organization have compiled a comprehensive report related to global online retail sales, defining continued growth patterns and trends in the industry. Headed “Global B2C E-Commerce and Online Payment Market 2014” the report perceives a decrease in overall growth rates in North America and Western Europe, as the expected growth in the online market sales, is absorbed into the emerging markets. It is forecast that the Asia-Pacific region will see the fastest growth in the B2C E-Commerce market and replace North America as the B2C E-Commerce sales region leader, within a few years. This effectively, offers a significant, motivational platform for new Internet services, to promote user market research-friendly applications.

Research and Market Sectors

The primary online sale providers in B2C E-Commerce, worldwide, includes the USA based Amazon.com Inc. which remains the E-Commerce revenues, global leader. The growth leaders in the world top ten are two online companies from China, both of which conducted Initial Public Offering (IPO) this year; Alibaba Group Holding Ltd. and JD Inc. Another company from the top ten is Cnova N.V., a recently formed E-Commerce subsidiary of the French Group Casino, with various store retailers developing and expanding their E-Commerce facilities worldwide. It is a further indication of how consumers are increasingly being attracted to the opportunities of online researching and expanding their awareness of what is available to them.

Service providers; for example those related to finance, foreign market trade and investment promote a variety of information and research opportunities to online users. In addition, they provide comprehensive and competitive strategies with market research tools, designed to promote worldwide business opportunities for entrepreneurs and established providers. General access, to accurate and supported market research facilities, is a critical aspect of business development and success today. The Marketing Research Association was founded in 1957 and is recognized as one of the leading and prominent associations in the opinion and marketing research profession. It serves the purpose of providing insights and intelligence that helps businesses make decisions regarding the provision of products and services to consumers and industries.

This organization knowledge of market conditions and competition is gained by researching relevant sectors, which provide advantages for entry into new and established industries. It enables effective strategies to be implemented; the assessment of global environments in the service sectors, as well as foreign market trade and investment barriers! Research, is utilized for promoting export opportunities and inward investment, helping determine how to execute competitive strategies, focus on objective policies and strengthen global opportunities. It is a medium that influences, administrates and enforces agreements, preferences, leveling trading environments and competitiveness in the international marketplace.

The retail industry aspect of online market research, is being transformed worldwide by M-Commerce with its mobile audience, rapidly increasing as the volume and varieties of products purchased on the mobile medium, increases. Researches conducted in the markets of North America and Europe, revealed that the M-Commerce penetration on the total online retail trade, had attained 10%, or more. It was also shown that in emerging markets, smart-phone and tablet penetration is fast increasing and contributing significantly to online shopping growth.

Financial Performance

Top 10 U.S. Market Research Organizations 2013

From The 2014 AMA Gold Top 50 Report:

Rank	Company	W.W Research Revenue in 2013 (million USD)
1	Nielsen Holdings N.V. - Nielsen Holdings N.V., Arbitron Inc.	6,054.0
2	Kantar Group - TNS, Millward Brown, etc.	3,363.7
3	IMS Health Inc.	2,544.0
4	Ipsos Group S.A.	2,274.0
5	GfK	1,985.5
6	IRI	845.1
7	Westat Inc.	582.5
8	dunnhumbyUSA LLC	462.0
9	The NPD Group Inc.	287.7
10	comScore Inc.	286.9

Top 10 U.K Market Research Organizations 2013

From the Market Research Society UK Research & Insight Industry League Tables

Rank	Company	Turnover in 2013 (million GBP)
1	TNS UK	194.140
2	Dunn Humby	165.220
3	Ipsos MORI	162.400
4	Gartner	121.036
5	GfK NOP	116.366

6	Millward Brown	105.043
7	IMS Health Group	93.231
8	ACNielsen	95.119
9	Wood Mackenzie Research & Consulting	85.120
10	Euromonitor	74.228

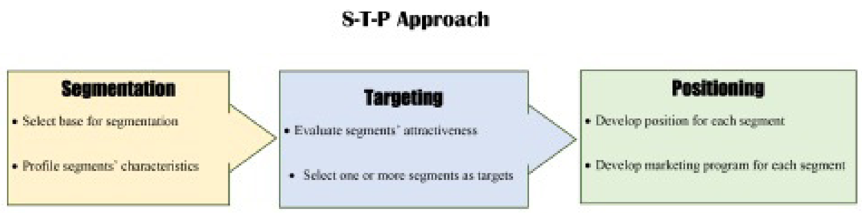
Global Market Research Turnover in 2014

From the ESOMAR Global Market Research Report 2014

Rank	Continent	Sales in 2013 (million USD)	Share
1	Europe	16,005	40%
2	North America	15,705	39%
3	Asia Pacific	5,998	15%
4	Latin America	1,920	5%
5	Africa	382.1	1%
6	Middle East	277	1%

Market Segmentation

Market segmentation is the process of dividing a broad consumer or business market, normally consisting of existing and potential customers, into sub-groups of consumers (known as *segments*) based on some type of shared characteristics. In dividing or segmenting markets, researchers typically look for shared characteristics such as common needs, common interests, similar lifestyles or even similar demographic profiles. The overall aim of segmentation is to identify *high yield segments* – that is, those segments that are likely to be the most profitable or that have growth potential – so that these can be selected for special attention (i.e. become target markets). Many different ways to segment a market have been identified. Business-to-business (B2B) sellers might segment the market into different types of businesses or countries. While business to consumer (B2C) seller might segment the market into demographic segments, lifestyle segments, behavioral segments or any other meaningful segment.


S-T-P Approach
Segmentation
• Select base for segmentation
• Profile segments' characteristics
Targeting
• Evaluate segments' attractiveness
• Select one or more segments as targets
Positioning
• Develop position for each segment
• Develop marketing program for each segment

The STP approach highlights the three areas of decision-making

Market segmentation assumes that different market segments require different marketing programs – that is, different offers, prices, promotion, distribution or some combination of marketing variables. Market segmentation is not only designed to identify the most profitable segments, but also to develop profiles of key segments in order to better understand their needs and purchase motivations. Insights from segmentation analysis are subsequently used to support marketing strategy development and planning. Many marketers use the S-T-P approach; Segmentation → Targeting → Positioning to provide the framework for marketing planning objectives. That is, a market is segmented, one or more segments are selected for targeting, and products or services are positioned in a way that resonates with the selected target market or markets.

Market Segmentation: Historical Overview

The business historian, Richard S. Tedlow, identifies four stages in the evolution of market segmentation:

The Model T Ford *Henry Ford famously said "Any customer can have a car painted any color that he wants so long as it is black"*

- Fragmentation (pre 1880s): The economy was characterised by small regional suppliers who sold goods on a local or regional basis
- Unification or Mass Marketing (1880s-1920s): As transportation systems improved, the economy became unified. Standardised, branded goods were distributed at a national level. Manufacturers tended to insist on strict standardisation in order to achieve scale economies with a view to penetrating markets in the early stages of a product's life cycle. e.g. the Model T Ford
- Segmentation (1920s-1980s): As market size increased, manufacturers were able to produce different models pitched at different quality points to meet the needs of various demographic and psychographic market segments. This is the era of market differentiation based on demographic, socio-economic and lifestyle factors.
- Hyper-segmentation (1980s+): a shift towards the definition of ever more narrow market segments. Technological advancements, especially in the area of digital communications, allows marketers to communicate to individual consumers or very small groups.

Contemporary market segmentation emerged in the twentieth century as marketers responded to two pressing issues. Demographic and purchasing data were available for groups but rarely for individuals and secondly, advertising and distribution channels were available for groups, but rarely for single consumers and so brand marketers approached the task from a tactical viewpoint. Thus, segmentation was essentially a brand-driven process. Until recently, most segmentation approaches have retained this tactical perspective in that they address immediate short-term questions; typically decisions about the current "market served" and are concerned with informing marketing mix decisions.

Criticisms of Market Segmentation

The limitations of conventional segmentation have been well documented in the literature. Perennial criticisms include:

- that it is no better than mass marketing at building brands
- that in competitive markets, segments rarely exhibit major differences in the way they use brands
- that it fails to identify sufficiently narrow clusters
- geographic/demographic segmentation is overly descriptive and lacks sufficient insights into the motivations necessary to drive communications strategy
- difficulties with market dynamics, notably the instability of segments over time and structural change which leads to segment creep and membership migration as individuals move from one segment to another

Market segmentation has many critics. But in spite of its limitations, market segmentation remains one of the enduring concepts in marketing and continues to be widely used in practice. One American study, for example, suggested that almost 60 percent of senior executives had used market segmentation in the past two years.

Market Segmentation Strategy

A key consideration for marketers is whether to segment or not to segment. Depending on company philosophy, resources, product type or market characteristics, a businesses may develop an undifferentiated approach or *differentiated approach*. In an undifferentiated approach (also known as *mass marketing*), the marketer ignores segmentation and develops a product that meets the needs of the largest number of buyers. In a differentiated approach the firm targets one or more market segments, and develops separate offers for each segment.

Even simple products like salt are highly differentiated in practice.

In consumer marketing, it is difficult to find examples of undifferentiated approaches. Even goods such as salt and sugar, which were once treated as commodities, are now highly differentiated. Consumers can purchase a variety of salt products; cooking salt, table salt, sea-salt, rock salt, kosher salt, mineral salt, herbal or vegetable salts, iodized salt, salt substitutes and if that is not enough choice, at the brand level, gourmet cooks are likely to make a major distinction between Maldon salt and other competing brands.The following table outlines the main strategic approaches.

Table 1: Main Strategic Approaches to Segmentation

Number of Segments	Segmentation Strategy	Comments
Zero	Undifferentiated strategy	Mass marketing
One	Focus strategy	Niche marketing
Two or more	Differentiated strategy	Multiple niches
Thousands	Hypersegmentation	One-to-one marketing

A number of factors are likely to affect a company's segmentation strategy:

- Company resources: When resources are restricted, a concentrated strategy may be more effective.
- Product variability: For highly uniform products (such as sugar or steel) an undifferentiated marketing may be more appropriate. For products that can be differentiated, (such as cars) then either a differentiated or concentrated approach is indicated.
- Product life cycle: For new products, one version may be used at the launch stage, but this may be expanded to a more segmented approach over time. As more competitors enter the market, it may be necessary to differentiate.
- Market characteristics: When all buyers have similar tastes, or are unwilling to pay a premium for different quality, then undifferentiated marketing is indicated.
- Competitive activity: When competitors apply differentiated or concentrated market segmentation, using undifferentiated marketing may prove to be fatal. A company should consider whether it can use a different market segmentation approach.

Market Segmentation Process: S-T-P

The process of segmenting the market is deceptively simple. Seven basic steps describe the entire process including segmentation, targeting and positioning. In practice, however, the task can be very laborious since it involves poring over loads of data, and requires a great deal of skill in analysis, interpretation and some judgement. Although a great deal of analysis needs to be undertaken, and many decisions need to be made, marketers tend to use the so-called S-T-P process, that is Segmentation→ Targeting → Positioning, as a broad framework for simplifying the process and outlined here:

Segmentation

1. Identify market (also known as the universe) to be segmented.

2. Identify, select and apply base or bases to be used in the segmentation

3. Develop segment profiles

Targeting

4. Evaluate each segment's attractiveness

5. Select segment or segments to be targeted

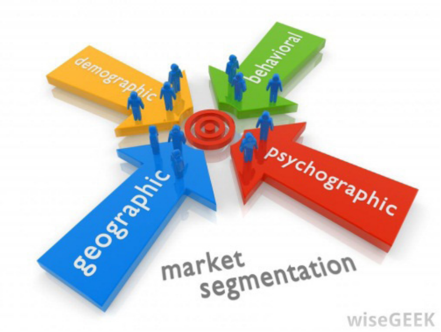
Positioning

6. Identify optimal positioning for each segment

7. Develop the marketing program for each segment

Bases for Segmenting Consumer Markets

A major step in the segmentation process is the selection of a suitable base. In this step, marketers are looking for a means of achieving internal homogeneity (similarity within the segments), and external heterogeneity (differences between segments). In other words, they are searching for a process that minimises differences between members of a segment and maximises differences between each segment. In addition, the segmentation approach must yield segments that are meaningful for the specific marketing problem or situation. For example, a person's hair colour may be a relevant base for a shampoo manufacturer, but it would not be relevant for a seller of financial services. Selecting the right base requires a good deal of thought and a basic understanding of the market to be segmented.



Typical bases used to segment markets

In reality, marketers can segment the market using any base or variable provided that it is identifiable, measurable, actionable and stable. For example, some fashion houses have segmented the market using women's dress size as a variable. However, the most common bases for segmenting consumer markets include: geographics, demographics, psychographics and behavior. Marketers normally select a single base for the segmentation analysis, although, some bases can be combined into a single segmentation with care. For example, geographics and demographics are often combined, but other bases are rarely combined. Given that psychographics includes demographic variables such as age, gender and income as well as attitudinal and behavioral variables, it makes little logical sense to combine psychographics with demographics or other bases. Any attempt to use combined bases needs careful consideration and a logical foundation.

The following sections provide a detailed description of the most common forms of consumer market segmentation.

Geographic Segmentation

Geographic segmentation divides markets according to geographic criteria. In practice, markets can be segmented as broadly as continents and as narrowly as neighborhoods or postal codes. Typical geographic variables include:

- Country e.g. USA, UK, China, Japan, South Korea, Malaysia, Singapore, Australia, New Zealand
- Region e.g. North, North-west, Mid-west, South, Central
- Population density: e.g. central business district (CBD), urban, suburban, rural, regional
- City or town size: e.g. under 1,000; 1,000- 5,000; 5,000 – 10,000... 1,000,000 – 3,000,000 and over 3,000,000
- Climatic zone: e.g. Mediterranean, Temperate, Sub-Tropical, Tropical, Polar,

The geo-cluster approach (also called *geo-demographic segmentation*) combines demographic data with geographic data to create richer, more detailed profiles. Geo-cluster models break down an area into groups of households based on the assumption that people in any given neighbourhood are likely to exhibit similarities in their demographic composition and other household characteristics.

Geographic segmentation may be considered the first step in international marketing, where marketers must decide whether to adapt their existing products and marketing programs for the unique needs of distinct geographic markets. Tourism Marketing Boards often segment international visitors based on their country of origin. By way of example, Tourism Australia undertakes marketing in 16 core geographic markets; of which China, UK, US, NZ and Japan have been identified as priority segments because they have the greatest potential for growth and are extremely profitable segments with higher than average expenditure per visit. Tourism Australia carries out extensive research on each of these segments and develops rich profiles of high priority segments to better understand their needs and how they make travel decisions. Insights from this analysis are used in travel product development, allocation of promotional budgets, advertising strategy and in broader urban planning decisions. For example, in light of the numbers of Japanese visitors, the city of Melbourne has erected Japanese signage in tourist precincts.

A number of proprietary geo-demographic packages are available for commercial use. Examples include Acorn in the United Kingdom, Experian's Mosaic (geodemography) Segmentation (active in North America, South America, UK, Europe, South Africa and parts of Asia-Pacific) or Helix Personas (Australia, New Zealand and Indonesia). It should be noted that all these commercial packages combine geographics with behavioural, demographic and attitudinal data and yield a very large number of segments. For instance, the NZ Helix Personas segments New Zealand's relatively small population into 51 discrete personality profiles across seven geographic communities (too numerous to itemise on this page). These commercial databases typically allow prospective clients access to limited scale demonstrations, and readers interested in learning more about how geo-demographic segmentation can benefit marketers and planners, are advised to experiment with the online demonstration software via these companies' websites.

Geographic segmentation is widely used in direct marketing campaigns to identify areas which are potential candidates for personal selling, letter-box distribution or direct mail. Geo-cluster segmentation is widely used by Governments and public sector departments such as urban planning, health authorities, police, criminal justice departments, telecommunications and public utility organisations such as water boards.

Demographic Segmentation

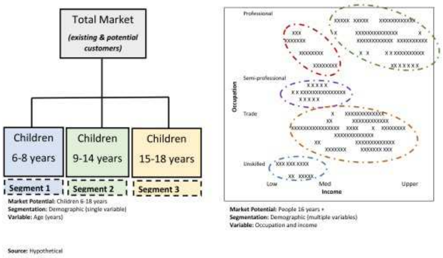
Segmentation according to demography is based on consumer- demographic variables such as age, income, family size, socio-economic status, etc. Demographic segmentation assumes that consumers with similar demographic profiles will exhibit similar purchasing patterns, motivations, interests and lifestyles and that these characteristics will translate into similar product/brand preferences. In practice, demographic segmentation can potentially employ any variable that is used by the nation's census collectors. Typical demographic variables and their descriptors are as follows:

- Age: e.g. Under 5, 5–8 years, 9–12 years, 13–17 years, 18-24, 25-29, 30-39, 40-49, 50-59, 60+
- Gender: Male, Female
- Occupation: Professional, self-employed, semi-professional, clerical/ admin, sales, trades, mining, primary producer, student, home duties, unemployed, retired
- Social class (or socio-economic status):
- Marital Status: Single, married, divorced, widowed
- Family Life-stage: Young single; Young married with no children; Young family with children under 5 years; Older married with children; Older married with no children living at home, Older living alone
- Family size/ number of dependants: 0, 1-2, 3-4, 5+
- Income: Under $10,000; 10,000- 20,000; 20,001- 30,000; 30,001-40,000, 40,001-50,000 etc
- Educational attainment: Primary school; Some secondary, Completed secondary, Some university, Degree; Post graduate or higher degree
- Home ownership: Renting, Own home with mortgage, Home owned outright
- Ethnicity: Asian, African, Aboriginal, Polynesian, Melanesian, Latin-American, African-American, American Indian etc
- Religion: Catholic, Protestant, Muslim, Jewish, Buddhist, Hindu, Other

The Scouting movement offers an excellent example of demographic segmentation in practice. Scouts develops different products based on relevant age groups. In Australia, the segments are Joeys for boys and girls aged 6–7 years; Cubs for children ages 8– 10 years; Scouts for those aged 11–14 years; Venturers ages 15–17 years and Rovers aged 18–25 years. The Scouting movement provides members of each cohort with different uniforms and develops different activity programs for each segment. Scouts even cater to the needs of the over 25s offering them roles as scout leaders or volunteers. Scouts' segmentation is an example of a simple demographic segmentation analysis which utilises just a single variable, namely age.

Scouts develop products for people of all ages

In practice, most demographic segmentation utilises a combination of demographic variables. For instance, a segmentation approach developed for New Zealand by Nielsen Research combines multiple demographic variables including age, life-stage and socio-economic status. The proprietary segmentation product, known as geoTribes, segments the NZ market into 15 tribes, namely: Rockafellas- Affluent mature families; Achievers-Ambitious younger and middle aged families; Fortunats- Financially secure retirees and pre-retirees; Crusaders-Career-oriented singles and couples; Preppies- Mature children of affluent parents; Indpendents- Young singles and couples; Suburban Splendour- Middle class mature families; Twixters- Mature children living at home; Debstars-Financially extended younger families; Boomers -White collar post family pre-retirees; True Blues -Blue collar mature families and pre-retiree singles or couples; Struggleville -Struggling young and middle aged families; Grey Power-Better off retirees; Survivors-Retirees living on minimal incomes and Slender Meanz-People living in underprivileged circumstances.



Market Potential: Children 6-18 years
Segmentation: Demographic (single variable)
Variable: Age (years)

Source: Hypothetical

Market Potential: People 16 years +
Segmentation: Demographic (multiple variables)
Variable: Occupation and income

Visualisation of two approaches to demographic segmentation using one and two variables

The use of multiple segmentation variables normally requires analysis of databases using sophisticated statistical techniques such as cluster analysis or principal components analysis. It should be noted that these types of analysis require very large sample sizes. However, data-collection is expensive for individual firms. For this reason, many companies purchase data from commercial market research firms, many of whom develop proprietary software to interrogate the data. Proprietary packages, such as that mentioned in the preceding geoTribes by Nielsen example, offer clients access to an extensive database along with a program that allows users to interrogate the data via a 'user-friendly' interface. In other words, users do not need a detailed understanding of the 'back-end' statistical procedures used to analyse the data and derive the market segments. However, users still need to be skilled in the interpretation of findings for use in marketing decision-making.

Psychographic Segmentation

Psychographic segmentation, which is sometimes called lifestyle segmentation, is measured by studying the activities, interests, and opinions (AIOs) of customers. It considers how people spend their leisure, and which external influences they are most responsive to and influenced by. Psychographics is a very widely used basis for segmentation, because it enables marketers to identify tightly defined market segments and better understand the motivations for selecting particular products.

One of the most well-known psychographic segmentation analyses is the so-called Values And Lifestyles Segments (VALS), a proprietary psychometric method that measures attitudes, behaviours and demographics that align with brand preferences and new product adoption. The approach was originally developed by SBI International in the 1970s, and the typologies or segments have undergone several revisions over the years. As it currently stands, the VALs segments that describe the adult American population are: Achievers (26%), Strivers (24%), Experiencers (21%), Innovators, Thinkers, Believers, Makers and Survivors. The VALs segments are country specific and the developer offers VALs segmentation typologies for use in China, Japan, the Dominican Republic, Nigeria, UK and Venezuela.

Outside the USA, other countries have developed their own brand of proprietary psychographics segmentation. In Australia and New Zealand, Roy Morgan Research has developed the Values Segments which describes ten segments based on mindset, demographics and behaviours. The Values Segments are: Visible Achievers (20%); Traditional Family Life (19%); Socially Aware (14%); Look-At-Me (11%); Conventional Family Life (10%); Something Better (9%); Young Optimists (7%); Fairer Deal (5%); Real Conservatives (3%) and Basic Needs (2%) Market research company, Nielsen offers PALS (Personal Aspirational Lifestyle Segments), a values-based segmentation that ranks future lifestyle priorities, such as the importance of career, family, wanting to have fun, or the desire for a balanced lifestyle. PALS divides the Australian market into six groups, namely Success Driven (25%); Balance Seekers (23%); Health Conscious (22%), Harmony Seekers (11%), Individualists (11%) and Fun Seekers (8%).

In Britain, the following segments based on British lifestyles were developed by McCann-Erickson: Avant-Guardians (interested in change); Pontificators (traditionalists); Chameleons (follow the crowd) and Sleepwalkers (contented underachievers).

While many of these proprietary psychographic segmentation analyses are well-known, the majority of studies based on psychographics are custom designed. That is the segments are developed for individual products at a specific time. One common thread among psychographic segmentation studies is that they use quirky names to describe the segments.

Behavioral Segmentation

Behavioral segmentation divides consumers into groups according to their observed behaviors. Many marketers believe that behavioral variables are superior to demographics and geographics for building market segments. Typical behavioral variables and their descriptors include:

- Purchase/Usage Occasion: e.g. regular occasion, special occasion, festive occasion, gift-giving
- Benefit-Sought: e.g. economy, quality, service level, convenience, access

- User Status: e.g. First-time user, Regular user, Non-user
- Usage Rate/ Purchase Frequency: e.g. Light user, heavy user, moderate user
- Loyalty Status: e.g. Loyal, switcher, non-loyal, lapsed
- Buyer Readiness: e.g. Unaware, aware, intention to buy
- Attitude to Product or Service: e.g. Enthusiast, Indifferent, Hostile; Price Conscious, Quality Conscious
- Adopter Status: e.g. Early adopter, late adopter, laggard

Note that these descriptors are merely commonly used examples. Marketers must ensure that they customize the variable and descriptors for both local conditions and for specific applications. For example, in the health industry, planners often segment broad markets according to 'health consciousness' and identify low, moderate and highly health conscious segments. This is an example of behavioral segmentation, using attitude to product or service as a key descriptor or variable.

Purchase/ Usage Occasion

Purchase or usage occasion segmentation focuses on analyzing occasions when consumers might purchase or consume a product. This approach customer-level and occasion-level segmentation models and provides an understanding of the individual customers' needs, behavior and value under different occasions of usage and time. Unlike traditional segmentation models, this approach assigns more than one segment to each unique customer, depending on the current circumstances they are under.

For example, Cadbury has segmented the market into five segments based on usage and behavior:

Cadbury segment consumers of their chocolate range according to usage occasion

- Immediate Eat (34%): Driven by the need to snack, indulge or an energy boost. Products that meet these needs include brands such as Kit-Kat, Mars Bars
- Home Stock (25%): Driven by the need to have something in the pantry to share with family in front of TV or for home snacking. Products that meet these needs are blocks or multi-packs
- Kids (17%): Driven by need for after school snacks, parties, treats. Products that meet these needs include Smarties and Milky Bars.

- Gift-giving (15%): Products purchased as gifts, needs include a token of appreciation, a romantic gesture or a special occasion. Products that meet these needs include boxed chocolates such as Cabury's Roses or Quality Street
- Seasonal (3.4%): Driven by need to give a present or create a festive atmosphere. The products are mainly purchased on feast days such as Christmas, Easter, Advent. Products that meet these needs include Easter eggs and Christmas tree decorations.

Benefit-sought

Benefit sought (sometimes called *needs-based segmentation*) divides markets into distinct needs or benefits sought by the consumer.

Other Types of Consumer Segmentation

In addition to geographics, demographics, pyschographics and behavioral bases, marketers occasionally turn to other means of segmenting the market, or to develop segment profiles.

Generational Segments

A generation is defined as "a cohort of people born within a similar span of time (15 years at the upper end) who share a comparable age and life stage and who were shaped by a particular span of time (events, trends and developments)." Generational segmentation refers to the process of dividing and analysing a population into cohorts based on their birth date. Generational segmentation assumes that people's values and attitudes are shaped by the key events that occurred during their lives and that these attitudes translate into product and brand preferences.

Demographers, studying population change, disagree about precise dates for each generation. Dating is normally achieved by identifying population peaks or troughs, which can occur at different times in each country. For example, in Australia the post-war population boom peaked in 1960, while the peak occurred somewhat later in the USA and Europe, with most estimates converging on 1964. Accordingly, Australian Boomers are normally defined as those born between 1945–1960; while American and European Boomers are normally defined as those born between 1945–64. Thus, the generational segments and their dates discussed here must be taken as approximations only.

The primary generational segments identified by marketers are:

- Builders: born 1920 to 1945
- Baby Boomers: born about 1945–1965
- Generation X: born about 1966–1976
- Generation Y: also known as Millenials; born about 1977–1994
- Generation Z: also known as Centennials; born 1995–2015

Table 2: Unique Characteristics of Selected Generations

Millenials	Generation X	Baby Boomers
Technology use (24%)	Technology use (12%)	Work Ethic (17%)
Music/ popular culture (11%)	Work ethic (11%)	Respectful (14%)
Liberal/ tolerant (7%)	Conservative/ traditional (7%)	Values/ morals (8%)
Smarter (6%)	Smarter (6%)	Smarter (5%)
Clothes (5%)	Respectful (5%)	n.a.

Cultural Segmentation

Cultural segmentation is used to classify markets according to cultural origin. Culture is a major dimension of consumer behavior and can be used to enhance customer insight and as a component of predictive models. Cultural segmentation enables appropriate communications to be crafted to particular cultural communities. Cultural segmentation can be applied to existing customer data to measure market penetration in key cultural segments by product, brand, channel as well as traditional measures of recency, frequency and monetary value. These benchmarks form an important evidence-base to guide strategic direction and tactical campaign activity, allowing engagement trends to be monitored over time.

Cultural segmentation can also be mapped according to state, region, suburb and neighborhood. This provides a geographical market view of population proportions and may be of benefit in selecting appropriately located premises, determining territory boundaries and local marketing activities.

Census data is a valuable source of cultural data but cannot meaningfully be applied to individuals. Name analysis (onomastics) is the most reliable and efficient means of describing the cultural origin of individuals. The accuracy of using name analysis as a surrogate for cultural background in Australia is 80-85%, after allowing for female name changes due to marriage, social or political reasons or colonial influence. The extent of name data coverage means a user will code a minimum of 99 percent of individuals with their most likely ancestral origin.

Selecting Target Markets

Selecting the markets to target is a key decision for marketers

Another major decision in developing the segmentation strategy is the selection of market segments that will become the focus of special attention (known as *target markets*). The marketer faces a number of important decisions:

- What criteria should be used to evaluate markets?
- How many markets to enter (one, two or more)?

- Which market segments are the most valuable?

When a marketer enters more than one market, the segments are often labelled the *primary target market*, *secondary target market*. The primary market is the target market selected as the main focus of marketing activities. The secondary target market is likely to be a segment that is not as large as the primary market, but has growth potential. Alternatively, the secondary target group might consist of a small number of purchasers that account for a relatively high proportion of sales volume perhaps due to purchase value or purchase frequency.

In terms of evaluating markets, three core considerations are essential:

- Segment size and growth
- Segment structural attractiveness
- Company objectives and resources.

There are no formulas for evaluating the attractiveness of market segments and a good deal of judgement must be exercised. Nevertheless, a number of considerations can be used to evaluate market segments for attractiveness.

Segment Size and Growth:

- How large is the market?
- Is the market segment substantial enough to be profitable?
- Segment size can be measured in number of customers, but superior measures are likely to include sales value or volume
- Is the market segment growing or contracting?
- What are the indications that growth will be sustained in the long term? Is any observed growth sustainable?
- Is the segment stable over time? (Segment must have sufficient time to reach desired performance level)

Segment Structural Attractiveness:

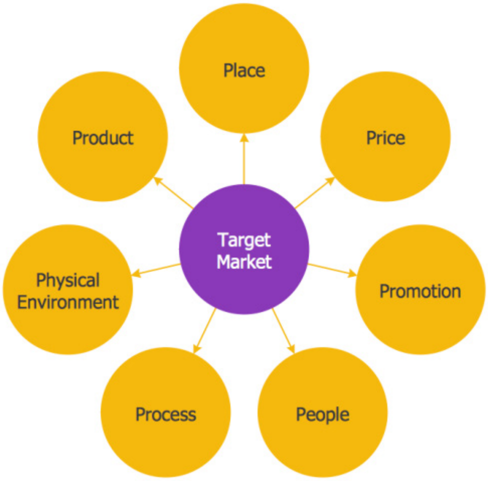
- To what extent are competitors targeting this market segment?
- Can we carve out a viable position to differentiate from any competitors?
- How responsive are members of the market segment to the marketing program?
- Is this market segment reachable and accessible? (i.e., with respect to distribution and promotion)

Company Objectives and Resources:

- Is this market segment aligned with our company's operating philosophy?

- Do we have the resources necessary to enter this market segment?
- Do we have prior experience with this market segment or similar market segments?
- Do we have the skills and/or know-how to enter this market segment successfully?

Market Segmentation and the Marketing Program



The marketing program is designed with the needs of the target market in mind

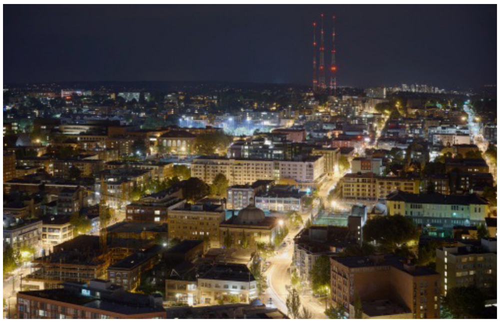
When the segments have been determined and separate offers developed for each of the core segments, the marketer's next task is to design a marketing program (also known as the marketing mix) that will resonate with the target market or markets. Developing the marketing program requires a deep knowledge of key market segment's purchasing habits, their preferred retail outlet, their media habits and their price sensitivity. The marketing program for each brand or product should be based on the understanding of the target market (or target markets) revealed in the market profile.

Table 3 provides a brief summary of the marketing mix used by Black and Decker, manufacturer of three brands of power tools and household appliances. Each brand is pitched at different quality points and targets different segments of the market. This table is designed to illustrate how the marketing program must be fine-tuned for each market segment.

Table 3: Black & Decker Market Segmentation

Segment	Product/ Brand	Product Strategy	Price Strategy	Promotion Strategy	Place Strategy
Home-owners/ D-I-Y	Black & Decker	Quality Adequate for occasional use	Lower priced	TV advertising during holidays	Mass distribution (lower-tier stores)
Weekend Warriors	Firestorm	Quality Adequate for regular use	Higher priced	Ads in D-I-Y magazines, shows	Selective distribution (top tier hardware stores)
Professional users	De Walt	Quality adequate for daily use	Highest price	Personal selling (sales reps call on job sites)	Selective distribution (top tier hardware stores)

Bases for Segmenting Business Markets



Businesses can be segmented using industry, size, turnover or any other meaningful variable

Businesses may be segmented according to industry, business size, turnover, number of employees or any other relevant variables.

Firmographics

Firmographics (also known as *emporographics* or *feature based segmentation*) is the business community's answer to demographic segmentation. It is commonly used in business-to-business markets (it's estimated that 81% of B2B marketers use this technique). Under this approach the target market is segmented based on features such as company size (either in terms of revenue or number of employees), industry sector or location (country and/or region).

Multi-variable Account Segmentation

In Sales Territory Management, using more than one criterion to characterize the organization's accounts, such as segmenting sales accounts by government, business, customer, etc. and account size or duration, in effort to increase time efficiency and sales volume.

Using Segmentation in Customer Retention

The basic approach to retention-based segmentation is that a company tags each of its active customers with four values:

Is this customer at high risk of canceling the company's service?

One of the most common indicators of high-risk customers is a drop off in usage of the company's service. For example, in the credit card industry this could be signaled through a customer's decline in spending on his or her card.

Is this customer at high risk of switching to a competitor to purchase product?

Many times customers move purchase preferences to a competitor brand. This may happen for many reasons those of which can be more difficult to measure. It is many times beneficial for the former company to gain meaningful insights, through data analysis, as to why this change of preference has occurred. Such insights can lead to effective strategies for winning back the customer or on how not to lose the target customer in the first place.

Is this customer worth retaining?

This determination boils down to whether the post-retention profit generated from the customer is predicted to be greater than the cost incurred to retain the customer, and includes evaluation of customer lifecycles.

What retention tactics should be used to retain this customer?

This analysis of customer lifecycles is usually included in the growth plan of a business to determine which tactics to implement to retain or let go of customers. Tactics commonly used range from providing special customer discounts to sending customers communications that reinforce the value proposition of the given service.

Segmentation: Algorithms and Approaches

The choice of an appropriate statistical method for the segmentation, depends on a number of factors including, the broad approach (a-priori or post-hoc), the availability of data, time constraints, the marketer's skill level and resources.

A-priori Segmentation

According to the Market Research Association (MRA), a priori research occurs when "a theoretical framework is developed before the research is conducted". In other words, the marketer has an idea about whether to segment the market geographically, demographically, psychographically or behaviorally before undertaking any research. For example, a marketer might want to learn more about the motivations and demographics of light and moderate users in an effort to understand what tactics could be used to increase usage rates. In this case, the target variable is known – the marketer has already segmented using a behavioral variable – user status. The next step would be to collect and analyse attitudinal data for light and moderate users. Typical analysis includes simple cross-tabulations, frequency distributions and occasionally logistic regression or CHAID analysis.

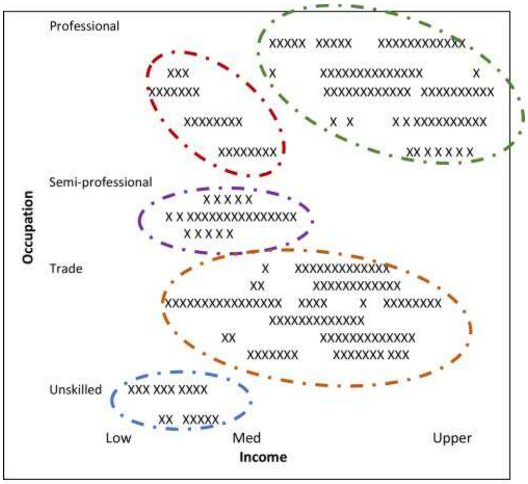
The main disadvantage of a-priori segmentation is that it does not explore other opportunities to identify market segments that could be more meaningful.

Post-hoc Segmentation

In contrast, post-hoc segmentation makes no assumptions about the optimal theoretical framework. Instead, the analyst's role is to determine the segments that are the most meaningful for a given marketing problem or situation. In this approach, the empirical data drives the segmentation selection. Analysts typically employ some type of clustering analysis or structural equation modeling to identify segments within the data. The figure alongside illustrates how segments might be formed using clustering, however note that this diagram only uses two variables, while in practice clustering employs a large number of variables. Post-hoc segmentation relies on access to rich data sets, usually with a very large number of cases.

Statistical Techniques used in Segmentation

Marketers often engage commercial research firms or consultancies to carry out segmentation analysis, especially if they lack the statistical skills to undertake the analysis. Some segmentation, especially post-hoc analysis, relies on sophisticated statistical analysis.



Market Potential (Universe): People 16 years +
Segmentation: Demographic (multiple variables)
Variables: Occupation and income

Source: Hypothetical

Visualisation of market segments formed using clustering methods.

Common statistical techniques for segmentation analysis include:

- Clustering algorithms such as K-means or other Cluster analysis
- Conjoint Analysis
- Logistic Regression (also known as Logit Regression)
- Chi-square Automatic Interaction Detector CHAID; a type of decision-tree
- Structural Equation Modeling (SEM)
- Multidimensional scaling and Canonical Analysis
- Statistical mixture models such as Latent Class Analysis
- Ensemble approaches such as Random Forests
- Other algorithms such as Neural Networks

Data Sources used for Segmentation

Marketers use a variety of data sources for segmentation studies and market profiling. Typical sources of information include:

Internal Databases

- Customer transaction records e.g. sale value per transaction, purchase frequency
- Patron membership records e.g. active members, lapsed members, length of membership
- Customer relationship management (CRM) databases
- In-house surveys

External Sources

- Commercial Surveys or tracking studies (available from major research companies such as Nielsen and Roy Morgan)
- Government agencies or Professional/ Industry associations
- Census data
- Observed purchase behavior – collected from online agencies such as Google
- Data-mining techniques

Market Trend

A market trend is a perceived tendency of financial markets to move in a particular direction over time. These trends are classified as *secular* for long time frames, *primary* for medium time frames, and *secondary* for short time frames. Traders attempt to identify market trends using technical analysis, a framework which characterizes market trends as predictable price tendencies within the market when price reaches support and resistance levels, varying over time.

A trend can only be determined in hindsight, since at any time prices in the future are not known.

Market Nomenclature

The terms “bull market” and “bear market” describe upward and downward market trends, respectively, and can be used to describe either the market as a whole or specific sectors and securities. The names perhaps correspond to the fact that a bull attacks by lifting its horns upward, while a bear strikes with its claws in a downward motion.

Etymology

The fighting styles of both animals may have a major impact on the names.

One hypothetical etymology points to London bearskin “jobbers” (market makers), who would sell bearskins before the bears had actually been caught in contradiction of the proverb *ne vendez pas la peau de l’ours avant de l’avoir tué* (“don’t sell the bearskin before you’ve killed the bear”)—an admonition against over-optimism. By the time of the South Sea Bubble of 1721, the bear was also associated with short selling; jobbers would sell bearskins they did not own in anticipation of falling prices, which would enable them to buy them later for an additional profit.

Some analogies that have been used as mnemonic devices:

- Bull is short for “bully”, in its now somewhat dated meaning of “excellent”.
- It relates to the speed of the animals: Bulls usually charge at very high speed, whereas bears normally are thought of as lazy and cautious movers—a misconception, because a bear, under the right conditions, can outrun a horse.

- They were originally used in reference to two old merchant banking families, the Barings and the Bulstrodes.
- The word “bull” plays off the market’s returns being “full”, whereas “bear” alludes to the market’s returns being “bare”.
- “Bull” symbolizes charging ahead with excessive confidence, whereas “bear” symbolizes preparing for winter and hibernation in doubt.

Secular Trends

A secular market trend is a long-term trend that lasts 5 to 25 years and consists of a series of primary trends. A secular bear market consists of smaller bull markets and larger bear markets; a secular bull market consists of larger bull markets and smaller bear markets.

In a secular bull market the prevailing trend is “bullish” or upward-moving. The United States stock market was described as being in a secular bull market from about 1983 to 2000 (or 2007), with brief upsets including the crash of 1987 and the market collapse of 2000-2002 triggered by the dot-com bubble.

In a secular bear market, the prevailing trend is “bearish” or downward-moving. An example of a secular bear market occurred in gold between January 1980 to June 1999, culminating with the Brown Bottom. During this period the nominal gold price fell from a high of $850/oz ($30/g) to a low of $253/oz ($9/g), and became part of the Great Commodities Depression.

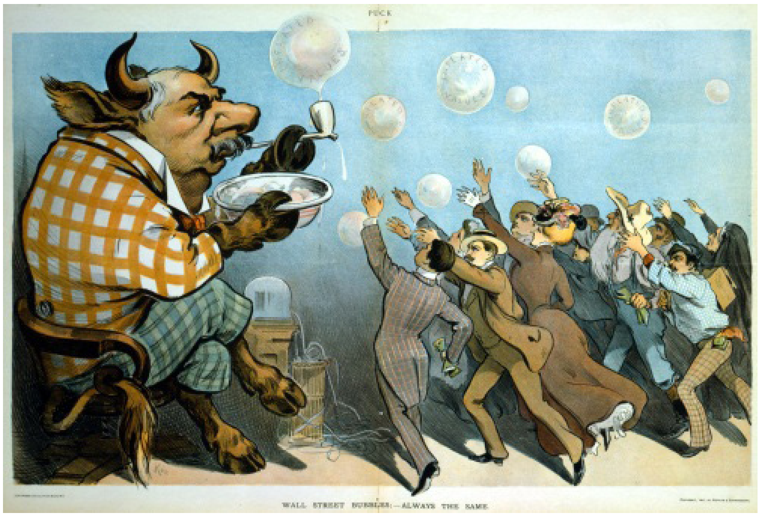
Primary Trends

Statues of the two symbolic beasts of finance, the bear and the bull, in front of the Frankfurt Stock Exchange.

A primary trend has broad support throughout the entire market (most sectors) and lasts for a year or more.

Bull Market

A bull market is a period of generally rising prices. The start of a bull market is marked by widespread pessimism. This point is when the “crowd” is the most “bearish”. The feeling of despondency changes to hope, “optimism”, and eventually euphoria, as the bull runs its course. This often leads the economic cycle, for example in a full recession, or earlier.


PUCK
WALL STREET BUBBLES:—ALWAYS THE SAME

A 1901 cartoon depicting financier J. P. Morgan as a bull with eager investors

An analysis of Morningstar, Inc. stock market data from 1926 to 2014 found that a typical bull market "lasted 8.5 years with an average cumulative total return of 458%", while annualized gains for bull markets range from 14.9% to 34.1%.

Examples

India's Bombay Stock Exchange Index, BSE SENSEX, was in a bull market trend for about five years from April 2003 to January 2008 as it increased from 2,900 points to 21,000 points. Notable bull markets marked the 1925-1929, 1953–1957 and the 1993-1997 periods when the U.S. and many other stock markets rose; while the first period ended abruptly with the start of the Great Depression, the end of the later time periods were mostly periods of soft landing, which became large bear markets.

Bear Market

A bear market is a general decline in the stock market over a period of time. It is a transition from high investor optimism to widespread investor fear and pessimism. According to The Vanguard Group, "While there's no agreed-upon definition of a bear market, one generally accepted measure is a price decline of 20% or more over at least a two-month period."

An analysis of Morningstar, Inc. stock market data from 1926 to 2014 found that a typical bear market "lasted 1.3 years with an average cumulative loss of -41%", while annualized declines for bear markets range from -19.7% to -47%.

Examples

A bear market followed the Wall Street Crash of 1929 and erased 89% (from 386 to 40) of the Dow Jones Industrial Average's market capitalization by July 1932, marking the start of the Great Depression. After regaining nearly 50% of its losses, a longer bear market from 1937 to 1942 occurred in which the market was again cut in half. Another long-term bear market occurred from about 1973 to 1982, encompassing the 1970s energy crisis and the high unemployment of the early 1980s. Yet another bear market occurred between March 2000 and October 2002. Recent examples occurred between October 2007 and March 2009, as a result of the financial crisis of 2007–08.

Market Top

A market top (or market high) is usually not a dramatic event. The market has simply reached the highest point that it will, for some time (usually a few years). It is retroactively defined as market participants are not aware of it as it happens. A decline then follows, usually gradually at first and later with more rapidity. William J. O'Neil and company report that since the 1950s a market top is characterized by three to five distribution days in a major market index occurring within a relatively short period of time. Distribution is a decline in price with higher volume than the preceding session.

Examples

The peak of the dot-com bubble (as measured by the NASDAQ-100) occurred on March 24, 2000. The index closed at 4,704.73. The Nasdaq peaked at 5,132.50 and the S&P 500 at 1525.20.

A recent peak for the broad U.S. market was October 9, 2007. The S&P 500 index closed at 1,565 and the Nasdaq at 2861.50.

Market Bottom

A market bottom is a trend reversal, the end of a market downturn, and the beginning of an upward moving trend (bull market).

It is very difficult to identify a bottom (referred to by investors as "bottom picking") while it is occurring. The upturn following a decline is often short-lived and prices might resume their decline. This would bring a loss for the investor who purchased stock(s) during a misperceived or "false" market bottom.

Baron Rothschild is said to have advised that the best time to buy is when there is "blood in the streets", i.e., when the markets have fallen drastically and investor sentiment is extremely negative.

Examples

Some examples of market bottoms, in terms of the closing values of the Dow Jones Industrial Average (DJIA) include:

- The Dow Jones Industrial Average hit a bottom at 1738.74 on 19 October 1987, as a result of the decline from 2722.41 on 25 August 1987. This day was called Black Monday (chart).
- A bottom of 7286.27 was reached on the DJIA on 9 October 2002 as a result of the decline from 11722.98 on 14 January 2000. This included an intermediate bottom of 8235.81 on 21 September 2001 (a 14% change from 10 September) which led to an intermediate top of 10635.25 on 19 March 2002 (chart). The "tech-heavy" Nasdaq fell a more precipitous 79% from its 5132 peak (10 March 2000) to its 1108 bottom (10 October 2002).

- A bottom of 6,440.08 (DJIA) on 9 March 2009 was reached after a decline associated with the subprime mortgage crisis starting at 14164.41 on 9 October 2007 (chart).

Secondary Trends

Secondary trends are short-term changes in price direction within a primary trend. The duration is a few weeks or a few months.

One type of secondary market trend is called a market correction. A correction is a short term price decline of 5% to 20% or so. An example occurred from April to June 2010, when the S&P 500 went from above 1200 to near 1000; this was hailed as the end of the bull market and start of a bear market, but it was not, and the market turned back up. A correction is a downward movement that is not large enough to be a bear market (ex post).

Another type of secondary trend is called a bear market rally (sometimes called “sucker’s rally” or “dead cat bounce”) which consist of a market price increase of only 10% or 20% and then the prevailing, bear market trend resumes. Bear market rallies occurred in the Dow Jones index after the 1929 stock market crash leading down to the market bottom in 1932, and throughout the late 1960s and early 1970s. The Japanese Nikkei 225 has been typified by a number of bear market rallies since the late 1980s while experiencing an overall long-term downward trend.

The Australian market in the beginning of 2015 has been described as a “meerkat market”, being timid with low consumer and business sentiment.

Causes

The price of assets such as stocks is set by supply and demand. By definition, the market balances buyers and sellers, so it’s impossible to literally have ‘more buyers than sellers’ or vice versa, although that is a common expression. For a surge in demand, the buyers will increase the price they are willing to pay, while the sellers will increase the price they wish to receive. For a surge in supply, the opposite happens.

Supply and demand are created when investors shift allocation of investment between asset types. For example, at one time, investors may move money from government bonds to “tech” stocks; at another time, they may move money from “tech” stocks to government bonds. In each case, this will affect the price of both types of assets.

Generally, investors try to follow a buy-low, sell-high strategy but often mistakenly end up buying high and selling low. Contrarian investors and traders attempt to “fade” the investors’ actions (buy when they are selling, sell when they are buying). A time when most investors are selling stocks is known as distribution, while a time when most investors are buying stocks is known as accumulation.

According to standard theory, a decrease in price will result in less supply and more demand, while an increase in price will do the opposite. This works well for most assets but it often works in reverse for stocks due to the mistake many investors make of buying high in a state of euphoria and selling low in a state of fear or panic as a result of the herding instinct. In case an increase in price causes an increase in demand, or a decrease in price causes an increase in supply, this destroys the expected negative feedback loop and prices will be unstable. This can be seen in a bubble or crash.

Investor Sentiment

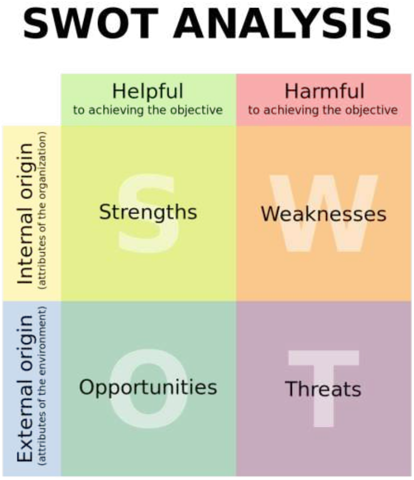
Investor sentiment is a contrarian stock market indicator.

When a high proportion of investors express a bearish (negative) sentiment, some analysts consider it to be a strong signal that a market bottom may be near. The predictive capability of such a signal is thought to be highest when investor sentiment reaches extreme values. Indicators that measure investor sentiment may include:

David Hirshleifer sees in the trend phenomenon a path starting with underreaction and ending in overreaction by investors / traders.

- Investor Intelligence Sentiment Index: If the Bull-Bear spread (% of Bulls - % of Bears) is close to a historic low, it may signal a bottom. Typically, the number of bears surveyed would exceed the number of bulls. However, if the number of bulls is at an extreme high and the number of bears is at an extreme low, historically, a market top may have occurred or is close to occurring. This contrarian measure is more reliable for its coincidental timing at market lows than tops.
- American Association of Individual Investors (AAII) sentiment indicator: Many feel that the majority of the decline has already occurred once this indicator gives a reading of minus 15% or below.
- Other sentiment indicators include the Nova-Ursa ratio, the Short Interest/Total Market Float, and the put/call ratio.

SWOT Analysis


SWOT ANALYSIS
Helpful
to achieving the objective
Harmful
to achieving the objective
Internal origin
(attributes of the organization)
External origin
(attributes of the environment)
S
Strengths
W
Weaknesses
O
Opportunities
T
Threats

A SWOT analysis, with its four elements in a 2×2 matrix.

SWOT analysis (alternatively SWOT matrix) is an acronym for *strengths*, *weaknesses*, *opportunities*, and *threats* and is a structured planning method that evaluates those four elements of a project or business venture. A SWOT analysis can be carried out for a company, product, place, industry, or person. It involves specifying the objective of the business venture or project and identifying the internal and external factors that are favorable and unfavorable to achieve that objective. Some authors credit SWOT to Albert Humphrey, who led a convention at the Stanford Research Institute (now SRI International) in the 1960s and 1970s using data from Fortune 500 companies. However, Humphrey himself did not claim the creation of SWOT, and the origins remain obscure. The degree to which the internal environment of the firm matches with the external environment is expressed by the concept of strategic fit.

- Strengths: characteristics of the business or project that give it an advantage over others
- Weaknesses: characteristics that place the business or project at a disadvantage relative to others
- Opportunities: elements in the environment that the business or project could exploit to its advantage
- Threats: elements in the environment that could cause trouble for the business or project

Identification of SWOTs is important because they can inform later steps in planning to achieve the objective. First, decision-makers should consider whether the objective is attainable, given the SWOTs. If the objective is *not* attainable, they must select a different objective and repeat the process.

Users of SWOT analysis must ask and answer questions that generate meaningful information for each category (strengths, weaknesses, opportunities, and threats) to make the analysis useful and find their competitive advantage.

Internal and External Factors

So it is said that if you know your enemies and know yourself, you can win a hundred battles without a single loss. If you only know yourself, but not your opponent, you may win or may lose. If you know neither yourself nor your enemy, you will always endanger yourself.

The Art of War by Sun Tzu

SWOT analysis aims to identify the key internal and external factors seen as important to achieving an objective. SWOT analysis groups key pieces of information into two main categories:

1. Internal factors – the *strengths* and *weaknesses* internal to the organization
2. External factors – the *opportunities* and *threats* presented by the environment external to the organization

Analysis may view the internal factors as strengths or as weaknesses depending upon their effect on the organization’s objectives. What may represent strengths with respect to one objective may be weaknesses (distractions, competition) for another objective. The factors may include all of the 4Ps as well as personnel, finance, manufacturing capabilities, and so on.

The external factors may include macroeconomic matters, technological change, legislation, and sociocultural changes, as well as changes in the marketplace or in competitive position. The results are often presented in the form of a matrix.

SWOT analysis is just one method of categorization and has its own weaknesses. For example, it may tend to persuade its users to compile lists rather than to think about actual important factors in achieving objectives. It also presents the resulting lists uncritically and without clear prioritization so that, for example, weak opportunities may appear to balance strong threats.

It is prudent not to eliminate any candidate SWOT entry too quickly. The importance of individual SWOTs will be revealed by the value of the strategies they generate. A SWOT item that produces valuable strategies is important. A SWOT item that generates no strategies is not important.

Use

The usefulness of SWOT analysis is not limited to profit-seeking organizations. SWOT analysis may be used in any decision-making situation when a desired end-state (objective) is defined. Examples include non-profit organizations, governmental units, and individuals. SWOT analysis may also be used in pre-crisis planning and preventive crisis management. SWOT analysis may also be used in creating a recommendation during a viability study/survey.

Strategy Building

SWOT analysis can be used effectively to build organizational or personal strategy. Steps necessary to execute strategy-oriented analysis involve identification of internal and external factors (using popular 2x2 matrix), selection and evaluation of the most important factors, and identification of relations existing between internal and external features.

For instance, strong relations between strengths and opportunities can suggest good conditions in the company and allow using an *aggressive* strategy. On the other hand, strong interactions between weaknesses and threats could be analyzed as a potential warning and advice for using a *defensive* strategy. The analysis of these relationships to determine which strategy to implement is often performed in the growth planning phase for a business.

Matching and Converting

One way of utilizing SWOT is matching and converting. Matching is used to find competitive advantage by matching the strengths to opportunities. Another tactic is to convert weaknesses or threats into strengths or opportunities. An example of a conversion strategy is to find new markets. If the threats or weaknesses cannot be converted, a company should try to minimize or avoid them.

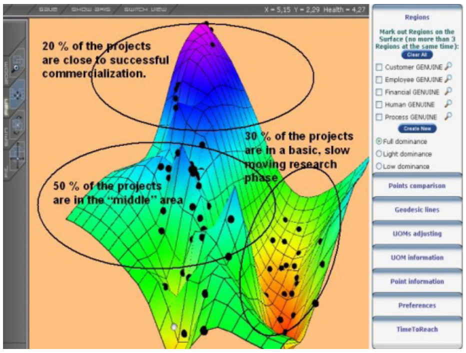
SWOT Variants

Various complementary analyses to SWOT have been proposed, such as the Growth-share matrix and Porter's five forces analysis.

TOWS

Heinz Weihrich said that some users found it difficult to translate the results of the SWOT analysis into meaningful actions that could be adopted within the wider corporate strategy. He introduced the TOWS Matrix, a conceptual framework that helps in finding the most efficient actions.

SWOT Landscape Analysis



20 % of the projects are close to successful commercialization.

30 % of the projects are in a basic, slow moving research phase

50 % of the projects are in the "middle" area

SAVE
SHOW AXIS
SWITCH VIEW
X = 5,15 Y = 2,29 Health = 4,27
ZOOM
PAN
SPIN
FIT
Regions
Mark out Regions on the Surface (no more than 3 Regions at the same time):
Clear All
☐ Customer GENUINE
☐ Employee GENUINE
☐ Financial GENUINE
☐ Human GENUINE
☐ Process GENUINE
Create New
● Full dominance
○ Light dominance
○ Low dominance
Points comparison
Geodesic lines
UOMs adjusting
UOM information
Point information
Preferences
TimeToReach

The SWOT landscape systematically deploys the relationships between overall objective and underlying SWOT-factors and provides an interactive, query-able 3D landscape.

The SWOT landscape graphs differentiate managerial situations by visualizing and foreseeing the dynamic performance of comparable objects according to findings by Brendan Kitts, Leif Edvinsson, and Tord Beding (2000).

Changes in relative performance are continually identified. Projects (or other units of measurements) that could be potential risk or opportunity objects are highlighted.

SWOT landscape also indicates which underlying strength and weakness factors have influence or likely will have highest influence in the context of value in use (e.g., capital value fluctuations).

Corporate Planning

As part of the development of strategies and plans to enable the organization to achieve its objectives, that organization will use a systematic/rigorous process known as corporate planning. SWOT alongside PEST/PESTLE can be used as a basis for the analysis of business and environmental factors.

- Set objectives – defining what the organization is going to do
- Environmental scanning
    - Internal appraisals of the organization's SWOT, this needs to include an assessment of the present situation as well as a portfolio of products/services and an analysis of the product/service life cycle
- Analysis of existing strategies, this should determine relevance from the results of an internal/external appraisal. This may include gap analysis of environmental factors
- Strategic Issues defined – key factors in the development of a corporate plan that the organization must address
- Develop new/revised strategies – revised analysis of strategic issues may mean the objectives need to change
- Establish critical success factors – the achievement of objectives and strategy implementation
- Preparation of operational, resource, projects plans for strategy implementation

- Monitoring results – mapping against plans, taking corrective action, which may mean amending objectives/strategies

Marketing

In many competitor analyses, marketers build detailed profiles of each competitor in the market, focusing especially on their relative competitive strengths and weaknesses using SWOT analysis. Marketing managers will examine each competitor's cost structure, sources of profits, resources and competencies, competitive positioning and product differentiation, degree of vertical integration, historical responses to industry developments, and other factors.

Marketing management often finds it necessary to invest in research to collect the data required to perform accurate marketing analysis. Accordingly, management often conducts market research (alternately marketing research) to obtain this information. Marketers employ a variety of techniques to conduct market research, but some of the more common include:

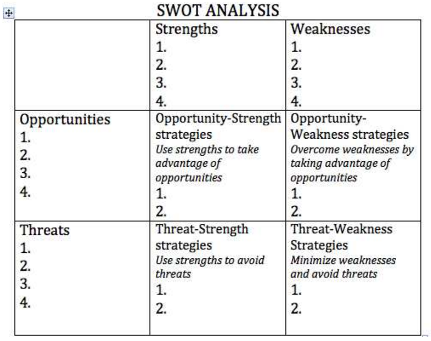
- Qualitative marketing research such as focus groups
- Quantitative marketing research such as statistical surveys
- Experimental techniques such as test markets
- Observational techniques such as ethnographic (on-site) observation
- Marketing managers may also design and oversee various environmental scanning and competitive intelligence processes to help identify trends and inform the company's marketing analysis.

Below is an example SWOT analysis of a market position of a small management consultancy with specialism in HRM.

Strengths	Weaknesses	Opportunities	Threats
Reputation in marketplace	Shortage of consultants at operating level rather than partner level	Well established position with a well-defined market niche	Large consultancies operating at a minor level
Expertise at partner level in HRM consultancy	Unable to deal with multidisciplinary assignments because of size or lack of ability	Identified market for consultancy in areas other than HRM	Other small consultancies looking to invade the marketplace

In Community Organization

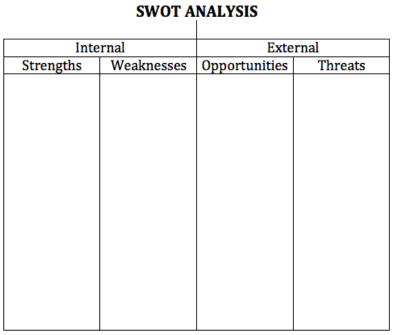
The SWOT analysis has been utilized in community work as a tool to identify positive and negative factors within organizations, communities, and the broader society that promote or inhibit successful implementation of social services and social change efforts. It is used as a preliminary resource, assessing strengths, weaknesses, opportunities, and threats in a community served by a nonprofit or community organization. This organizing tool is best used in collaboration with community workers and/or community members before developing goals and objectives for a program design or implementing an organizing strategy.The SWOT analysis is a part of the planning for social change process and will not provide a strategic plan if used by itself. After a SWOT analysis is completed, a social change organization can turn the SWOT list into a series of recommendations to consider before developing a strategic plan.



SWOT ANALYSIS

	Strengths 1. 2. 3. 4.	Weaknesses 1. 2. 3. 4.
Opportunities 1. 2. 3. 4.	Opportunity-Strength strategies *Use strengths to take advantage of opportunities* 1. 2.	Opportunity-Weakness strategies *Overcome weaknesses by taking advantage of opportunities* 1. 2.
Threats 1. 2. 3. 4.	Threat-Strength strategies *Use strengths to avoid threats* 1. 2.	Threat-Weakness Strategies *Minimize weaknesses and avoid threats* 1. 2.

one example of a SWOT Analysis used in community organizing



SWOT ANALYSIS

Internal		External	
Strengths	Weaknesses	Opportunities	Threats

A simple SWOT Analysis used in Community Organizing

Strengths and Weaknesses: *These are the internal factors within an organization.*

- Human resources - staff, volunteers, board members, target population
- Physical resources - your location, building, equipment
- Financial - grants, funding agencies, other sources of income
- Activities and processes - programs you run, systems you employ
- Past experiences - building blocks for learning and success, your reputation in the community

Opportunities and Threats: *These are external factors stemming from community or societal forces.*

- Future trends in your field or the culture
- The economy - local, national, or international
- Funding sources - foundations, donors, legislatures
- Demographics - changes in the age, race, gender, culture of those you serve or in your area
- The physical environment (Is your building in a growing part of town? Is the bus company cutting routes?)

- Legislation (Do new federal requirements make your job harder...or easier?)
- Local, national, or international events

Although the SWOT analysis was originally designed as an organizational method for business and industries, it has been replicated in various community work as a tool for identifying external and internal support to combat internal and external opposition. The SWOT analysis is necessary to provide direction to the next stages of the change process. It has been utilized by community organizers and community members to further social justice in the context of Social Work practice.

Application in Community Organization

Elements to Consider

Elements to consider in a SWOT analysis include understanding the community that a particular organization is working with. This can be done via public forums, listening campaigns, and informational interviews. Data collection will help inform the community members and workers when developing the SWOT analysis. A needs and assets assessment are tooling that can be used to identify the needs and existing resources of the community. When these assessments are done and data has been collected, an analysis of the community can be made that informs the SWOT analysis.

Steps for Implementation

A SWOT analysis is best developed in a group setting such as a work or community meeting. A facilitator can conduct the meeting by first explaining what a SWOT analysis is as well as identifying the meaning of each term.

One way of facilitating the development of a SWOT analysis includes developing an example SWOT with the larger group then separating each group into smaller teams to present to the larger group after set amount of time. This allows for individuals, who may be silenced in a larger group setting, to contribute. Once the allotted time is up, the facilitator may record all the factors of each group onto a large document such as a poster board, and then the large group, as a collective, can go work through each of the threats and weaknesses to explore options that may be used to combat negative forces with the strengths and opportunities present within the organization and community. A SWOT meeting allows participants to creatively brainstorm, identify obstacles, and possibly strategize solutions/way forward to these limitations.

When to use SWOT Analysis

The uses of a SWOT analysis by a community organization are as follows: to organize information, provide insight into barriers that may be present while engaging in social change processes, and identify strengths available that can be activated to counteract these barriers.

- *A SWOT analysis can be used to:*
- Explore new solutions to problems
- Identify barriers that will limit goals/objectives
- Decide on direction that will be most effective

- Reveal possibilities and limitations for change
- To revise plans to best navigate systems, communities, and organizations
- As a brainstorming and recording device as a means of communication
- To enhance “credibility of interpretation” to be utilized in presentation to leaders or key supporters.

Benefits

The SWOT analysis in social work practice framework is beneficial because it helps organizations decide whether or not an objective is obtainable and therefore enables organizations to set achievable goals, objectives, and steps to further the social change or community development effort. It enables organizers to take visions and produce practical and efficient outcomes that effect long-lasting change, and it helps organizations gather meaningful information to maximize their potential. Completing a SWOT analysis is a useful process regarding the consideration of key organizational priorities, such as gender and cultural diversity and fundraising objectives.

Limitations

Some findings from Menon et al. (1999) and Hill and Westbrook (1997) have suggested that SWOT may harm performance and that “no-one subsequently used the outputs within the later stages of the strategy”.

Other critiques include the misuse of the SWOT analysis as a technique that can be quickly designed without critical thought leading to a misrepresentation of strengths, weaknesses, opportunities, and threats within an organization’s internal and external surroundings.

Another limitation includes the development of a SWOT analysis simply to defend previously decided goals and objectives. This misuse leads to limitations on brainstorming possibilities and “real” identification of barriers. This misuse also places the organization’s interest above the well-being of the community. Further, a SWOT analysis should be developed as a collaborative with a variety of contributions made by participants including community members. The design of a SWOT analysis by one or two community workers is limiting to the realities of the forces, specifically external factors, and devalues the possible contributions of community members.

Marketing Research

Marketing research is “the process or set of processes that links the producers, customers, and end users to the marketer through information — information used to identify and define marketing opportunities and problems; generate, refine, and evaluate marketing actions; monitor marketing performance; and improve understanding of marketing as a process. Marketing research specifies the information required to address these issues, designs the method for collecting information, manages and implements the data collection process, analyzes the results, and communicates the findings and their implications.”

It is the systematic gathering, recording, and analysis of qualitative and quantitative data about issues relating to marketing products and services. The goal of marketing research is to identify and assess how changing elements of the marketing mix impacts customer behavior. The term is commonly interchanged with market research; however, expert practitioners may wish to draw a distinction, in that *market* research is concerned specifically with markets, while *marketing* research is concerned specifically about marketing processes.

Marketing research is often partitioned into two sets of categorical pairs, either by target market:

- Consumer marketing research, and
- Business-to-business (B2B) marketing research

Or, alternatively, by methodological approach:

- Qualitative marketing research, and
- Quantitative marketing research

Consumer marketing research is a form of applied sociology that concentrates on understanding the preferences, attitudes, and behaviors of consumers in a market-based economy, and it aims to understand the effects and comparative success of marketing campaigns. The field of consumer marketing research as a statistical science was pioneered by Arthur Nielsen with the founding of the ACNielsen Company in 1923.

Thus, marketing research may also be described as the systematic and objective identification, collection, analysis, and dissemination of information for the purpose of assisting management in decision making related to the identification and solution of problems and opportunities in marketing.

Role

The task of marketing research (MR) is to provide management with relevant, accurate, reliable, valid, and current market information. Competitive marketing environment and the ever-increasing costs attributed to poor decision making require that marketing research provide sound information. Sound decisions are not based on gut feeling, intuition, or even pure judgment.

Managers make numerous strategic and tactical decisions in the process of identifying and satisfying customer needs. They make decisions about potential opportunities, target market selection, market segmentation, planning and implementing marketing programs, marketing performance, and control. These decisions are complicated by interactions between the controllable marketing variables of product, pricing, promotion, and distribution. Further complications are added by uncontrollable environmental factors such as general economic conditions, technology, public policies and laws, political environment, competition, and social and cultural changes. Another factor in this mix is the complexity of consumers. Marketing research helps the marketing manager link the marketing variables with the environment and the consumers. It helps remove some of the uncertainty by providing relevant information about the marketing variables, environment, and consumers. In the absence of relevant information, consumers' response to marketing programs cannot be predicted reliably or accurately. Ongoing marketing research programs provide information on controllable and non-controllable factors and consumers; this information enhances the effectiveness of decisions made by marketing managers.

Traditionally, marketing researchers were responsible for providing the relevant information and marketing decisions were made by the managers. However, the roles are changing and marketing researchers are becoming more involved in decision making, whereas marketing managers are becoming more involved with research. The role of marketing research in managerial decision making is explained further using the framework of the “DECIDE” model.

History

Marketing research has evolved in the decades since Arthur Nielsen established it as a viable industry, one that would grow hand-in-hand with the B2B and B2C economies. Markets naturally evolve, and since the birth of ACNielsen, when research was mainly conducted by in-person focus groups and pen-and-paper surveys, the rise of the Internet and the proliferation of corporate websites have changed the means by which research is executed.

Web analytics were born out of the need to track the behaviour of site visitors and, as the popularity of e-commerce and web advertising grew,businesses demanded details on the information created by new practices in web data collection, such as click-through and exit rates. As the Internet boomed,websites became larger and more complex and the possibility of two-way communication between businesses and their consumers became a reality. Provided with the capacity to interact with online customers, Researchers were able to collect large amounts of data that were previously unavailable, further propelling the Marketing Research Industry.

In the new millennium, as the Internet continued to develop and websites became more interactive, data collection and analysis became more commonplace for those Marketing Research Firms whose clients had a web presence. With the explosive growth of the online marketplace came new competition for companies; no longer were businesses merely competing with the shop down the road — competition was now represented by a global force. Retail outlets were appearing online and the previous need for bricks-and-mortar stores was diminishing at a greater pace than online competition was growing.With so many online channels for consumers to make purchases, companies needed newer and more compelling methods, in combination with messages that resonated more effectively, to capture the attention of the average consumer.

Having access to web data did not automatically provide companies with the rationale behind the behaviour of users visiting their sites,which provoked the marketing research industry to develop new and better ways of tracking, collecting and interpreting information. This led to the development of various tools like online focus groups and pop-up or website intercept surveys. These types of services allowed companies to dig deeper into the motivations of consumers, augmenting their insights and utilizing this data to drive market share.

As information around the world became more accessible, increased competition led companies to demand more of Market Researchers. It was no longer sufficient to follow trends in web behaviour or track sales data; companies now needed access to consumer behaviour throughout the entire purchase process.This meant the Marketing Research Industry, again, needed to adapt to the rapidly changing needs of the marketplace, and to the demands of companies looking fora competitive edge.

Today, Marketing Research has adapted to innovations in technology and the corresponding ease with which information is available. B2B and B2C companies are working hard to stay competitive and they now demand both quantitative ("What") and qualitative ("Why?") marketing research in order to better understand their target audience and the motivations behind customer behaviours.

This demand is driving Marketing Researchers to develop new platforms for interactive, two-way communication between their firms and consumers. Mobile devices such as SmartPhones are the best example of an emerging platform that enables businesses to connect with their customers throughout the entire buying process. Innovative research firms, such as *OnResearch* with their *OnMobile* app, are now providing businesses with the means to reach consumers from the point of initial investigation through to the decision and, ultimately, the purchase.

As personal mobile devices become more capable and widespread, the Marketing Research Industry will look to further capitalize on this trend. Mobile devices present the perfect channel for Research Firms to retrieve immediate impressions from buyers and to provide their clients with a holistic view of the consumers within their target markets, and beyond. Now, more than ever,innovation is the key to success for Marketing Researchers. Marketing Research Clients are beginning to demand highly personalized and specifically-focused products from the MR firms; big data is great for identifying general market segments, but is less capable of identifying key factors of niche markets, which now defines the competitive edge companies are looking for in this mobile-digital age.

Characteristics

First, marketing *research is systematic*. Thus systematic planning is required at all the stages of the marketing research process. The procedures followed at each stage are methodologically sound, well documented, and, as much as possible, planned in advance. Marketing research uses the scientific method in that data are collected and analyzed to test prior notions or hypotheses. Experts in marketing research have shown that studies featuring multiple and often competing hypotheses yield more meaningful results than those featuring only one dominant hypothesis.

Marketing research is *objective*. It attempts to provide accurate information that reflects a true state of affairs. It should be conducted impartially. While research is always influenced by the researcher's research philosophy, it should be free from the personal or political biases of the researcher or the management. Research which is motivated by personal or political gain involves a breach of professional standards. Such research is deliberately biased so as to result in predetermined findings. The objective nature of marketing research underscores the importance of ethical considerations. Also, researchers should always be objective with regard to the selection of information to be featured in reference texts because such literature should offer a comprehensive view on marketing. Research has shown, however, that many marketing textbooks do not feature important principles in marketing research.

Related Business Research

Other forms of business research include:

- Market research is broader in scope and examines all aspects of a business environment. It asks questions about competitors, market structure, government regulations, economic trends, technological advances, and numerous other factors that make up the business environment. Sometimes the term refers more particularly to the financial analysis of companies, industries, or sectors. In this case, financial analysts usually carry out the research and provide the results to investment advisors and potential investors.

- Product research — This looks at what products can be produced with available technology, and what new product innovations near-future technology can develop.
- Advertising research - is a specialized form of marketing research conducted to improve the efficacy of advertising. Copy testing, also known as "pre-testing," is a form of customized research that predicts in-market performance of an ad before it airs, by analyzing audience levels of attention, brand linkage, motivation, entertainment, and communication, as well as breaking down the ad's flow of attention and flow of emotion. Pre-testing is also used on ads still in rough (ripomatic or animatic) form. (Young, p. 213)

Classification

Organizations engage in marketing research for two reasons: (1) to identify and (2) solve marketing problems. This distinction serves as a basis for classifying marketing research into problem identification research and problem solving research.

Problem identification research is undertaken to help identify problems which are, perhaps, not apparent on the surface and yet exist or are likely to arise in the future like company image, market characteristics, sales analysis, short-range forecasting, long range forecasting, and business trends research. Research of this type provides information about the marketing environment and helps diagnose a problem. For example, The findings of problem solving research are used in making decisions which will solve specific marketing problems.

The Stanford Research Institute, on the other hand, conducts an annual survey of consumers that is used to classify persons into homogeneous groups for segmentation purposes. The National Purchase Diary panel (NPD) maintains the largest diary panel in the United States.

Standardized services are research studies conducted for different client firms but in a standard way. For example, procedures for measuring advertising effectiveness have been standardized so that the results can be compared across studies and evaluative norms can be established. The Starch Readership Survey is the most widely used service for evaluating print advertisements; another well-known service is the Gallup and Robinson Magazine Impact Studies. These services are also sold on a syndicated basis.

- Customized services offer a wide variety of marketing research services customized to suit a client's specific needs. Each marketing research project is treated uniquely.
- Limited-service suppliers specialize in one or a few phases of the marketing research project. Services offered by such suppliers are classified as field services, coding and data entry, data analysis, analytical services, and branded products. Field services collect data through the internet, traditional mail, in-person, or telephone interviewing, and firms that special-

ize in interviewing are called field service organizations. These organizations may range from small proprietary organizations which operate locally to large multinational organizations with WATS line interviewing facilities. Some organizations maintain extensive interviewing facilities across the country for interviewing shoppers in malls.

- Coding and data entry services include editing completed questionnaires, developing a coding scheme, and transcribing the data on to diskettes or magnetic tapes for input into the computer. NRC Data Systems provides such services.
- Analytical services include designing and pretesting questionnaires, determining the best means of collecting data, designing sampling plans, and other aspects of the research design. Some complex marketing research projects require knowledge of sophisticated procedures, including specialized experimental designs, and analytical techniques such as conjoint analysis and multidimensional scaling. This kind of expertise can be obtained from firms and consultants specializing in analytical services.
- Data analysis services are offered by firms, also known as tab houses, that specialize in computer analysis of quantitative data such as those obtained in large surveys. Initially most data analysis firms supplied only tabulations (frequency counts) and cross tabulations (frequency counts that describe two or more variables simultaneously). With the proliferation of software, many firms now have the capability to analyze their own data, but, data analysis firms are still in demand.
- Branded marketing research products and services are specialized data collection and analysis procedures developed to address specific types of marketing research problems. These procedures are patented, given brand names, and marketed like any other branded product.

Types

Marketing research techniques come in many forms, including:

- Ad Tracking – periodic or continuous in-market research to monitor a brand's performance using measures such as brand awareness, brand preference, and product usage. (Young, 2005)
- Advertising Research – used to predict copy testing or track the efficacy of advertisements for any medium, measured by the ad's ability to get attention (measured with Attention-Tracking), communicate the message, build the brand's image, and motivate the consumer to purchase the product or service. (Young, 2005)
- Brand awareness research — the extent to which consumers can recall or recognise a brand name or product name
- Brand association research — what do consumers associate with the brand?
- Brand attribute research — what are the key traits that describe the brand promise?
- Brand name testing - what do consumers feel about the names of the products?
- Buyer decision making process— to determine what motivates people to buy and what

decision-making process they use; over the last decade, Neuromarketing emerged from the convergence of neuroscience and marketing, aiming to understand consumer decision making process

- Commercial eye tracking research — examine advertisements, package designs, websites, etc. by analyzing visual behavior of the consumer
- Concept testing - to test the acceptance of a concept by target consumers
- Coolhunting (also known as trendspotting) - to make observations and predictions in changes of new or existing cultural trends in areas such as fashion, music, films, television, youth culture and lifestyle
- Copy testing – predicts in-market performance of an ad before it airs by analyzing audience levels of attention, brand linkage, motivation, entertainment, and communication, as well as breaking down the ad's flow of attention and flow of emotion. (Young, p 213)
- Customer satisfaction research - quantitative or qualitative studies that yields an understanding of a customer's satisfaction with a transaction
- Demand estimation — to determine the approximate level of demand for the product
- Distribution channel audits — to assess distributors' and retailers' attitudes toward a product, brand, or company
- Internet strategic intelligence — searching for customer opinions in the Internet: chats, forums, web pages, blogs... where people express freely about their experiences with products, becoming strong opinion formers.
- Marketing effectiveness and analytics — Building models and measuring results to determine the effectiveness of individual marketing activities.
- Mystery consumer or mystery shopping - An employee or representative of the market research firm anonymously contacts a salesperson and indicates he or she is shopping for a product. The shopper then records the entire experience. This method is often used for quality control or for researching competitors' products.
- Positioning research — how does the target market see the brand relative to competitors? - what does the brand stand for?
- Price elasticity testing — to determine how sensitive customers are to price changes
- Sales forecasting — to determine the expected level of sales given the level of demand. With respect to other factors like Advertising expenditure, sales promotion etc.
- Segmentation research - to determine the demographic, psychographic, cultural, and behavioural characteristics of potential buyers
- Online panel - a group of individual who accepted to respond to marketing research online
- Store audit — to measure the sales of a product or product line at a statistically selected store sample in order to determine market share, or to determine whether a retail store provides adequate service

- Test marketing — a small-scale product launch used to determine the likely acceptance of the product when it is introduced into a wider market
- Viral Marketing Research - refers to marketing research designed to estimate the probability that specific communications will be transmitted throughout an individual's Social Network. Estimates of Social Networking Potential (SNP) are combined with estimates of selling effectiveness to estimate ROI on specific combinations of messages and media.

All of these forms of marketing research can be classified as either problem-identification research or as problem-solving research.

There are two main sources of data — primary and secondary. Primary research is conducted from scratch. It is original and collected to solve the problem in hand. Secondary research already exists since it has been collected for other purposes. It is conducted on data published previously and usually by someone else. Secondary research costs far less than primary research, but seldom comes in a form that exactly meets the needs of the researcher.

A similar distinction exists between exploratory research and conclusive research. Exploratory research provides insights into and comprehension of an issue or situation. It should draw definitive conclusions only with extreme caution. Conclusive research draws conclusions: the results of the study can be generalized to the whole population.

Exploratory research is conducted to explore a problem to get some basic idea about the solution at the preliminary stages of research. It may serve as the input to conclusive research. Exploratory research information is collected by focus group interviews, reviewing literature or books, discussing with experts, etc. This is unstructured and qualitative in nature. If a secondary source of data is unable to serve the purpose, a convenience sample of small size can be collected. Conclusive research is conducted to draw some conclusion about the problem. It is essentially, structured and quantitative research, and the output of this research is the input to management information systems (MIS).

Exploratory research is also conducted to simplify the findings of the conclusive or descriptive research, if the findings are very hard to interpret for the marketing managers.

Methods

Methodologically, marketing research uses the following types of research designs:

Based on questioning

- Qualitative marketing research - generally used for exploratory purposes — small number of respondents — not generalizable to the whole population — statistical significance and confidence not calculated — examples include focus groups, in-depth interviews, and projective techniques
- Quantitative marketing research - generally used to draw conclusions — tests a specific hypothesis - uses random sampling techniques so as to infer from the sample to the population — involves a large number of respondents — examples include surveys and questionnaires. Techniques include choice modelling, maximum difference preference scaling, and covariance analysis.

Based on observations

- Ethnographic studies — by nature qualitative, the researcher observes social phenomena in their natural setting — observations can occur cross-sectionally (observations made at one time) or longitudinally (observations occur over several time-periods) - examples include product-use analysis and computer cookie traces.
- Experimental techniques - by nature quantitative, the researcher creates a quasi-artificial environment to try to control spurious factors, then manipulates at least one of the variables — examples include purchase laboratories and test markets

Researchers often use more than one research design. They may start with secondary research to get background information, then conduct a focus group (qualitative research design) to explore the issues. Finally they might do a full nationwide survey (quantitative research design) in order to devise specific recommendations for the client.

Business to Business

Business to business (B2B) research is inevitably more complicated than consumer research. The researchers need to know what type of multi-faceted approach will answer the objectives, since seldom is it possible to find the answers using just one method. Finding the right respondents is crucial in B2B research since they are often busy, and may not want to participate. Encouraging them to "open up" is yet another skill required of the B2B researcher. Last, but not least, most business research leads to strategic decisions and this means that the business researcher must have expertise in developing strategies that are strongly rooted in the research findings and acceptable to the client.

There are four key factors that make B2B market research special and different from consumer markets:

- The decision making unit is far more complex in B2B markets than in consumer markets
- B2B products and their applications are more complex than consumer products
- B2B marketers address a much smaller number of customers who are very much larger in their consumption of products than is the case in consumer markets
- Personal relationships are of critical importance in B2B markets.

Small Businesses and Nonprofits

Marketing research does not only occur in huge corporations with many employees and a large budget. Marketing information can be derived by observing the environment of their location and the competitions location. Small scale surveys and focus groups are low cost ways to gather information from potential and existing customers. Most secondary data (statistics, demographics, etc.) is available to the public in libraries or on the internet and can be easily accessed by a small business owner.

Below are some steps that could be done by SME (Small Medium Entreprise) to analyze the market:

1. Provide secondary and or primary data (if necessary);
2. Analyze Macro & Micro Economic data (e.g. Supply & Demand, GDP,Price change, Economic growth, Sales by sector/industries,interest rate, number of investment/ divestment, I/O, CPI, Social anlysis,etc.);
3. Implement the marketing mix concept, which is consist of: Place, Price, Product,Promotion, People, Process, Physical Evidence and also Political & social situation to analyze global market situation);
4. Analyze market trends, growth, market size, market share, market competition (e.g. SWOT analysis, B/C Analysis,channel mapping identities of key channels, drivers of customers loyalty and satisfaction, brand perception, satisfaction levels, current competitor-channel relationship analysis, etc.),etc.;
5. Determine market segment, market target, market forecast and market position;
6. Formulating market strategy & also investigating the possibility of partnership/ collaboration (e.g. Profiling & SWOT analysis of potential partners, evaluating business partnership.)
7. Combine those analysis with the SME's business plan/ business model analysis (e.g. Business description, Business process, Business strategy, Revenue model, Business expansion, Return of Investment, Financial analysis (Company History, Financial assumption, Cost/ Benefit Analysis, Projected profit & Loss, Cashflow, Balance sheet & business Ratio,etc.).

Note as important : Overall analysis should be based on 6W+1H (What, When, Where, Which, Who, Why and How) question.

International Plan

International Marketing Research follows the same path as domestic research, but there are a few more problems that may arise. Customers in international markets may have very different customs, cultures, and expectations from the same company. In this case, Marketing Research relies more on primary data rather than secondary information. Gathering the primary data can be hindered by language, literacy and access to technology. Basic Cultural and Market intelligence information will be needed to maximize the research effectiveness. Some of the steps that would help overcoming barriers include; 1. Collect secondary information on the country under study from reliable international source e.g. WHO and IMF 2. Collect secondary information on the product/service under study from available sources 3. Collect secondary information on product manufacturers and service providers under study in relevant country 4. Collect secondary information on culture and common business practices 5. Ask questions to get better understanding of reasons behind any recommendations for a specific methodology

Common Terms

Market research techniques resemble those used in political polling and social science research. Meta-analysis (also called the Schmidt-Hunter technique) refers to a statistical method of combining data from multiple studies or from several types of studies. Conceptualization means the process of converting vague mental images into definable concepts. Operationalization is the process of converting concepts into specific observable behaviors that a researcher can measure. Precision refers to the exactness of any given measure. Reliability refers to the likelihood that a given operationalized construct will yield the same results if re-measured. Validity refers to the extent to which a measure provides data that captures the meaning of the operationalized construct as defined in the study. It asks, "Are we measuring what we intended to measure?"

- Applied research sets out to prove a specific hypothesis of value to the clients paying for the research. For example, a cigarette company might commission research that attempts to show that cigarettes are good for one's health. Many researchers have ethical misgivings about doing applied research.
- Sugging (from SUG, for "selling under the guise" of market research) forms a sales technique in which sales people pretend to conduct marketing research, but with the real purpose of obtaining buyer motivation and buyer decision-making information to be used in a subsequent sales call.
- Frugging comprises the practice of soliciting funds under the pretense of being a research organization.

Careers

Some of the positions available in marketing research include vice president of marketing research, research director, assistant director of research, project manager, field work director, statistician/data processing specialist, senior analyst, analyst, junior analyst and operational supervisor.

The most common entry-level position in marketing research for people with bachelor's degrees (e.g., BBA) is as operational supervisor. These people are responsible for supervising a well-defined set of operations, including field work, data editing, and coding, and may be involved in programming and data analysis. Another entry-level position for BBAs is assistant project manager. An assistant project manager will learn and assist in questionnaire design, review field instructions, and monitor timing and costs of studies. In the marketing research industry, however, there is a growing preference for people with master's degrees. Those with MBA or equivalent degrees are likely to be employed as project managers.

A small number of business schools also offer a more specialized Master of Marketing Research (MMR) degree. An MMR typically prepares students for a wide range of research methodologies and focuses on learning both in the classroom and the field.

The typical entry-level position in a business firm would be junior research analyst (for BBAs) or research analyst (for MBAs or MMRs). The junior analyst and the research analyst learn about the particular industry and receive training from a senior staff member, usually the marketing research manager. The junior analyst position includes a training program to prepare individuals for the responsibilities of a research analyst, including coordinating with the marketing department and sales force to develop goals for product exposure. The research analyst responsibilities include checking all data for accuracy, comparing and contrasting new research with established norms, and analyzing primary and secondary data for the purpose of market forecasting.

As these job titles indicate, people with a variety of backgrounds and skills are needed in marketing research. Technical specialists such as statisticians obviously need strong backgrounds in statistics and data analysis. Other positions, such as research director, call for managing the work of others and require more general skills. To prepare for a career in marketing research, students usually:

- Take all the marketing courses.
- Take courses in statistics and quantitative methods.
- Acquire computer skills.
- Take courses in psychology and consumer behavior.
- Acquire effective written and verbal communication skills.
- Think creatively.

Corporate Hierarchy

- Vice-President of Marketing Research: This is the senior position in marketing research. The VP is responsible for the entire marketing research operation of the company and serves on the top management team. Sets the objectives and goals of the marketing research department.
- Research Director: Also a senior position, the director has the overall responsibility for the development and execution of all the marketing research projects.
- Assistant Director of Research: Serves as an administrative assistant to the director and supervises some of the other marketing research staff members.
- (Senior) Project Manager: Has overall responsibility for design, implementation, and management of research projects.
- Statistician/Data Processing Specialist: Serves as an expert on theory and application of statistical techniques. Responsibilities include experimental design, data processing, and analysis.
- Senior Analyst: Participates in the development of projects and directs the operational execution of the assigned projects. Works closely with the analyst, junior analyst, and other personnel in developing the research design and data collection. Prepares the final report. The primary responsibility for meeting time and cost constraints rests with the senior analyst.
- Analyst: Handles the details involved in executing the project. Designs and pretests the questionnaires and conducts a preliminary analysis of the data.
- Junior Analyst: Handles routine assignments such as secondary data analysis, editing and coding of questionnaires, and simple statistical analysis.
- Field Work Director: Responsible for the selection, training, supervision, and evaluation of interviewers and other field workers.

References

- McQuarrie, Edward (2005), The market research toolbox: a concise guide for beginners (2nd ed.), SAGE, ISBN 978-1-4129-1319-5
- Drake, Philip (2008). McDonald & Wasko, ed. Distribution and Marketing in Contemporary Hollywood. Malden, MA: Blackwell Publishing. pp. 63–82. ISBN 978-1-4051-3388-3.
- 'What is geographic segmentation' Kotler, Philip, and Kevin Lane Keller. Marketing Management. Prentice Hall, 2006. ISBN 978-0-13-145757-7
- Reid, Robert D.; Bojanic, David C. (2009). Hospitality Marketing Management (5 ed.). John Wiley and Sons. p. 139. ISBN 978-0-470-08858-6. Retrieved 2013-06-08.
- Armstrong. M. A handbook of Human Resource Management Practice (10th edition) 2006, Kogan Page , London ISBN 0-7494-4631-5
- Bradley, Nigel Marketing Research. Tools and Techniques.Oxford University Press, Oxford, 2007 ISBN 0-19-928196-3 ISBN 978-0-19-928196-1
- Marder, Eric The Laws of Choice—Predicting Customer Behavior (The Free Press division of Simon & Schuster, 1997. ISBN 0-684-83545-2
- Kotler, Philip and Armstrong, Gary Principles of Marketing Pearson, Prentice Hall, New Jersey, 2007 ISBN 978-0-13-239002-6, ISBN 0-13-239002-7
- Berghoff, Hartmut, Philip Scranton, and Uwe Spiekermann, eds., The Rise of Marketing and Market Research (New York: Palgrave Macmillan, 2012), ISBN 978-0-230-34106-7
- Hunt, Shelby; Arnett, Dennis (16 June 2004). "Market Segmentation Strategy, Competitive Advantage, and Public Policy". 12 (1). Australasian Marketing Journal: 1–25. Retrieved 18 March 2016.
- Ommani, Ahmad (30 September 2011). "SWOT analysis for business management". 5 (22). African Journal of Business Management: 9448–9454. Retrieved 17 March 2016.

6

Essential Aspects of Business Intelligence

The essential aspects of business intelligence are context analysis, business performance management, business process discovery, information system, organization intelligence and process mining. The method to analyze the environment of any business is known as context analysis. The topics discussed in this section are of great importance to broaden the existing knowledge on business intelligence.

Context Analysis

Context analysis is a method to analyze the environment in which a business operates. Environmental scanning mainly focuses on the macro environment of a business. But context analysis considers the entire environment of a business, its internal and external environment. This is an important aspect of business planning. One kind of context analysis, called SWOT analysis, allows the business to gain an insight into their strengths and weaknesses and also the opportunities and threats posed by the market within which they operate. The main goal of a context analysis, SWOT or otherwise, is to analyze the environment in order to develop a strategic plan of action for the business.

Context analysis also refers to a method of sociological analysis associated with Scheflen (1963) which believes that 'a given act, be it a glance at [another] person, a shift in posture, or a remark about the weather, has no intrinsic meaning. Such acts can only be understood when taken in relation to one another.' (Kendon, 1990: 16). This is not discussed here; only Context Analysis in the business sense is.

Define Market or Subject

The first step of the method is to define a particular market (or subject) one wishes to analyze and focus all analysis techniques on what was defined. A subject, for example, can be a newly proposed product idea.

Trend Analysis

The next step of the method is to conduct a trend analysis. Trend analysis is an analysis of macro environmental factors in the external environment of a business, also called PEST analysis. It consists of analyzing political, economical, social, technological and demographic trends. This can be done by first determining which factors, on each level, are relevant for the chosen subject and to score each item as to specify its importance. This allows the business to identify those factors that can influence them. They can't control these factors but they can try to cope with them by adapting themselves. The trends (factors) that are addressed in PEST analysis are Political, Economical, Social and Technological; but for context analysis Demographic trends are also of importance. Demographic trends are those factors that have to do with the population, like for example average age, religion, education etc. Demographic information is of importance if, for example during market research, a business wants to determine a particular market segment to target. The other trends are described in environmental scanning and PEST analysis. Trend analysis only covers part of the external environment. Another important aspect of the external environment that a business should consider is its competition. This is the next step of the method, competitor analysis.

Competitor Analysis

As one can imagine, it is important for a business to know who its competition is, how they do their business and how powerful they are so that they can be on the defense and offense. In Competitor analysis a couple of techniques are introduced how to conduct such an analysis. Here I will introduce another technique which involves conducting four sub analyses, namely: determination of competition levels, competitive forces, competitor behavior and competitor strategy.

Competition Levels

Businesses compete on several levels and it is important for them to analyze these levels so that they can understand the demand. Competition is identified on four levels:

- Consumer needs: level of competition that refers to the needs and desires of consumers. A business should ask: What are the desires of the consumers?
- General competition: The kind of consumer demand. For example: do consumers prefer shaving with electric razor or a razor blade?
- Brand: This level refers to brand competition. Which brands are preferable to a consumer?
- Product: This level refers to the type of demand. Thus what types of products do consumers prefer?

Another important aspect of a competition analysis is to increase the consumer insight. For example: [Ducati] has, by interviewing a lot of their customers, concluded that their main competitor is not another bicycle, but sport-cars like [Porsche] or [GM]. This will of course influence the competition level within this business.

Competitive Forces

These are forces that determine the level of competition within a particular market. There are six forces that have to be taken into consideration, power of the competition, threat of new entrants, bargaining power of buyers and suppliers, threat of substitute products and the importance of complementary products. This analysis is described in Porter 5 forces analysis.

Competitor Behavior

Competitor behaviors are the defensive and offensive actions of the competition.

Competitor Strategy

These strategies refer to how an organization competes with other organizations. And these are: low price strategy and product differentiation strategy.

Opportunities and Threats

The next step, after the trend analysis and competitor analysis are conducted, is to determine threats and opportunities posed by the market. The trends analysis revealed a set of trends that can influence the business in either a positive or a negative manner. These can thus be classified as either opportunities or threats. Likewise, the competitor analysis revealed positive and negative competition issues that can be classified as opportunities or threats.

Organization Analysis

The last phase of the method is an analysis of the internal environment of the organization, thus the organization itself. The aim is to determine which skills, knowledge and technological fortes the business possesses. This entails conducting an internal analysis and a competence analysis.

Internal Analysis

The internal analysis, also called SWOT analysis, involves identifying the organizations strengths and weaknesses. The strengths refer to factors that can result in a market advantage and weaknesses to factors that give a disadvantage because the business is unable to comply with the market needs.

Competence Analysis

Competences are the combination of a business' knowledge, skills and technology that can give them the edge versus the competition. Conducting such an analysis involves identifying market related competences, integrity related competences and functional related competences.

SWOT-i Matrix

The previous sections described the major steps involved in context analysis. All these steps resulted in data that can be used for developing a strategy. These are summarized in a SWOT-i matrix. The trend and competitor analysis revealed the opportunities and threats posed by the market. The organization analysis revealed the competences of the organization and also its strengths and weaknesses. These strengths, weaknesses, opportunities and threats summarize the entire context analysis. A SWOT-i matrix, depicted in the table below, is used to depict these and to help visualize the strategies that are to be devised. SWOT- i stand for Strengths, Weaknesses, Opportunities, Threats and Issues. The Issues refer to strategic issues that will be used to devise a strategic plan.

	Opportunities (O_1, O_2, ..., O_n)	**Threats (T_1, T_2, ..., T_n)**
Strengths (S_1, S_2, ..., S_n)	$S_1O_1 \ldots S_nO_1$ … $S_1O_n \ldots S_nO_n$	$S_1T_1 \ldots S_nT_1$ … $S_1T_n \ldots S_nT_n$
Weaknesses (W_1, W_2, ..., W_n)	$W_1O_1 \ldots W_nO_1$ … $W_1O_n \ldots W_nO_n$	$W_1T_1 \ldots W_nT_1$ … $W_1T_n \ldots W_nT_n$

This matrix combines the strengths with the opportunities and threats, and the weaknesses with the opportunities and threats that were identified during the analysis. Thus the matrix reveals four clusters:

- Cluster strengths and opportunities: use strengths to take advantage of opportunities.
- Cluster strengths and threats: use strengths to overcome the threats
- Cluster weaknesses and opportunities: certain weaknesses hamper the organization from taking advantage of opportunities therefore they have to look for a way to turn those weaknesses around.
- Cluster weaknesses and threats: there is no way that the organization can overcome the threats without having to make major changes.

Strategic Plan

The ultimate goal of context analysis is to develop a strategic plan. The previous sections described all the steps that form the stepping stones to developing a strategic plan of action for the organization .The trend and competitor analysis gives insight to the opportunities and threats in the market and the internal analysis gives insight to the competences of the organization. And these were combined in the SWOT-i matrix. The SWOT-i matrix helps identify issues that need to be dealt with. These issues need to be resolved by formulating an objective and a plan to reach that objective, a strategy.

Example

Joe Arden is in the process of writing a business plan for his business idea, Arden Systems. Arden Systems will be a software business that focuses on the development of software for small businesses. Joe realizes that this is a tough market because there are many software companies that develop business software. Therefore, he conducts context analysis to gain insight into the environment of the business in order to develop a strategic plan of action to achieve competitive advantage within the market.

Define Market

First step is to define a market for analysis. Joe decides that he wants to focus on small businesses consisting of at most 20 employees.

Trend Analysis

Next step is to conduct trend analysis. The macro environmental factors that Joe should take into consideration are as follows:

- Political trend: Intellectual property rights
- Economical trend: Economic growth
- Social trend: Reduce operational costs; Ease for conducting business administration

- Technological trend: Software suites; Web applications
- Demographic trend: Increase in the graduates of IT related studies

Competitor Analysis

Following trend analysis is competitor analysis. Joe analyzes the competition on four levels to gain insight into how they operate and where advantages lie.

- Competition level:
    - Consumer need: Arden Systems will be competing on the fact that consumers want efficient and effective conducting of a business
    - Brand: There are software businesses that have been making business software for a while and thus have become very popular in the market. Competing based on brand will be difficult.
    - Product: They will be packaged software like the major competition.
- Competitive forces: Forces that can affect Arden Systems are in particular:
    - The bargaining power of buyers: the extent to which they can switch from one product to the other.
    - Threat of new entrants: it is very easy for someone to develop a new software product that can be better than Arden's.
    - Power of competition: the market leaders have most of the cash and customers; they have to power to mold the market.
- Competitor behavior: The focus of the competition is to take over the position of the market leader.
- Competitor strategy: Joe intends to compete based on product differentiation.

Opportunities and Threats

Now that Joe has analyzed the competition and the trends in the market he can define opportunities and threats.

- Opportunities:
    - Because the competitors focus on taking over the leadership position, Arden can focus on those segments of the market that the market leader ignores. This allows them to take over where the market leader shows weakness.
    - The fact that there are new IT graduates, Arden can employ or partner with someone that may have a brilliant idea.
- Threats:
    - IT graduates with fresh idea's can start their own software businesses and form a major competition for Arden Systems.

Organization Analysis

After Joe has identified the opportunities and threats of the market he can try to figure out what Arden System's strengths and weaknesses are by doing an organization analysis.

- Internal analysis:
    - Strength: Product differentiation
    - Weakness: Lacks innovative people within the organization
- Competence analysis:
    - Functional related competence: Arden Systems provides system functionalities that fit small businesses.
    - Market-related competence: Arden Systems has the opportunity to focus on a part of the market which is ignored.

SWOT-i Matrix

After the previous analyses, Joe can create a SWOT-i matrix to perform SWOT analysis.

	Opportunities	Threats
Strengths	Product differentiation, market leader ignores market segment	
Weaknesses		Lack of innovation, increase in IT graduates

Strategic Plan

After creating the SWOT-i matrix, Joe is now able to devise a strategic plan.

- Focus all software development efforts to that part of the market which is ignored by market leaders, small businesses.
- Employ recent innovative It graduates to stimulate the innovation within Arden Systems.

Business Performance Management

Business performance management is a set of performance management and analytic processes that enables the management of an organization's performance to achieve one or more pre-selected goals. Synonyms for "business performance management" include "corporate performance management (CPM)" and "enterprise performance management".

Business performance management is contained within approaches to business process management.

Business performance management has three main activities:

1. selection of goals,
2. consolidation of measurement information relevant to an organization's progress against these goals, and
3. interventions made by managers in light of this information with a view to improving future performance against these goals.

Although presented here sequentially, typically all three activities will run concurrently, with interventions by managers affecting the choice of goals, the measurement information monitored, and the activities being undertaken by the organization.

Because business performance management activities in large organizations often involve the collation and reporting of large volumes of data, many software vendors, particularly those offering business intelligence tools, market products intended to assist in this process. As a result of this marketing effort, business performance management is often incorrectly understood as an activity that necessarily relies on software systems to work, and many definitions of business performance management explicitly suggest software as being a definitive component of the approach.

This interest in business performance management from the software community is sales-driven - "The biggest growth area in operational BI analysis is in the area of business performance management."

Since 1992, business performance management has been strongly influenced by the rise of the balanced scorecard framework. It is common for managers to use the balanced scorecard framework to clarify the goals of an organization, to identify how to track them, and to structure the mechanisms by which interventions will be triggered. These steps are the same as those that are found in BPM, and as a result balanced scorecard is often used as the basis for business performance management activity with organizations.

In the past, owners have sought to drive strategy down and across their organizations, transform these strategies into actionable metrics and use analytics to expose the cause-and-effect relationships that, if understood, could give insight into decision-making.

History

Reference to non-business performance management occurs in Sun Tzu's *The Art of War*. Sun Tzu claims that to succeed in war, one should have full knowledge of one's own strengths and weaknesses as well as those of one's enemies. Lack of either set of knowledge might result in defeat. Parallels between the challenges in business and those of war include:

- collecting data - both internal and external
- discerning patterns and meaning in the data (analyzing)
- responding to the resultant information

Prior to the start of the Information Age in the late 20th century, businesses sometimes took the trouble to laboriously collect data from non-automated sources. As they lacked computing resources to properly analyze the data, they often made commercial decisions primarily on the basis of intuition.

As businesses started automating more and more systems, more and more data became available. However, collection often remained a challenge due to a lack of infrastructure for data exchange or due to incompatibilities between systems. Reports on the data gathered sometimes took months to generate. Such reports allowed informed long-term strategic decision-making. However, short-term tactical decision-making often continued to rely on intuition.

In 1989 Howard Dresner, a research analyst at Gartner, popularized “business intelligence” (BI) as an umbrella term to describe a set of concepts and methods to improve business decision-making by using fact-based support systems. Performance management builds on a foundation of BI, but marries it to the planning-and-control cycle of the enterprise - with enterprise planning, consolidation and modeling capabilities.

Increasing standards, automation, and technologies have led to vast amounts of data becoming available. Data warehouse technologies have allowed the building of repositories to store this data. Improved ETL and enterprise application integration tools have increased the timely collecting of data. OLAP reporting technologies have allowed faster generation of new reports which analyze the data. As of 2010, business intelligence has become the art of sieving through large amounts of data, extracting useful information and turning that information into actionable knowledge.

Definition and Scope

Business performance management consists of a set of management and analytic processes, supported by technology, that enable businesses to define strategic goals and then measure and manage performance against those goals. Core business performance management processes include financial planning, operational planning, business modeling, consolidation and reporting, analysis, and monitoring of key performance indicators linked to strategy.

Business performance management involves consolidation of data from various sources, querying, and analysis of the data, and putting the results into practice.

Frameworks

Various frameworks for implementing business performance management exist. The discipline gives companies a top-down framework by which to align planning and execution, strategy and tactics, and business-unit and enterprise objectives. Reactions may include the Six Sigma strategy, balanced scorecard, activity-based costing (ABC), Objectives and Key Results (OKR), Total Quality Management, economic value-add, integrated strategic measurement and Theory of Constraints.

The balanced scorecard is the most widely adopted performance management framework.

Metrics and Key Performance Indicators

Some of the areas from which bank management may gain knowledge by using business performance management include:

- customer-related numbers:

  - new customers acquired
  - status of existing customers
  - attrition of customers (including breakup by reason for attrition)
- turnover generated by segments of the customers - possibly using demographic filters
- outstanding balances held by segments of customers and terms of payment - possibly using demographic filters
- collection of bad debts within customer relationships
- demographic analysis of individuals (potential customers) applying to become customers, and the levels of approval, rejections and pending numbers
- delinquency analysis of customers behind on payments
- profitability of customers by demographic segments and segmentation of customers by profitability
- campaign management
- real-time dashboard on key operational metrics
  - overall equipment effectiveness
- clickstream analysis on a website
- key product portfolio trackers
- marketing-channel analysis
- sales-data analysis by product segments
- callcenter metrics

Though the above list describes what a bank might monitor, it could refer to a telephone company or to a similar service-sector company.

Items of generic importance include:

1. consistent and correct KPI-related data providing insights into operational aspects of a company
2. timely availability of KPI-related data
3. KPIs designed to directly reflect the efficiency and effectiveness of a business
4. information presented in a format which aids decision-making for management and decision-makers
5. ability to discern patterns or trends from organized information

Business performance management integrates the company's processes with CRM or ERP. Companies should become better able to gauge customer satisfaction, control customer trends and influence shareholder value.

Application Software Types

People working in business intelligence have developed tools that ease the work of business performance management, especially when the business-intelligence task involves gathering and analyzing large amounts of unstructured data.

Tool categories commonly used for business performance management include:

- MOLAP — Multidimensional online analytical processing, sometimes simply called "analytics" (based on dimensional analysis and the so-called "hypercube" or "cube")
- scorecarding, dashboarding and data visualization
- data warehouses
- document warehouses
- text mining
- DM — data mining
- BPO — business performance optimization
- EPM — enterprise performance management
- EIS — executive information systems
- DSS — decision support systems
- MIS — management information systems
- SEMS — strategic enterprise management software
- EOI — Operational intelligence Enterprise Operational Intelligence Software

Design and Implementation

Questions asked when implementing a business performance management program include:

- Goal-alignment queries

  Determine the short- and medium-term purpose of the program. What strategic goal(s) of the organization will the program address? What organizational mission/vision does it relate to? A hypothesis needs to be crafted that details how this initiative will eventually improve results / performance (i.e. a strategy map).

- Baseline queries

  Assess current information-gathering competency. Does the organization have the capability to monitor important sources of information? What data is being collected and how is it being stored? What are the statistical parameters of this data, e.g., how much random variation does it contain? Is this being measured?

- Cost and risk queries

Estimate the financial consequences of a new BI initiative. Assess the cost of the present operations and the increase in costs associated with the BPM initiative. What is the risk that the initiative will fail? This risk assessment should be converted into a financial metric and included in the planning.

- Customer and stakeholder queries

  Determine who will benefit from the initiative and who will pay. Who has a stake in the current procedure? What kinds of customers / stakeholders will benefit directly from this initiative? Who will benefit indirectly? What quantitative / qualitative benefits follow? Is the specified initiative the best or only way to increase satisfaction for all kinds of customers? How will customer benefits be monitored? What about employees, shareholders, and distribution channel members?

- Metrics-related queries

  Information requirements need operationalization into clearly defined metrics. Decide which metrics to use for each piece of information being gathered. Are these the best metrics and why? How many metrics need to be tracked? If this is a large number (it usually is), what kind of system can track them? Are the metrics standardized, so they can be benchmarked against performance in other organizations? What are the industry standard metrics available?

- Measurement methodology-related queries

  Establish a methodology or a procedure to determine the best (or acceptable) way of measuring the required metrics. How frequently will data be collected? Are there any industry standards for this? Is this the best way to do the measurements? How do we know that?

- Results-related queries

  Monitor the BPM program to ensure that it meets objectives. The program itself may require adjusting. The program should be tested for accuracy, reliability, and validity. How can it be demonstrated that the BI initiative, and not something else, contributed to a change in results? How much of the change was probably random?

Business Process Discovery

Business process discovery (BPD) related to process mining is a set of techniques that automatically construct a representation of an organization's current business processes and its major process variations. These techniques use evidence found in the existing technology systems that run business processes within an organization.

Business Process Discovery Techniques

Business process discovery techniques embody the following properties:

- Emergent paradigm - Current methods are based on top-down structured manual interviews relying on second-hand representations of the business process/system behaviors. An automated discovery process relies on collecting data from the information system over a period of time. This data can then be analyzed to form a process model.
- Automated process discovery – By automating the analysis of the data, the subjectivity of current manual process analysis techniques is removed. The automated system has an ingrained methodology that — through repeated trials — has been shown to accurately discover processes and process variations without bias.
- Accurate information- Since the information is collected from the actual source it cannot be inaccurate, as opposed to gathering it from second party representation.
- Complete information - An automated process captures all the information that is occurring within the system and represents them by time, date, user, etc.... Since the information is collected from real-time interactions, it is not subject to lost or selective memory issues. This includes completeness regarding exceptions in the processes. Often, exceptions are treated as statistical “noise,” which may exclude important inefficiencies in business processes.
- Standardized Process - The automated collection of information yields process data which can be grouped, quantified and classified. This supplies a basis for the development and monitoring of both current and new processes, to which benchmarks can be assigned. These benchmarks are the root of both new process design and the determination of problem root cause. Additionally, standardized process data can set the stage for efforts at continuous process improvement.

Application / Techniques

Business Process Discovery complements and builds upon the work in many other fields.

- Process discovery is one of the three main types of process mining. The other two types of process mining are conformance checking and model extension/enhancement. All of these techniques aim at extracting process related knowledge from event logs. In the case of process discovery, there is no prior process model; the model is discovered based on event logs. Conformance checking aims at finding differences between a given process model and event log. This way it is possible to quantify compliance and analyze discrepancies. Enhancement takes an a priori model and improves or extends it using information from the event log, e.g., show bottlenecks.
- Business process discovery is the next level of understanding in the emerging field of business analytics, which allows organizations to view, analyze and adjust the underlying structure and processes that go into day-to-day operations. This discovery includes information gathering of all of the components of a business process, including technology, people, department procedures and protocols.
- Business process discovery creates a process master which complements business process analysis (BPA). BPA tools and methodologies are well suited to top-down hierarchical process decomposition, and analysis of to-be processes. BPD provides a bottoms-up analysis

that marries to the top-down to provide a complete business process, organized hierarchically by BPA.

- Business Intelligence provides organizations with reporting and analytics on the data in their organizations. However, BI has no process model, awareness or analytics. BPD complements BI by providing an explicit process view to current operations, and providing analytics on that process model to help organizations identify and act upon business process inefficiencies, or anomalies.
- Web analytics are a limited example of BPD in that web analytics reconstruct the web-user's process as they interact with a Web-site. However, these analytics are limited to the process as is contained within the session, from the users perspective and with respect to just the web-based system and process.
- Business triage provides a framework for categorizing the processes identified by business process analysis (BPA) based on their relative importance to achieving a stated, measurable goal or outcome. Utilizing the same categories employed by military medical and disaster medical services, business processes are categorized as:
    - Essential/critical (red process) - Process essential for achieving outcomes/goals
    - Important/urgent (yellow process) - Process which speeds achieving outcomes/goals
    - Optional/supportive (green process) - Process not needed to achieve outcomes/goals

Resources are allocated based on the process category with resources first dedicated to red processes, then yellow processes and finally green processes. In the event that resources become limited, resources are first withheld from Green Processes, then Yellow Processes. Resources are only withheld from Red Processes if failure to achieve outcomes/goals is acceptable.

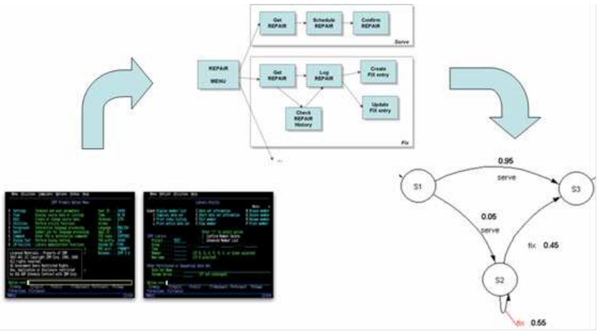
The Purpose / Example

A small example may illustrate the Business Process Discovery technology that is required today. Automated Business Process Discovery tools capture the required data, and transform it into a structured dataset for the actual diagnosis; A major challenge is the grouping of repetitive actions from the users into meaningful events. Next, these Business process discovery tools propose probabilistic process models. Probabilistic behavior is essential for the analysis and the diagnosis of the processes. The following shows an example where a probabilistic repair-process is recovered from user actions. The "as-is" process model shows exactly where the pain is in this business. Five percent faulty repairs is a bad sign, but worse, the repetitive fixes that are needed to complete those repairs are cumbersome.

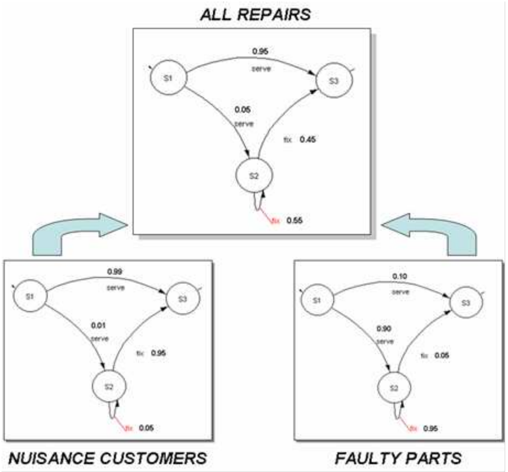
History

- Business intelligence (BI) emerged more than 20 years ago and is critical for reporting what is happening within an organization's systems. Yet current BI applications and data mining technologies are not always suited for evaluating the level of detail required to analyze unstructured data and the human dynamics of business processes.

- Six-Sigma and other quantitative approaches to business process improvement have been employed for over a decade with varying degrees of success. A major limitation to the success of these approaches is the availability of accurate data to form the basis of the analysis. With BPD, many six-sigma organizations are finding the ability to extend their analysis into major business processes effectively.
- Process mining According to researchers at Eindhoven University of Technology, (PM) emerged as a scientific discipline around 1990 when techniques like the Alpha algorithm made it possible to extract process models (typically represented as Petri nets) from event logs. However, the recognition of this wanna-be scientific discipline is extremely limited within few countries. As the hype of Process Mining carried by Eindhoven University of Technology growing, more and more criticisms have emerged pointing out that Process Mining is no more than a set of algorithms which solves a specific and simple business problem: business process discovery and auxiliary evaluation methods. Today, there are over 100 process mining algorithms that are able to discover process models that also include concurrency, e.g., genetic process discovery techniques, heuristic mining algorithms, region-based mining algorithms, and fuzzy mining algorithms.


Get REPAIR
Schedule REPAIR
Confirm REPAIR
Serve
REPAIR MENU
Get REPAIR
Log REPAIR
Create FIX entry
Check REPAIR History
Update FIX entry
Fix
0.95
serve
S1
S3
0.05
serve
fix 0.45
S2
fix 0.55

A deeper analysis of the "as-is" process data may reveal which are the faulty parts that are responsible for the overall behavior in this example. It may lead to the discovery of subgroups of repairs that actually need management focus for improvement.


ALL REPAIRS
0.95
serve
S1
S3
0.05
serve
fix 0.45
S2
fix 0.55
0.99
serve
S1
S3
0.01
serve
fix 0.95
S2
fix 0.05
0.10
serve
S1
S3
0.90
serve
fix 0.05
S2
fix 0.95
NUISANCE CUSTOMERS
FAULTY PARTS

In this case, it would become obvious that the faulty parts are also responsible for the repetitive fixes. Similar applications have been documented, such as a Healthcare Insurance Provider case where in 4 months the ROI of Business Process Analysis was earned from precisely comprehending its claims handling process and discovering the faulty parts.

Information System

An information system (IS) is any organized system for the collection, organization, storage and communication of information. More specifically, it is the study of complementary networks that people and organizations use to collect, filter, process, create and distribute data.

"An information system (IS) is a group of components that interact to produce information"

A computer information system is a system composed of people and computers that processes or interprets information. The term is also sometimes used in more restricted senses to refer to only the software used to run a computerized database or to refer to only a computer system.

Information system is an academic study of systems with a specific reference to information and the complementary networks of hardware and software that people and organizations use to collect, filter, process, create and also distribute data. An emphasis is placed on an Information System having a definitive Boundary, Users, Processors, Stores, Inputs, Outputs and the aforementioned communication networks.

Any specific information system aims to support operations, management and decision-making. An information system is the information and communication technology (ICT) that an organization uses, and also the way in which people interact with this technology in support of business processes.

Some authors make a clear distinction between information systems, computer systems, and business processes. Information systems typically include an ICT component but are not purely concerned with ICT, focusing instead on the end use of information technology. Information systems are also different from business processes. Information systems help to control the performance of business processes.

Alter argues for advantages of viewing an information system as a special type of work system. A work system is a system in which humans or machines perform processes and activities using resources to produce specific products or services for customers. An information system is a work system whose activities are devoted to capturing, transmitting, storing, retrieving, manipulating and displaying information.

As such, information systems inter-relate with data systems on the one hand and activity systems on the other. An information system is a form of communication system in which data represent and are processed as a form of social memory. An information system can also be considered a semi-formal language which supports human decision making and action.

Information systems are the primary focus of study for organizational informatics.

Overview

Silver et al. (1995) provided two views on IS that includes software, hardware, data, people, and procedures. Zheng provided another system view of information system which also adds processes and essential system elements like environment, boundary, purpose, and interactions. The Association for Computing Machinery defines "Information systems specialists [as] focus[ing] on integrating information technology solutions and business processes to meet the information needs of businesses and other enterprises."

There are various types of information systems, for example: transaction processing systems, decision support systems, knowledge management systems, learning management systems, database management systems, and office information systems. Critical to most information systems are information technologies, which are typically designed to enable humans to perform tasks for which the human brain is not well suited, such as: handling large amounts of information, performing complex calculations, and controlling many simultaneous processes.

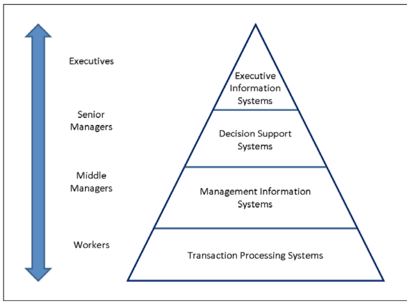
Information technologies are a very important and malleable resource available to executives. Many companies have created a position of chief information officer (CIO) that sits on the executive board with the chief executive officer (CEO), chief financial officer (CFO), chief operating officer (COO), and chief technical officer (CTO). The CTO may also serve as CIO, and vice versa. The chief information security officer (CISO) focuses on information security management.

The six components that must come together in order to produce an information system are:

1. Hardware: The term hardware refers to machinery. This category includes the computer itself, which is often referred to as the central processing unit (CPU), and all of its support equipments. Among the support equipments are input and output devices, storage devices and communications devices.
2. Software: The term software refers to computer programs and the manuals (if any) that support them. Computer programs are machine-readable instructions that direct the circuitry within the hardware parts of the system to function in ways that produce useful information from data. Programs are generally stored on some input / output medium, often a disk or tape.
3. Data: Data are facts that are used by programs to produce useful information. Like programs, data are generally stored in machine-readable form on disk or tape until the computer needs them.
4. Procedures: Procedures are the policies that govern the operation of a computer system. "Procedures are to people what software is to hardware" is a common analogy that is used to illustrate the role of procedures in a system.
5. People: Every system needs people if it is to be useful. Often the most over-looked element of the system are the people, probably the component that most influence the success or failure of information systems. This includes "not only the users, but those who operate and service the computers, those who maintain the data, and those who support the network of computers." <Kroenke, D. M. (2015). MIS Essentials. Pearson Education>
6. Feedback: it is another component of the IS, that defines that an IS may be provided with a feedback (Although this component isn't necessary to function).

Data is the bridge between hardware and people. This means that the data we collect is only data, until we involve people. At that point, data is now information.

Types of Information System


Executives
Senior Managers
Middle Managers
Workers
Executive Information Systems
Decision Support Systems
Management Information Systems
Transaction Processing Systems

A four level

The “classic” view of Information systems found in the textbooks in the 1980s was of a pyramid of systems that reflected the hierarchy of the organization, usually transaction processing systems at the bottom of the pyramid, followed by management information systems, decision support systems, and ending with executive information systems at the top. Although the pyramid model remains useful, since it was first formulated a number of new technologies have been developed and new categories of information systems have emerged, some of which no longer fit easily into the original pyramid model.

Some examples of such systems are:

- data warehouses
- enterprise resource planning
- enterprise systems
- expert systems
- search engines
- geographic information system
- global information system
- office automation.

A computer(-based) information system is essentially an IS using computer technology to carry out some or all of its planned tasks. The basic components of computer-based information systems are:

- *Hardware*- these are the devices like the monitor, processor, printer and keyboard, all of which work together to accept, process, show data and information.
- *Software*- are the programs that allow the hardware to process the data.
- *Databases*- are the gathering of associated files or tables containing related data.
- *Networks*- are a connecting system that allows diverse computers to distribute resources.
- *Procedures*- are the commands for combining the components above to process information and produce the preferred output.

The first four components (hardware, software, database, and network) make up what is known as the information technology platform. Information technology workers could then use these components to create information systems that watch over safety measures, risk and the management of data. These actions are known as information technology services.

Certain information systems support parts of organizations, others support entire organizations, and still others, support groups of organizations. Recall that each department or functional area within an organization has its own collection of application programs, or information systems. These functional area information systems (FAIS) are supporting pillars for more general IS namely, business intelligence systems and dashboards. As the name suggest, each FAIS support a particular function within the organization, e.g.: accounting IS, finance IS, production/operation management (POM) IS, marketing IS, and human resources IS. In finance and accounting, managers use IT systems to forecast revenues and business activity, to determine the best sources and uses of funds, and to perform audits to ensure that the organization is fundamentally sound and that all financial reports and documents are accurate. Other types of organizational information systems are FAIS, Transaction processing systems, enterprise resource planning, office automation system, management information system, decision support system, expert system, executive dashboard, supply chain management system, and electronic commerce system. Dashboards are a special form of IS that support all managers of the organization. They provide rapid access to timely information and direct access to structured information in the form of reports. Expert systems attempt to duplicate the work of human experts by applying reasoning capabilities, knowledge, and expertise within a specific domain.

Information System Development

Information technology departments in larger organizations tend to strongly influence the development, use, and application of information technology in the organizations. A series of methodologies and processes can be used to develop and use an information system. Many developers now use an engineering approach such as the system development life cycle (SDLC), which is a systematic procedure of developing an information system through stages that occur in sequence. Recent research aims at enabling and measuring the ongoing, collective development of such systems within an organization by the entirety of human actors themselves. An information system can be developed in house (within the organization) or outsourced. This can be accomplished by outsourcing certain components or the entire system. A specific case is the geographical distribution of the development team (offshoring, global information system).

A computer-based information system, following a definition of Langefors, is a technologically implemented medium for:

- recording, storing, and disseminating linguistic expressions,
- as well as for drawing conclusions from such expressions.

Geographic information systems, land information systems, and disaster information systems are examples of emerging information systems, but they can be broadly considered as spatial information systems. System development is done in stages which include:

- Problem recognition and specification

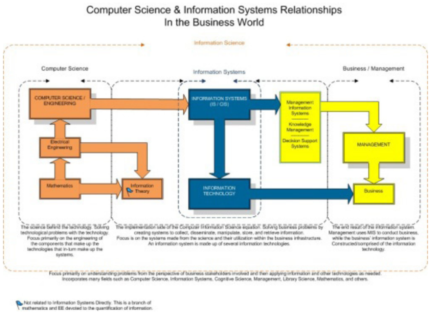
- Information gathering
- Requirements specification for the new system
- System design
- System construction
- System implementation
- Review and maintenance.

As an Academic Discipline

The field of study called *information systems* encompasses a variety of topics including systems analysis and design, computer networking, information security, database management and decision support systems. *Information management* deals with the practical and theoretical problems of collecting and analyzing information in a business function area including business productivity tools, applications programming and implementation, electronic commerce, digital media production, data mining, and decision support. *Communications and networking* deals with the telecommunication technologies. Information systems bridges business and computer science using the theoretical foundations of information and computation to study various business models and related algorithmic processes on building the IT systems within a computer science discipline. Computer information system(s) (CIS) is a field studying computers and algorithmic processes, including their principles, their software and hardware designs, their applications, and their impact on society, whereas IS emphasizes functionality over design.

Several IS scholars have debated the nature and foundations of Information Systems which has its roots in other reference disciplines such as Computer Science, Engineering, Mathematics, Management Science, Cybernetics, and others. Information systems also can be defined as a collection of hardware, software, data, people and procedures that work together to produce quality information.

Differentiating IS from Related Disciplines


Computer Science & Information Systems Relationships
In the Business World
Information Science
Computer Science
Information Systems
Business / Management
COMPUTER SCIENCE / ENGINEERING
Electrical Engineering
Mathematics
Information Theory
INFORMATION SYSTEMS (IS / CIS)
INFORMATION TECHNOLOGY
Management Information Systems
Knowledge Management
Decision Support Systems
MANAGEMENT
Business
The science behind the technology. Solving technological problems with the technology. Focus primarily on the engineering of the components that make up the technologies that in-turn make up the systems.
The implementation side of the Computer Information Science equation. Solving business problems by creating systems to collect, disseminate, manipulate, store, and retrieve information. Focus is on the systems made from the science and their utilization within the business infrastructure. An information system is made up of several information technologies.
The end result of the information system. Management uses MIS to conduct business, while the business' information system is Constructed/comprised of the information technology.
Focus primarily on understanding problems from the perspective of business stakeholders involved and then applying information and other technologies as needed. Incorporates many fields such as Computer Science, Information Systems, Cognitive Science, Management, Library Science, Mathematics, and others.
Not related to Information Systems Directly. This is a branch of mathematics and EE devoted to the quantification of information.

Information Systems relationship to Information Technology, Computer Science, Information Science, and Business.

Similar to computer science, other disciplines can be seen as both related and foundation disciplines of IS. The domain of study of IS involves the study of theories and practices related to the social and technological phenomena, which determine the development, use, and effects of information systems in organization and society. But, while there may be considerable overlap of the disciplines at the boundaries, the disciplines are still differentiated by the focus, purpose, and orientation of their activities.

In a broad scope, the term *Information Systems* is a scientific field of study that addresses the range of strategic, managerial, and operational activities involved in the gathering, processing, storing, distributing, and use of information and its associated technologies in society and organizations. The term information systems is also used to describe an organizational function that applies IS knowledge in industry, government agencies, and not-for-profit organizations. *Information Systems* often refers to the interaction between algorithmic processes and technology. This interaction can occur within or across organizational boundaries. An information system is the technology an organization uses and also the way in which the organizations interact with the technology and the way in which the technology works with the organization's business processes. Information systems are distinct from information technology (IT) in that an information system has an information technology component that interacts with the processes' components.

One problem with that approach is that it prevents the IS field from being interested in non-organizational use of ICT, such as in social networking, computer gaming, mobile personal usage, etc. A different way of differentiating the IS field from its neighbours is to ask, "Which aspects of reality are most meaningful in the IS field and other fields?" This approach, based on philosophy, helps to define not just the focus, purpose and orientation, but also the dignity, destiny and responsibility of the field among other fields. *International Journal of Information Management*, 30, 13-20.

Career Pathways

Information Systems have a number of different areas of work:

- IS strategy
- IS management
- IS development
- IS iteration
- IS organization

There is a wide variety of career paths in the information systems discipline. "Workers with specialized technical knowledge and strong communications skills will have the best prospects. Workers with management skills and an understanding of business practices and principles will have excellent opportunities, as companies are increasingly looking to technology to drive their revenue."

Information technology is important to the operation of contemporary businesses, it offers many employment opportunities. The information systems field includes the people in organizations who design and build information systems, the people who use those systems, and the people responsible for managing those systems. The demand for traditional IT staff such as programmers, business analysts, systems analysts, and designer is significant. Many well-paid jobs exist in areas of Information technology. At the top of the list is the chief information officer (CIO).

The CIO is the executive who is in charge of the IS function. In most organizations, the CIO works with the chief executive officer (CEO), the chief financial officer (CFO), and other senior executives. Therefore, he or she actively participates in the organization's strategic planning process.

Research

Information systems research is generally interdisciplinary concerned with the study of the effects of information systems on the behaviour of individuals, groups, and organizations. Hevner et al. (2004) categorized research in IS into two scientific paradigms including *behavioural science* which is to develop and verify theories that explain or predict human or organizational behavior and *design science* which extends the boundaries of human and organizational capabilities by creating new and innovative artifacts.

Salvatore March and Gerald Smith proposed a framework for researching different aspects of Information Technology including outputs of the research (research outputs) and activities to carry out this research (research activities). They identified research outputs as follows:

1. *Constructs* which are concepts that form the vocabulary of a domain. They constitute a conceptualization used to describe problems within the domain and to specify their solutions.
2. A *model* which is a set of propositions or statements expressing relationships among constructs.
3. A *method* which is a set of steps (an algorithm or guideline) used to perform a task. Methods are based on a set of underlying constructs and a representation (model) of the solution space.
4. An *instantiation* is the realization of an artifact in its environment.

Also research activities including:

1. *Build* an artifact to perform a specific task.
2. *Evaluate* the artifact to determine if any progress has been achieved.
3. Given an artifact whose performance has been evaluated, it is important to determine why and how the artifact worked or did not work within its environment. Therefore, *theorize* and *justify* theories about IT artifacts.

Although Information Systems as a discipline has been evolving for over 30 years now, the core focus or identity of IS research is still subject to debate among scholars. There are two main views around this debate: a narrow view focusing on the IT artifact as the core subject matter of IS research, and a broad view that focuses on the interplay between social and technical aspects of IT that is embedded into a dynamic evolving context. A third view calls on IS scholars to pay balanced attention to both the IT artifact and its context.

Since the study of information systems is an applied field, industry practitioners expect information systems research to generate findings that are immediately applicable in practice. This is not always the case however, as information systems researchers often explore behavioral issues in much more depth than practitioners would expect them to do. This may render information systems research results difficult to understand, and has led to criticism.

In the last ten years the business trend is represented by the considerable increasing of Information Systems Function (ISF) role, especially with regard the enterprise strategies and operations supporting. It became a key-factor to increase productivity and to support new value creation. To study an information system itself, rather than its effects, information systems models are used, such as EATPUT.

The international body of Information Systems researchers, the Association for Information Systems (AIS), and its Senior Scholars Forum Subcommittee on Journals (23 April 2007), proposed a 'basket' of journals that the AIS deems as 'excellent', and nominated: *Management Information Systems Quarterly* (MISQ), *Information Systems Research* (ISR), *Journal of the Association for Information Systems* (JAIS), *Journal of Management Information Systems* (JMIS), *European Journal of Information Systems* (EJIS), and *Information Systems Journal* (ISJ).

A number of annual information systems conferences are run in various parts of the world, the majority of which are peer reviewed. The AIS directly runs the International Conference on Information Systems (ICIS) and the Americas Conference on Information Systems (AMCIS), while AIS affiliated conferences include the Pacific Asia Conference on Information Systems (PACIS), European Conference on Information Systems (ECIS), the Mediterranean Conference on Information Systems (MCIS), the International Conference on Information Resources Management (Conf-IRM) and the Wuhan International Conference on E-Business (WHICEB). AIS chapter conferences include Australasian Conference on Information Systems (ACIS), Information Systems Research Conference in Scandinavia (IRIS), Information Systems International Conference (ISICO), Conference of the Italian Chapter of AIS (itAIS), Annual Mid-Western AIS Conference (MWAIS) and Annual Conference of the Southern AIS (SAIS). EDSIG, which is the special interest group on education of the AITP, organizes the Conference on Information Systems and Computing Education and the Conference on Information Systems Applied Research which are both held annually in November.

The Impact on Economic Models

- Microeconomic theory model
- Transaction cost theory
- Agency theory

Organizational Intelligence

Organizational Intelligence (OI) is the capability of an organization to comprehend and conclude knowledge relevant to its business purpose. In other words, it is the intellectual capacity of the entire organizations. With relevant organizational intelligence comes great potential value for companies and therefore organizations find study where their strengths and weaknesses lie in responding to change and complexity. Organizational Intelligence embraces both knowledge management and organizational learning, as it is the application of knowledge management concepts to a business environment, additionally including learning mechanisms, comprehension models and business value network models, such as the balanced scorecard concept. Organizational Intelligence consists of the ability to make sense of complex situations and act effectively, to interpret and act upon relevant events and signals in the environment. It also includes the ability to develop, share and use knowledge relevant to its business purpose as well as the ability to reflect and learn from experience

While organizations in the past have been viewed as compilations of tasks, products, employees, profit centers and processes, today they are seen as intelligent systems that are designed to manage knowledge. Scholars have shown that organizations engage in learning processes using tacit forms of intuitive knowledge, hard data stored in computer networks and information gleaned from the environment, all of which are used to make sensible decisions. Because this complex process involves large numbers of people interacting with diverse information systems, organizational intelligence is more than the aggregate intelligence of organizational members; it is the intelligence of the organization itself as a larger system.

Organizational Intelligence vs Operational Intelligence

Organizational Intelligence and operational intelligence are usually seen as subsets of business analytics, since both are types of know-how that have the goal of improving business performance across the enterprise. Operational Intelligence is often linked to or compared with real-time business intelligence (BI) since both deliver visibility and insight into business operations. Operational Intelligence differs from BI in being primarily activity-centric, whereas BI is primarily data-centric and relies on a database (or Hadoop cluster) as well as after-the-fact and report-based approaches to identifying patterns in data. By definition, Operational Intelligence works in real-time and transforms unstructured data streams—from log file, sensor, network and service data—into real-time, actionable intelligence.

While Operational Intelligence is activity-focused and BI is data-focused, Organizational Intelligence differs from these other approaches in being workforce- or organization-focused. Organizational Intelligence helps companies understand the relationships that drive their business—by identifying communities as well as employee workflow and collaborative communications patterns across geographies, divisions, and internal and external organizations.

Information Process

There are many aspects that organizations must consider in the three steps that they take to gain information. Without these considerations, organizations may experience strategic challenges.

Acquiring Information

First of all, organizations must acquire applicable information to make beneficial predictions. An organization must ask what they already know and need to know. They must also know the timeframe in which the information is needed and where and to find it. To make the best judgements, they must also evaluate the value of the information. Seemingly valuable information that costs more to find than gain from can hurt the company. If judged valuable, the organization must find the most efficient means of acquiring it.

Processing Information

After acquiring the right information, an organization must know how to properly process it. They need to know how they can make new information more retrievable and how they can make sure that the information gets disseminated to the right people. The organization must figure out how to secure it and how long and if long, how they need to preserve it.

Utilization of Information

The last step includes the utilization of the information. An organization should ask themselves if they are looking at the right information and if so, if they are placing them in the right context. They must consider the possible environmental changes alter the informational value and determine all the relevant connections and patterns. Not forgetting to know if they are including the right people in the decision making process and if there are any technology that can improve the decision making.

Organizational Ignorance

There are briefly four dimensions of problems that many organizations face when dealing with information. This is also referred to as organizational ignorance.

Uncertainty

An organization may be uncertain when it does not possess enough or the right information. To exemplify, a company may be uncertain in a competitive landscape because it does not have enough information to see how the competitors will act. This does not imply that the context of the situation is complex or unclear. Uncertainty can even exist when the range of possibilities is small and simple. There are different degrees of uncertainty. First of all an organization can be completely determined (complete certainty), have some probabilities (risk), probabilities estimated with lesser confidence (subjective uncertainty), unknown probabilities (traditional uncertainty) or undefined (complete uncertainty). However even with the lack of clarity, uncertainty assumes that the context of the problem is clear and well-understood.

Complexity

An organization may be processing more information than they can manage. Complexity doesn’t always correlate with vagueness or unpredictability. Rather, it occurs when there are too much or when the scope is too large to process. Organizations with complexity problems have interrelated variables, solutions and methods. Managing these problems is dependent of the individuals and the organizations. For instance, uninformed and novices must deal with each elements and relationships one by one but experts can perceive the situation better and find familiar patterns more easily. Organizations facing complexity must have the capacity to locate, map, collect, share, exploit on what the organizations need to know.

Ambiguity

An organization may not have a conceptual framework for interpreting the information. If uncertainty represents not having answers, and complexity represents difficulty in finding them, ambiguity represents not being able to formulate the right questions. Ambiguity cannot be resolved by increasing the amount of information. An organization must be able to interpret and explain the information in collective agreement. Hypotheses should be continuously made and discussed and key communication activities such as face-to-face conversations must be made. Resolving ambiguity in the earlier stages than competitors gives organizations much advantage because it helps organizations to make more appropriate and strategic decisions and have better awareness.

Equivocality

An organization may be having competing frameworks for interpreting a job. Equivocality refers to multiple interpretations of the field. Each interpretation is unambiguous but differ from each other and they may be mutually exclusive or in conflict. Equivocality result not only because everyone's experiences and values are unique but also from unreliable or conflicting preferences and goals, different interests or vague roles and responsibilities.

Information Organization and Culture

A culture of the organization describes how the organization will work in order to succeed. It can simply be described as the organization's atmosphere or values. Organizational culture is important because it can be used as a successful leadership tool to shape and improve the organization. Once the culture is settled, it can be used by the leader to deliver his/her vision to the organization. Moreover, if the leader deeply understands the organizational culture, he/she can also use it to predict a future outcome in certain situations.

Control

An organization with control culture is company oriented and reality oriented. They will succeed by controlling and keeping restrictions. The organization will value timeliness of information, security and hierarchical standardization. They make plans and maintain a process. This organization has stability, predictability and authority. For example, an organization with control culture can be monarchy.

Competence

An organization with competence culture is company oriented and possibility oriented. They will succeed by being the best with exclusivity of the information. The organization values efficiency, accuracy and achievement. They look for creativity and expertise from the people in the organization. For example, an organization with competence culture can be...

Cultivation

An organization with cultivation culture is people oriented and possibility oriented. They will succeed by growing people, who fulfill the shared vision. The organization values self-actualization and brilliance. They also prioritizes the idea from people. For example, an organization with cultivation culture can be technological utopianism.

Collaboration

An organization with collaboration culture is people oriented and reality oriented. They will succeed by working together. The organization values affiliation and teamwork. They also prioritizes people in the organization. This organization has accessibility and inclusiveness of information. For example, an organization with collaboration culture can be anarchy.

Organizational Intelligence and Innovation

An organization's leadership effectiveness is closely related to the organization's intelligence and innovation. There are six leadership factors that determines organization' atmosphere : flexibility (how freely people can communicate with each other and innovate), responsibility (sense of loyalty to the organization), the standards set by people in the organization, appropriate feedback and rewards, the clear vision shared by people and the amount of commitment to the goal. Combination of these factors result in six different leadership styles: Coercive/Commanding, Authoritative/ Visionary, Affiliative, Democratic, Coaching and Pacesetting.

Furthermore, organizational intelligence is a collection of individual intelligence. The leadership style of the organization and its atmosphere are related to the organization's innovation. Innovation happens when there are new information getting shared and processed efficiently in the organization.

Theories

Round Table

In King Arthur's Round Table, Harvard professor David Perkins uses the metaphor of the Round Table to discuss how collaborative conversations create smarter organizations. The Round Table is one of the most familiar stories of Arthurian legend since it's meant to signal the shift in power from a king who normally sat at the head of a long table and made long pronouncements while everyone else listened. By reducing hierarchy and making collaboration easier, Arthur discovered an important source of power—organizational intelligence—that allowed him to unite medieval England.

Lawnmower Paradox

The lawnmower paradox, another metaphor from Perkins' book, describes the fact that, while pooling physical effort is easy, pooling mental effort is hard. "It's a lot easier for 10 people to collaborate on mowing a large lawn than for 10 people to collaborate on designing a lawnmower." An organization's intelligence is reflected by the types of conversations—face-to-face and electronic, from the mailroom to the boardroom—which members have with one another. "At the top, top level, organizational intelligence depends on ways of interacting with one another that show good knowledge processing and positive symbolic conduct."

Harold Wilensky argued that organizational intelligence benefited from healthy argument and constructive rivalry.

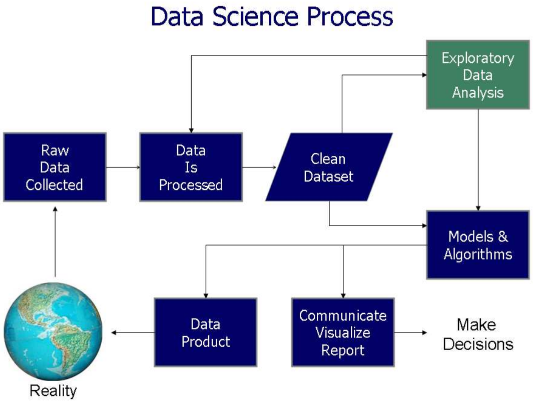
Data Visualization

Data visualization or data visualisation is viewed by many disciplines as a modern equivalent of visual communication. It involves the creation and study of the visual representation of data, meaning “information that has been abstracted in some schematic form, including attributes or variables for the units of information”.

A primary goal of data visualization is to communicate information clearly and efficiently via statistical graphics, plots and information graphics. Numerical data may be encoded using dots, lines, or bars, to visually communicate a quantitative message. Effective visualization helps users analyze and reason about data and evidence. It makes complex data more accessible, understandable and usable. Users may have particular analytical tasks, such as making comparisons or understanding causality, and the design principle of the graphic (i.e., showing comparisons or showing causality) follows the task. Tables are generally used where users will look up a specific measurement, while charts of various types are used to show patterns or relationships in the data for one or more variables.

Data visualization is both an art and a science . It is viewed as a branch of descriptive statistics by some, but also as a grounded theory development tool by others. The rate at which data is generated has increased. Data created by internet activity and an expanding number of sensors in the environment, such as satellites, are referred to as “Big Data”. Processing, analyzing and communicating this data present a variety of ethical and analytical challenges for data visualization. The field of data science and practitioners called data scientists have emerged to help address this challenge.

Overview



Data visualization is one of the steps in analyzing data and presenting it to users.

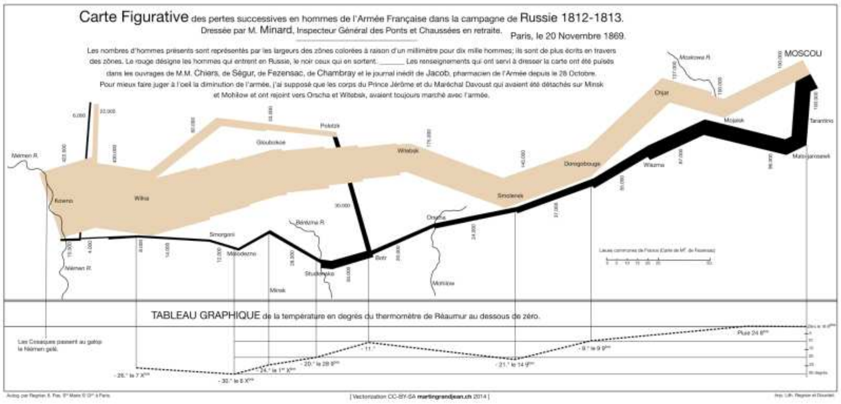
Data visualization refers to the techniques used to communicate data or information by encoding it as visual objects (e.g., points, lines or bars) contained in graphics. The goal is to communicate information clearly and efficiently to users. It is one of the steps in data analysis or data science. According to Friedman (2008) the “main goal of data visualization is to communicate information clearly and effectively through graphical means. It doesn’t mean that data visualization needs to look boring to be functional or extremely sophisticated to look beautiful. To convey ideas effectively, both aesthetic form and functionality need to go hand in hand, providing insights into a rather sparse and complex data set by communicating its key-aspects in a more intuitive way. Yet designers often fail to achieve a balance between form and function, creating gorgeous data visualizations which fail to serve their main purpose — to communicate information”.

Indeed, Fernanda Viegas and Martin M. Wattenberg have suggested that an ideal visualization should not only communicate clearly, but stimulate viewer engagement and attention.

Not limited to the communication of an information, a well-crafted data visualization is also a way to a better understanding of the data (in a data-driven research perspective), as it helps uncover trends, realize insights, explore sources, and tell stories.

Data visualization is closely related to information graphics, information visualization, scientific visualization, exploratory data analysis and statistical graphics. In the new millennium, data visualization has become an active area of research, teaching and development. According to Post et al. (2002), it has united scientific and information visualization.

Characteristics of Effective Graphical Displays



Charles Joseph Minard's 1869 diagram of Napoleon's March - an early example of an information graphic.

Professor Edward Tufte explained that users of information displays are executing particular *analytical tasks* such as making comparisons or determining causality. The *design principle* of the information graphic should support the analytical task, showing the comparison or causality.

In his 1983 book *The Visual Display of Quantitative Information*, Edward Tufte defines 'graphical displays' and principles for effective graphical display in the following passage: "Excellence in statistical graphics consists of complex ideas communicated with clarity, precision and efficiency. Graphical displays should:

- show the data
- induce the viewer to think about the substance rather than about methodology, graphic design, the technology of graphic production or something else
- avoid distorting what the data has to say
- present many numbers in a small space
- make large data sets coherent
- encourage the eye to compare different pieces of data
- reveal the data at several levels of detail, from a broad overview to the fine structure

- serve a reasonably clear purpose: description, exploration, tabulation or decoration
- be closely integrated with the statistical and verbal descriptions of a data set.

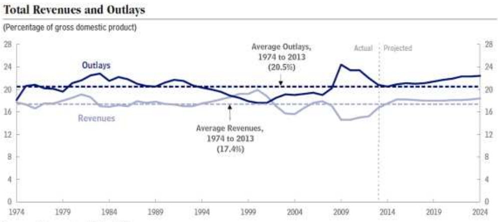
Graphics *reveal* data. Indeed graphics can be more precise and revealing than conventional statistical computations."

For example, the Minard diagram shows the losses suffered by Napoleon's army in the 1812–1813 period. Six variables are plotted: the size of the army, its location on a two-dimensional surface (x and y), time, direction of movement, and temperature. The line width illustrates a comparison (size of the army at points in time) while the temperature axis suggests a cause of the change in army size. This multivariate display on a two dimensional surface tells a story that can be grasped immediately while identifying the source data to build credibility. Tufte wrote in 1983 that: "It may well be the best statistical graphic ever drawn."

Not applying these principles may result in misleading graphs, which distort the message or support an erroneous conclusion. According to Tufte, chartjunk refers to extraneous interior decoration of the graphic that does not enhance the message, or gratuitous three dimensional or perspective effects. Needlessly separating the explanatory key from the image itself, requiring the eye to travel back and forth from the image to the key, is a form of "administrative debris." The ratio of "data to ink" should be maximized, erasing non-data ink where feasible.

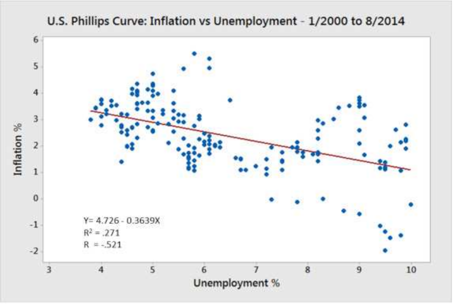
The Congressional Budget Office summarized several best practices for graphical displays in a June 2014 presentation. These included: a) Knowing your audience; b) Designing graphics that can stand alone outside the context of the report; and c) Designing graphics that communicate the key messages in the report.

Quantitative Messages



Source: Congressional Budget Office.

A time series illustrated with a line chart demonstrating trends in U.S. federal spending and revenue over time.



Source Data: FRED Database
Inflation: CPI for All Urban Consumers

A scatterplot illustrating negative correlation between two variables (inflation and unemployment) measured at points in time.

Author Stephen Few described eight types of quantitative messages that users may attempt to understand or communicate from a set of data and the associated graphs used to help communicate the message:

1. Time-series: A single variable is captured over a period of time, such as the unemployment rate over a 10-year period. A line chart may be used to demonstrate the trend.
2. Ranking: Categorical subdivisions are ranked in ascending or descending order, such as a ranking of sales performance (the *measure*) by sales persons (the *category*, with each sales person a *categorical subdivision*) during a single period. A bar chart may be used to show the comparison across the sales persons.
3. Part-to-whole: Categorical subdivisions are measured as a ratio to the whole (i.e., a percentage out of 100%). A pie chart or bar chart can show the comparison of ratios, such as the market share represented by competitors in a market.
4. Deviation: Categorical subdivisions are compared against a reference, such as a comparison of actual vs. budget expenses for several departments of a business for a given time period. A bar chart can show comparison of the actual versus the reference amount.
5. Frequency distribution: Shows the number of observations of a particular variable for given interval, such as the number of years in which the stock market return is between intervals such as 0-10%, 11-20%, etc. A histogram, a type of bar chart, may be used for this analysis. A boxplot helps visualize key statistics about the distribution, such as median, quartiles, outliers, etc.
6. Correlation: Comparison between observations represented by two variables (X,Y) to determine if they tend to move in the same or opposite directions. For example, plotting unemployment (X) and inflation (Y) for a sample of months. A scatter plot is typically used for this message.
7. Nominal comparison: Comparing categorical subdivisions in no particular order, such as the sales volume by product code. A bar chart may be used for this comparison.
8. Geographic or geospatial: Comparison of a variable across a map or layout, such as the unemployment rate by state or the number of persons on the various floors of a building. A cartogram is a typical graphic used.

Analysts reviewing a set of data may consider whether some or all of the messages and graphic types above are applicable to their task and audience. The process of trial and error to identify meaningful relationships and messages in the data is part of exploratory data analysis.

Visual Perception and Data Visualization

A human can distinguish differences in line length, shape orientation, and color (hue) readily without significant processing effort; these are referred to as “pre-attentive attributes.” For example, it may require significant time and effort (“attentive processing”) to identify the number of times the digit “5” appears in a series of numbers; but if that digit is different in size, orientation, or color, instances of the digit can be noted quickly through pre-attentive processing.

Effective graphics take advantage of pre-attentive processing and attributes and the relative strength of these attributes. For example, since humans can more easily process differences in line length than surface area, it may be more effective to use a bar chart (which takes advantage of line length to show comparison) rather than pie charts (which use surface area to show comparison).

Human Perception/Cognition and Data Visualization

There is a human side to data visualization. With the "studying [of] human perception and cognition ..." we are better able to understand the target of the data which we display. Cognition refers to processes in human beings like perception, attention, learning, memory, thought, concept formation, reading, and problem solving. The basis of data visualization evolved because as a picture is worth a thousand words, data displayed graphically allows for an easier comprehension of the information. Proper visualization provides a different approach to show potential connections, relationships, etc. which are not as obvious in non-visualized quantitative data. Visualization becomes a means of data exploration. Human brain neurons involve multiple functions but 2/3 of the brain's neurons are dedicated to vision. With a well-developed sense of sight, analysis of data can be made on data, whether that data is quantitative or qualitative. Effective visualization follows from understanding the processes of human perception and being able to apply this to intuitive visualizations is important. Understanding how humans see and organize the world is critical to effectively communicating data to the reader. This leads to more intuitive designs.

History of Data Visualization

There is a history of data visualization: beginning in the 2nd century C.E. with data arrangement into columns and rows and evolving to the initial quantitative representations in the 17th century. According to the Interaction Design Foundation, French philosopher and mathematician René Descartes laid the ground work for Scotsman William Playfair. Descartes developed a two-dimensional coordinate system for displaying values, which in the late 18th century Playfair saw potential for graphical communication of quantitative data. In the second half of the 20th century, Jacques Bertin used quantitative graphs to represent information "intuitively, clearly, accurately, and efficiently". John Tukey and more notably Edward Tufte pushed the bounds of data visualization. Tukey with his new statistical approach: exploratory data analysis and Tufte with his book "The Visual Display of Quantitative Information", the path was paved for refining data visualization techniques for more than statisticians. With the progression of technology came the progression of data visualization; starting with hand drawn visualizations and evolving into more technical applications – including interactive designs leading to software visualization. Programs like SAS, SOFA, R, Minitab, and more allow for data visualization in the field of statistics. Other data visualization applications, more focused and unique to individuals, programming languages such as D3, Python and JavaScript help to make the visualization of quantitative data a possibility.

Terminology

Data visualization involves specific terminology, some of which is derived from statistics. For example, author Stephen Few defines two types of data, which are used in combination to support a meaningful analysis or visualization:

- Categorical: Text labels describing the nature of the data, such as "Name" or "Age". This term also covers qualitative (non-numerical) data.

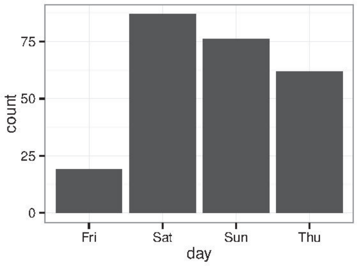
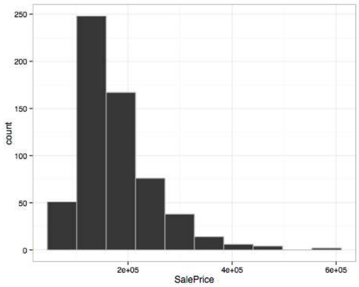
- Quantitative: Numerical measures, such as “25” to represent the age in years.

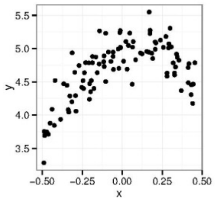
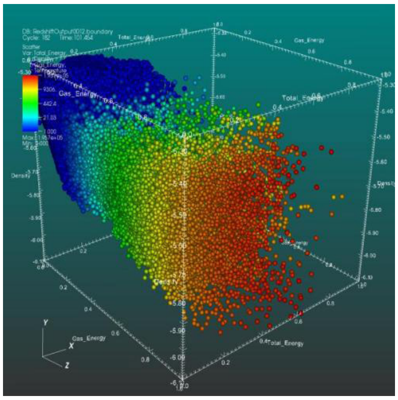
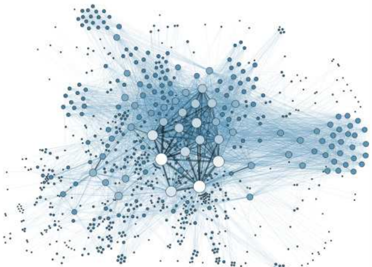
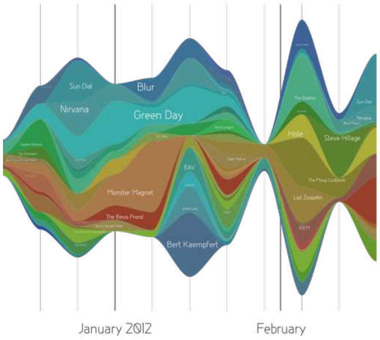
Two primary types of information displays are tables and graphs.

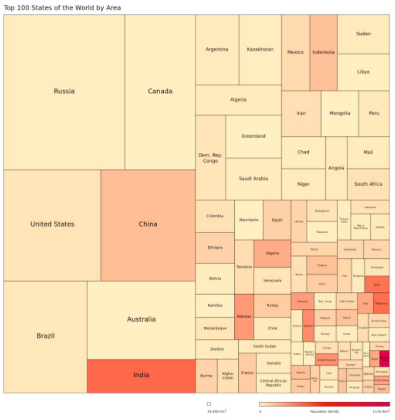
- A *table* contains quantitative data organized into rows and columns with categorical labels. It is primarily used to look up specific values. In the example above, the table might have categorical column labels representing the name (a *qualitative variable*) and age (a *quantitative variable*), with each row of data representing one person (the sampled *experimental unit* or *category subdivision*).
- A *graph* is primarily used to show relationships among data and portrays values encoded as *visual objects* (e.g., lines, bars, or points). Numerical values are displayed within an area delineated by one or more *axes*. These axes provide *scales* (quantitative and categorical) used to label and assign values to the visual objects. Many graphs are also referred to as *charts*.

KPI Library has developed the “Periodic Table of Visualization Methods,” an interactive chart displaying various data visualization methods. It includes six types of data visualization methods: data, information, concept, strategy, metaphor and compound.

Examples of Diagrams used for Data Visualization

	Name	Visual Dimensions	Example Usages
 Bar chart of tips by day of week	Bar chart	• length/count • category • (color)	• Comparison of values, such as sales performance for several persons or businesses in a single time period. For a single variable measured over time (trend) a line chart is preferable.
 Histogram of housing prices	Histogram	• bin limits • count/length • (color)	• Determining frequency of annual stock market percentage returns within particular ranges (bins) such as 0-10%, 11-20%, etc. The height of the bar represents the number of observations (years) with a return % in the range represented by the bin.

 Basic scatterplot of two variables	Scatter plot	- x position - y position - (symbol/ glyph) - (color) - (size)	- Determining the relationship (e.g., correlation) between unemployment (x) and inflation (y) for multiple time periods.
 Scatter Plot	Scatter plot (3D)	- position x - position y - position z - color	
 Network Analysis	Network	- nodes size - nodes color - ties thickness - ties color - spatialization	- Finding clusters in the network (e.g. grouping Facebook friends into different clusters). - Determining the most influential nodes in the network (e.g. A company wants to target a small group of people on Twitter for a marketing campaign).
 Streamgraph	Streamgraph	- width - color - time (flow)	

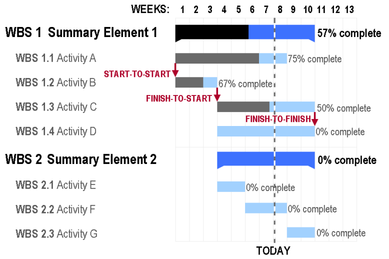
<table>
<tr><td>

Treemap</td><td>Treemap</td><td>

- size
- color

</td><td>

- disk space by location / file type

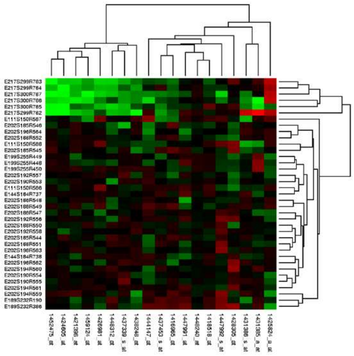
</td></tr>
<tr><td>

Gantt Chart</td><td>Gantt chart</td><td>

- color
- time (flow)

</td><td>

- schedule / progress, e.g. in project planning

</td></tr>
<tr><td>
Heat map</td><td>Heat map</td><td>

- row
- column
- cluster
- color

</td><td>

- Analyzing risk, with green, yellow and red representing low, medium, and high risk, respectively.

</td></tr>
</table>

Other Perspectives

There are different approaches on the scope of data visualization. One common focus is on information presentation, such as Friedman (2008) presented it. In this way Friendly (2008) presumes two main parts of data visualization: statistical graphics, and thematic cartography. In this line the “Data Visualization: Modern Approaches” (2007) article gives an overview of seven subjects of data visualization:

- Articles & resources
- Displaying connections
- Displaying data

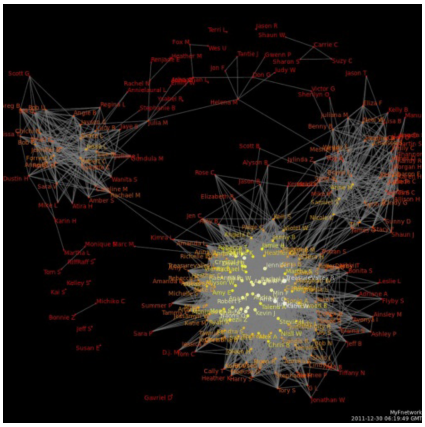
- Displaying news
- Displaying websites
- Mind maps
- Tools and services

All these subjects are closely related to graphic design and information representation.

On the other hand, from a computer science perspective, Frits H. Post (2002) categorized the field into a number of sub-fields:

- Information visualization
- Interaction techniques and architectures
- Modelling techniques
- Multiresolution methods
- Visualization algorithms and techniques
- Volume visualization

Data Presentation Architecture



A data visualization from social media

Data presentation architecture (DPA) is a skill-set that seeks to identify, locate, manipulate, format and present data in such a way as to optimally communicate meaning and proper knowledge.

Historically, the term *data presentation architecture* is attributed to Kelly Lautt: “Data Presentation Architecture (DPA) is a rarely applied skill set critical for the success and value of Business Intelligence. Data presentation architecture weds the science of numbers, data and statistics in discovering valuable information from data and making it usable, relevant and actionable with the arts of data visualization, communications, organizational psychology and change management in order to provide business intelligence solutions with the data scope, delivery timing, format and visualizations that will most effectively support and drive operational, tactical and strategic behaviour toward understood business (or organizational) goals. DPA is neither an IT nor a business skill set but exists as a separate field of expertise. Often confused with data visualization, data presentation architecture is a much broader skill set that includes determining what data on what schedule and in what exact format is to be presented, not just the best way to present data that has already been chosen (which is data visualization). Data visualization skills are one element of DPA."

Objectives

DPA has two main objectives:

- To use data to provide knowledge in the most efficient manner possible (minimize noise, complexity, and unnecessary data or detail given each audience's needs and roles)
- To use data to provide knowledge in the most effective manner possible (provide relevant, timely and complete data to each audience member in a clear and understandable manner that conveys important meaning, is actionable and can affect understanding, behavior and decisions)

Scope

With the above objectives in mind, the actual work of data presentation architecture consists of:

- Creating effective delivery mechanisms for each audience member depending on their role, tasks, locations and access to technology
- Defining important meaning (relevant knowledge) that is needed by each audience member in each context
- Determining the required periodicity of data updates (the currency of the data)
- Determining the right timing for data presentation (when and how often the user needs to see the data)
- Finding the right data (subject area, historical reach, breadth, level of detail, etc.)
- Utilizing appropriate analysis, grouping, visualization, and other presentation formats

Related Fields

DPA work shares commonalities with several other fields, including:

- Business analysis in determining business goals, collecting requirements, mapping processes.
- Business process improvement in that its goal is to improve and streamline actions and decisions in furtherance of business goals
- Data visualization in that it uses well-established theories of visualization to add or highlight meaning or importance in data presentation.
- Graphic or user design: As the term DPA is used, it falls just short of design in that it does

not consider such detail as colour palates, styling, branding and other aesthetic concerns, unless these design elements are specifically required or beneficial for communication of meaning, impact, severity or other information of business value. For example:

  - choosing locations for various data presentation elements on a presentation page (such as in a company portal, in a report or on a web page) in order to convey hierarchy, priority, importance or a rational progression for the user is part of the DPA skill-set.
  - choosing to provide a specific colour in graphical elements that represent data of specific meaning or concern is part of the DPA skill-set
- Information architecture, but information architecture's focus is on unstructured data and therefore excludes both analysis (in the statistical/data sense) and direct transformation of the actual content (data, for DPA) into new entities and combinations.
- Solution architecture in determining the optimal detailed solution, including the scope of data to include, given the business goals
- Statistical analysis or data analysis in that it creates information and knowledge out of data

Data Profiling

Data profiling is the process of examining the data available in an existing information data source (e.g. a database or a file) and collecting statistics or small but informative summaries about that data. The purpose of these statistics may be to:

1. Find out whether existing data can easily be used for other purposes
2. Improve the ability to search the data by tagging it with keywords, descriptions, or assigning it to a category
3. Give metrics on data quality including whether the data conforms to particular standards or patterns
4. Assess the risk involved in integrating data for new applications, including the challenges of joins
5. Discover metadata of the source database, including value patterns and distributions, key candidates, foreign-key candidates, and functional dependencies
6. Assess whether known metadata accurately describes the actual values in the source database
7. Understanding data challenges early in any data intensive project, so that late project surprises are avoided. Finding data problems late in the project can lead to delays and cost overruns.
8. Have an enterprise view of all data, for uses such as master data management where key data is needed, or data governance for improving data quality.

Introduction

Data profiling refers to analyzing candidate data sources for a data warehouse in order to clarify the structure, content, relationships and derivation rules of the data. Profiling helps not only to understand anomalies and assess data quality, but also to discover, register, and assess enterprise metadata. Thus, the purpose of data profiling is both to validate metadata when it is available and to discover metadata when it is not. The result of the analysis is used both strategically, to determine suitability of the candidate source systems and give the basis for an early go/no-go decision, and tactically, to identify problems for later solution design, and to level sponsors' expectations.

How to do Data Profiling

Data profiling utilizes different kinds of descriptive statistics such as minimum, maximum, mean, mode, percentile, standard deviation, frequency, and variation as well as other aggregates such as count and sum. Additional metadata information obtained during data profiling could be the data type, length, discrete values, uniqueness, occurrence of null values, typical string patterns, and abstract type recognition. The metadata can then be used to discover problems such as illegal values, misspelling, missing values, varying value representation, and duplicates.

Different analyses are performed for different structural levels. E.g. single columns could be profiled individually to get an understanding of frequency distribution of different values, type, and use of each column. Embedded value dependencies can be exposed in a cross-columns analysis. Finally, overlapping value sets possibly representing foreign key relationships between entities can be explored in an inter-table analysis.

Normally, purpose-built tools are used for data profiling to ease the process. The computation complexity increases when going from single column, to single table, to cross-table structural profiling. Therefore, performance is an evaluation criterion for profiling tools.

When to Conduct Data Profiling

According to Kimball, data profiling is performed several times and with varying intensity throughout the data warehouse developing process. A light profiling assessment should be undertaken as soon as candidate source systems have been identified immediately after the acquisition of the DW/BI business requirements. The purpose is to clarify at an early stage if the right data is available at the appropriate detail level and that anomalies can be handled subsequently. If this is not the case the project may be terminated.

More detailed profiling is done prior to the dimensional modelling process in order to see what is required to convert data into the dimensional model. Detailed profiling extends into the ETL system design process in order to determine what data to extract and which filters to apply.

Additionally, data may be conducted in the data warehouse development process after data has been loaded into staging, the data marts, etc. Conducting data at these stages helps ensure that data cleaning and transformations have been done correctly according to requirements.

Benefits

The benefits of data profiling are to improve data quality, shorten the implementation cycle of major projects, and improve understanding of data for users. Discovering business knowledge embedded in data itself is one of the significant benefits derived from data profiling. Data profiling is one of the most effective technologies for improving data accuracy in corporate databases.

Although data profiling is effective and useful for each sector of our daily life, it can be challenging not to slip into "analysis paralysis".

Data Cleansing

Data cleansing, data cleaning, or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database and refers to identifying incomplete, incorrect, inaccurate or irrelevant parts of the data and then replacing, modifying, or deleting the dirty or coarse data. Data cleansing may be performed interactively with data wrangling tools, or as batch processing through scripting.

After cleansing, a data set should be consistent with other similar data sets in the system. The inconsistencies detected or removed may have been originally caused by user entry errors, by corruption in transmission or storage, or by different data dictionary definitions of similar entities in different stores. Data cleansing differs from data validation in that validation almost invariably means data is rejected from the system at entry and is performed at the time of entry, rather than on batches of data.

The actual process of data cleansing may involve removing typographical errors or validating and correcting values against a known list of entities. The validation may be strict (such as rejecting any address that does not have a valid postal code) or fuzzy (such as correcting records that partially match existing, known records). Some data cleansing solutions will clean data by cross checking with a validated data set. A common data cleansing practice is data enhancement, where data is made more complete by adding related information. For example, appending addresses with any phone numbers related to that address. Data cleansing may also involve activities like, harmonization of data, and standardization of data. For example, harmonization of short codes (st, rd, etc.) to actual words (street, road, etcetera). Standardization of data is a means of changing a reference data set to a new standard, ex, use of standard codes.

Motivation

Administratively, incorrect or inconsistent data can lead to false conclusions and misdirected investments on both public and private scales. For instance, the government may want to analyze population census figures to decide which regions require further spending and investment on infrastructure and services. In this case, it will be important to have access to reliable data to avoid erroneous fiscal decisions.

In the business world, incorrect data can be costly. Many companies use customer information databases that record data like contact information, addresses, and preferences. For instance, if the addresses are inconsistent, the company will suffer the cost of resending mail or even losing customers.

The profession of forensic accounting and fraud investigating uses data cleansing in preparing its data and is typically done before data is sent to a data warehouse for further investigation.

There are packages available so you can cleanse/wash address data while you enter it into your system. This is normally done via an API and will prompt staff as they type the address.

Data Quality

High-quality data needs to pass a set of quality criteria. Those include:

- Validity: The degree to which the measures conform to defined business rules or constraints. When modern database technology is used to design data-capture systems, validity is fairly easy to ensure: invalid data arises mainly in legacy contexts (where constraints were not implemented in software) or where inappropriate data-capture technology was used (e.g., spreadsheets, where it is very hard to limit what a user chooses to enter into a cell). Data constraints fall into the following categories:
    - *Data-Type Constraints* – e.g., values in a particular column must be of a particular datatype, e.g., Boolean, numeric (integer or real), date, etc.
    - *Range Constraints:* typically, numbers or dates should fall within a certain range. That is, they have minimum and/or maximum permissible values.
    - *Mandatory Constraints:* Certain columns cannot be empty.
    - *Unique Constraints:* A field, or a combination of fields, must be unique across a dataset. For example, no two persons can have the same social security number.
    - *Set-Membership constraints*: The values for a column come from a set of discrete values or codes. For example, a person's gender may be Female, Male or Unknown (not recorded).
    - *Foreign-key constraints*: This is the more general case of set membership. The set of values in a column is defined in a column of another table that contains unique values. For example, in a US taxpayer database, the "state" column is required to belong to one of the US's defined states or territories: the set of permissible states/territories is recorded in a separate States table. The term foreign key is borrowed from relational database terminology.
    - Regular expression patterns: Occasionally, text fields will have to be validated this way. For example, phone numbers may be required to have the pattern (999) 999-9999.
    - Cross-field validation: Certain conditions that utilize multiple fields must hold. For example, in laboratory medicine, the sum of the components of the differential white blood cell count must be equal to 100 (since they are all percentages). In a hospital database, a patient's date of discharge from hospital cannot be earlier than the date of admission.

- Decleansing is detecting errors and syntactically removing them for better programming.
- Accuracy: The degree of conformity of a measure to a standard or a true value. Accuracy is very hard to achieve through data-cleansing in the general case, because it requires accessing an external source of data that contains the true value: such "gold standard" data is often unavailable. Accuracy has been achieved in some cleansing contexts, notably customer contact data, by using external databases that match up zip codes to geographical locations (city and state), and also help verify that street addresses within these zip codes actually exist.
- Completeness: The degree to which all required measures are known. Incompleteness is almost impossible to fix with data cleansing methodology: one cannot infer facts that were not captured when the data in question was initially recorded. (In some contexts, e.g., interview data, it may be possible to fix incompleteness by going back to the original source of data, i,e., re-interviewing the subject, but even this does not guarantee success because of problems of recall - e.g., in an interview to gather data on food consumption, no one is likely to remember exactly what one ate six months ago. In the case of systems that insist certain columns should not be empty, one may work around the problem by designating a value that indicates "unknown" or "missing", but supplying of default values does not imply that the data has been made complete.
- Consistency: The degree to which a set of measures are equivalent in across systems. Inconsistency occurs when two data items in the data set contradict each other: e.g., a customer is recorded in two different systems as having two different current addresses, and only one of them can be correct. Fixing inconsistency is not always possible: it requires a variety of strategies - e.g., deciding which data were recorded more recently, which data source is likely to be most reliable (the latter knowledge may be specific to a given organization), or simply trying to find the truth by testing both data items (e.g., calling up the customer).
- Uniformity: The degree to which a set data measures are specified using the same units of measure in all systems. In datasets pooled from different locales, weight may be recorded either in pounds or kilos, and must be converted to a single measure using an arithmetic transformation.

The term Integrity encompasses accuracy, consistency and some aspects of validation but is rarely used by itself in data-cleansing contexts because it is insufficiently specific. (For example, "referential integrity" is a term used to refer to the enforcement of foreign-key constraints above.)

The Process of Data Cleansing

- Data auditing: The data is audited with the use of statistical and database methods to detect anomalies and contradictions: this eventually gives an indication of the characteristics of the anomalies and their locations. Several commercial software packages will let you specify constraints of various kinds (using a grammar that conforms to that of a standard programming language, e.g., JavaScript or Visual Basic) and then generate code that checks the data for violation of these constraints. This process is referred to below in the bullets "workflow specification" and "workflow execution." For users who lack access to high-end cleansing software, Microcomputer database packages such as Microsoft Access or File Maker Pro will also let you perform such checks, on a constraint-by-constraint basis, interactively with little or no programming required in many cases.

- Workflow specification: The detection and removal of anomalies is performed by a sequence of operations on the data known as the workflow. It is specified after the process of auditing the data and is crucial in achieving the end product of high-quality data. In order to achieve a proper workflow, the causes of the anomalies and errors in the data have to be closely considered.
- Workflow execution: In this stage, the workflow is executed after its specification is complete and its correctness is verified. The implementation of the workflow should be efficient, even on large sets of data, which inevitably poses a trade-off because the execution of a data-cleansing operation can be computationally expensive.
- Post-processing and controlling: After executing the cleansing workflow, the results are inspected to verify correctness. Data that could not be corrected during execution of the workflow is manually corrected, if possible. The result is a new cycle in the data-cleansing process where the data is audited again to allow the specification of an additional workflow to further cleanse the data by automatic processing.

Good quality source data has to do with “Data Quality Culture” and must be initiated at the top of the organization. It is not just a matter of implementing strong validation checks on input screens, because almost no matter how strong these checks are, they can often still be circumvented by the users. There is a nine-step guide for organizations that wish to improve data quality:

- Declare a high level commitment to a data quality culture
- Drive process reengineering at the executive level
- Spend money to improve the data entry environment
- Spend money to improve application integration
- Spend money to change how processes work
- Promote end-to-end team awareness
- Promote interdepartmental cooperation
- Publicly celebrate data quality excellence
- Continuously measure and improve data quality

Decleanse

Parsing: for the detection of syntax errors. A parser decides whether a string of data is acceptable within the allowed data specification. This is similar to the way a parser works with grammars and languages.

- Data transformation: Data transformation allows the mapping of the data from its given format into the format expected by the appropriate application. This includes value conversions or translation functions, as well as normalizing numeric values to conform to minimum and maximum values.

- Duplicate elimination: Duplicate detection requires an algorithm for determining whether data contains duplicate representations of the same entity. Usually, data is sorted by a key that would bring duplicate entries closer together for faster identification.
- Statistical methods: By analyzing the data using the values of mean, standard deviation, range, or clustering algorithms, it is possible for an expert to find values that are unexpected and thus erroneous. Although the correction of such data is difficult since the true value is not known, it can be resolved by setting the values to an average or other statistical value. Statistical methods can also be used to handle missing values which can be replaced by one or more plausible values, which are usually obtained by extensive data augmentation algorithms.

Data Cleansing System

The essential job of this system is to find a suitable balance between fixing dirty data and maintaining the data as close as possible to the original data from the source production system. This is a challenge for the Extract, transform, load architect.

The system should offer an architecture that can cleanse data, record quality events and measure/control quality of data in the data warehouse.

A good start is to perform a thorough data profiling analysis that will help define to the required complexity of the data cleansing system and also give an idea of the current data quality in the source system(s).

Quality Screens

Part of the data cleansing system is a set of diagnostic filters known as quality screens. They each implement a test in the data flow that, if it fails records an error in the Error Event Schema. Quality screens are divided into three categories:

- Column screens. Testing the individual column, e.g. for unexpected values like NULL values; non-numeric values that should be numeric; out of range values; etc.
- Structure screens. These are used to test for the integrity of different relationships between columns (typically foreign/primary keys) in the same or different tables. They are also used for testing that a group of columns is valid according to some structural definition it should adhere.
- Business rule screens. The most complex of the three tests. They test to see if data, maybe across multiple tables, follow specific business rules. An example could be, that if a customer is marked as a certain type of customer, the business rules that define this kind of customer should be adhered.

When a quality screen records an error, it can either stop the dataflow process, send the faulty data somewhere else than the target system or tag the data. The latter option is considered the best solution because the first option requires, that someone has to manually deal with the issue each time it occurs and the second implies that data are missing from the target system (integrity) and it is often unclear, what should happen to these data.

Criticism of Existing Tools and Processes

The main reasons cited are:

- Project costs: costs typically in the hundreds of thousands of dollars
- Time: lack of enough time to deal with large-scale data-cleansing software
- Security: concerns over sharing information, giving an application access across systems, and effects on legacy systems

Error Event Schema

This schema is the place, where all error events thrown by quality screens, are recorded. It consists of an Error Event Fact table with foreign keys to three dimension tables that represent date (when), batch job (where) and screen (who produced error). It also holds information about exactly when the error occurred and the severity of the error. In addition there is an Error Event Detail Fact table with a foreign key to the main table that contains detailed information about in which table, record and field the error occurred and the error condition.

Challenges and Problems

- Error correction and loss of information: The most challenging problem within data cleansing remains the correction of values to remove duplicates and invalid entries. In many cases, the available information on such anomalies is limited and insufficient to determine the necessary transformations or corrections, leaving the deletion of such entries as a primary solution. The deletion of data, though, leads to loss of information; this loss can be particularly costly if there is a large amount of deleted data.
- Maintenance of cleansed data: Data cleansing is an expensive and time-consuming process. So after having performed data cleansing and achieving a data collection free of errors, one would want to avoid the re-cleansing of data in its entirety after some values in data collection change. The process should only be repeated on values that have changed; this means that a cleansing lineage would need to be kept, which would require efficient data collection and management techniques.
- Data cleansing in virtually integrated environments: In virtually integrated sources like IBM’s DiscoveryLink, the cleansing of data has to be performed every time the data is accessed, which considerably increases the response time and lowers efficiency.
- Data-cleansing framework: In many cases, it will not be possible to derive a complete data-cleansing graph to guide the process in advance. This makes data cleansing an iterative process involving significant exploration and interaction, which may require a framework in the form of a collection of methods for error detection and elimination in addition to data auditing. This can be integrated with other data-processing stages like integration and maintenance.

Process Mining

Process mining is a process management technique that allows for the analysis of business processes based on event logs. During process mining, specialized data-mining algorithms are applied to event log datasets in order to identify trends, patterns and details contained in event logs recorded by an information system. Process mining aims to improve process efficiency and understanding of processes. Process mining is also known as *Automated Business Process Discovery* (ABPD).

Overview

Process mining techniques are often used when no formal description of the process can be obtained by other approaches, or when the quality of existing documentation is questionable. For example, application of process mining methodology to the audit trails of a workflow management system, the transaction logs of an enterprise resource planning system, or the electronic patient records in a hospital can result in models describing processes, organizations, and products. Event log analysis can also be used to compare event logs with *prior* model(s) to understand whether the observations conform to a prescriptive or descriptive model.

Contemporary management trends such as BAM (Business Activity Monitoring), BOM (Business Operations Management), and BPI (business process intelligence) illustrate the interest in supporting diagnosis functionality in the context of Business Process Management technology (e.g., Workflow Management Systems and other *process-aware* information systems).

Application

Process mining follows the options established in business process engineering, then goes beyond those options by providing feedback for business process modeling:

- process analysis filters, orders and compresses logfiles for further insight into the connex of process operations.
- process design may be supported by feedback from process monitoring (action or event recording or logging)
- process enactment uses results from process mining based on logging for triggering further process operations

Classification

There are three classes of process mining techniques. This classification is based on whether there is a prior model and, if so, how the prior model is used during process mining.

- *Discovery*: Previous (*a priori*) models do not exist. Based on an event log, a new model is constructed or discovered based on low-level events. For example, using the alpha algorithm (a didactically driven approach). Many established techniques exist for automatically constructing process models (for example, Petri net, pi-calculus expression) based on an event log. Recently, process mining research has started targeting the other perspectives (e.g., data, resources, time, etc.). One example is the technique described in (Aalst, Reijers, & Song, 2005), which can be used to construct a social network.

- *Conformance checking*: Used when there is an *a priori* model. The existing model is compared with the process event log; discrepancies between the log and the model are analyzed. For example, there may be a process model indicating that purchase orders of more than 1 million Euro require two checks. Another example is the checking of the so-called "four-eyes" principle. Conformance checking may be used to detect deviations to enrich the model. An example is the extension of a process model with performance data, i.e., some *a priori* process model is used to project the potential bottlenecks. Another example is the *decision miner* described in (Rozinat & Aalst, 2006b) which takes an *a priori* process model and analyzes every choice in the process model. For each choice the event log is consulted to see which information is typically available the moment the choice is made. Then classical data mining techniques are used to see which data elements influence the choice. As a result, a decision tree is generated for each choice in the process.
- *Extension*: Used when there is an *a priori* model. The model is extended with a new aspect or perspective, so that the goal is *not* to check conformance, but rather to improve the existing model. An example is the extension of a process model with performance data, i.e., some prior process model dynamically annotated with performance data.

Software for Process Mining

A software framework for the evaluation of process mining algorithms has been developed at the Eindhoven University of Technology by Wil van der Aalst and others, and is available as an open source toolkit.

- Process Mining
- ProM Framework
- ProM Import Framework

Process Mining functionality is also offered by the following commercial vendors:

- Interstage Automated Process Discovery, a Process Mining service offered by Fujitsu, Ltd. as part of the Interstage Integration Middleware Suite.
- Disco is a complete Process Mining software by Fluxicon.
- ARIS Process Performance Manager, a Process Mining and Process Intelligence Tool offered by Software AG as part of the Process Intelligence Solution.
- QPR ProcessAnalyzer, Process Mining software for Automated Business Process Discovery (ABPD).
- Perceptive Process Mining, the Process Mining solution by Perceptive Software (formerly Futura Reflect / Pallas Athena Reflect).
- Celonis Process Mining, the Process Mining solution offered by Celonis
- SNP Business Process Analysis, the SAP-focused Process Mining solution by SNP Schnei-

der-Neureither & Partner AG

- minit is a Process Mining software offered by Gradient ECM
- myInvenio cloud and on-premises solution by Cognitive Technology Ltd.
- LANA is a process mining tool featuring discovery and conformance checking.
- ProcessGold Enterprise Platform, an integration of Process Mining & Business Intelligence.

Competitive Intelligence

Competitive intelligence is the action of defining, gathering, analyzing, and distributing intelligence about products, customers, competitors, and any aspect of the environment needed to support executives and managers making strategic decisions for an organization.

Competitive intelligence essentially means understanding and learning what is happening in the world outside the business so one can be as competitive as possible. It means learning as much as possible, as soon as possible, about one's industry in general, one's competitors, or even one's county's particular zoning rules. In short, it empowers anticipating and facing challenges head on.

Key points of this definition:

1. Competitive intelligence is an ethical and legal business practice, as opposed to industrial espionage, which is illegal.
2. The focus is on the external business environment
3. There is a process involved in gathering information, converting it into intelligence and then using it in decision making. Some CI professionals erroneously emphasise that if the intelligence gathered is not usable or actionable, it is not intelligence.

A more focused definition of CI regards it as the organizational function responsible for the early identification of risks and opportunities in the market before they become *obvious*. Experts also call this process the early signal analysis. This definition focuses attention on the difference between dissemination of widely available factual information (such as market statistics, financial reports, newspaper clippings) performed by functions such as libraries and information centers, and competitive intelligence which is a *perspective* on developments and events aimed at yielding a competitive edge.

The term CI is often viewed as synonymous with competitor analysis, but competitive intelligence is more than analyzing competitors: it is about making the organization more competitive relative to its entire environment and stakeholders: customers, competitors, distributors, technologies, and macroeconomic data.

Historic Development

The literature associated with the field of competitive intelligence is best exemplified by the detailed bibliographies that were published in the Society of Competitive Intelligence Professionals' refereed academic journal called *The Journal of Competitive Intelligence and Management*. Although elements of organizational intelligence collection have been a part of business for many years, the history of competitive intelligence arguably began in the U.S. in the 1970s, although the literature on the field pre-dates this time by at least several decades. In 1980, Michael Porter published the study *Competitive-Strategy: Techniques for Analyzing Industries and Competitors* which is widely viewed as the foundation of modern competitive intelligence. This has since been extended most notably by the pair of Craig Fleisher and Babette Bensoussan, who through several popular books on competitive analysis have added 48 commonly applied competitive intelligence analysis techniques to the practitioner's tool box. In 1985, Leonard Fuld published his best selling book dedicated to competitor intelligence. However, the institutionalization of CI as a formal activity among American corporations can be traced to 1988, when Ben and Tamar Gilad published the first organizational model of a formal corporate CI function, which was then adopted widely by US companies. The first professional certification program (CIP) was created in 1996 with the establishment of The Fuld-Gilad-Herring Academy of Competitive Intelligence in Cambridge, Massachusetts.

In 1986 the *Society of Competitive Intelligence Professionals* (SCIP) was founded in the United States and grew in the late 1990s to around 6,000 members worldwide, mainly in the United States and Canada, but with large numbers especially in the UK and Australia. Due to financial difficulties in 2009, the organization merged with Frost & Sullivan under the Frost & Sullivan Institute. SCIP has since been renamed "Strategic & Competitive Intelligence Professionals" to emphasise the strategic nature of the subject, and also to refocus the organisation's general approach, while keeping the existing SCIP brandname and logo. A number of efforts have been made to discuss the field's advances in post-secondary (university) education, covered by several authors including Blenkhorn & Fleisher, Fleisher, Fuld, Prescott, and McGonagle. Although the general view would be that competitive intelligence concepts can be readily found and taught in many business schools around the globe, there are still relatively few dedicated academic programs, majors, or degrees in the field, a concern to academics in the field who would like to see it further researched. These issues were widely discussed by over a dozen knowledgeable individuals in a special edition of the Competitive Intelligence Magazine that was dedicated to this topic. In France, a Specialized Master in Economic Intelligence and Knowledge Management was created in 1995 within the CERAM Business School, now SKEMA Business School, in Paris, with the objective of delivering a full and professional training in Economic Intelligence. A Centre for Global Intelligence and Influence was created in September 2011 in the same School.

On the other hand, practitioners and companies regard professional accreditation as especially important in this field. In 2011, SCIP recognized the Fuld-Gilad-Herring Academy of Competitive Intelligence's CIP certification process as its global, dual-level (CIP-I and CIP-II) certification program.

Global developments have also been uneven in competitive intelligence. Several academic journals, particularly the *Journal of Competitive Intelligence and Management* in its third volume, provided coverage of the field's global development. For example, in 1997 the *École de guerre économique* (fr) (School of economic warfare*) was founded in Paris, France. It is the first European institution which teaches the tactics of economic warfare within a globalizing world. In Germany, competitive intelligence was unattended until the early 1990s. The term "competitive intelligence" first appeared in German literature in 1997. In 1995 a German SCIP chapter was* *founded, which is now second in terms of membership in Europe. In summer 2004 the Institute for Competitive Intelligence was founded, which provides a postgraduate certification program for Competitive Intelligence Professionals. Japan is currently the only country that officially maintains an economic intelligence agency (JETRO). It was founded by the Ministry of International Trade and Industry (MITI) in 1958.*

Accepting the importance of competitive intelligence, major multinational corporations, such as ExxonMobil, Procter & Gamble, and Johnson and Johnson, have created formal CI units. Importantly, organizations execute competitive intelligence activities not only as a safeguard to protect against market threats and changes, but also as a method for finding new opportunities and trends.

Organizations use competitive intelligence to compare themselves to other organizations ("competitive benchmarking"), to identify risks and opportunities in their markets, and to pressure-test their plans against market response (war gaming), which enable them to make informed decisions. Most firms today realize the importance of knowing what their competitors are doing and how the industry is changing, and the information gathered allows organizations to understand their strengths and weaknesses.

One of the major activities involved in corporate competitive intelligence is use of ratio analysis, using key performance indicators (KPI). Organizations compare annual reports of their competitors on certain KPI and ratios, which are intrinsic to their industry. This helps them track their performance, vis-a-vis their competitors.

The actual importance of these categories of information to an organization depends on the contestability of its markets, the organizational culture, the personality and biases of its top decision makers, and the reporting structure of competitive intelligence within the company.

Strategic Intelligence (SI) focuses on the longer term, looking at issues affecting a company's competitiveness over the course of a couple of years. The actual time horizon for SI ultimately depends on the industry and how quickly it's changing. The general questions that SI answers are, 'Where should we as a company be in X years?' and 'What are the strategic risks and opportunities facing us?' This type of intelligence work involves among others the identification of weak signals and application of methodology and process called Strategic Early Warning (SEW), first introduced by Gilad, followed by Steven Shaker and Victor Richardson, Alessandro Comai and Joaquin Tena, and others. According to Gilad, 20% of the work of competitive intelligence practitioners should be dedicated to strategic early identification of weak signals within a SEW framework.

Tactical Intelligence: the focus is on providing information designed to improve shorter-term decisions, most often related with the intent of growing market share or revenues. Generally, it is the type of information that you would need to support the sales process in an organization. It investigates various aspects of a product/product line marketing:

- Product – what are people selling?
- Price – what price are they charging?
- Promotion – what activities are they conducting for promoting this product?
- Place – where are they selling this product?

- Other – sales force structure, clinical trial design, technical issues, etc.

With the right amount of information, organizations can avoid unpleasant surprises by anticipating competitors' moves and decreasing response time. Examples of competitive intelligence research is evident in daily newspapers, such as the *Wall Street Journal*, *Business Week*, and *Fortune*. Major airlines change hundreds of fares daily in response to competitors' tactics. They use information to plan their own marketing, pricing, and production strategies.

Resources, such as the Internet, have made gathering information on competitors easy. With a click of a button, analysts can discover future trends and market requirements. However competitive intelligence is much more than this, as the ultimate aim is to lead to competitive advantage. As the Internet is mostly public domain material, information gathered is less likely to result in insights that will be unique to the company. In fact there is a risk that information gathered from the Internet will be misinformation and mislead users, so competitive intelligence researchers are often wary of using such information.

As a result, although the Internet is viewed as a key source, most CI professionals should spend their time and budget gathering intelligence using primary research—networking with industry experts, from trade shows and conferences, from their own customers and suppliers, and so on. Where the Internet is used, it is to gather sources for primary research as well as information on what the company says about itself and its online presence (in the form of links to other companies, its strategy regarding search engines and online advertising, mentions in discussion forums and on blogs, etc.). Also, important are online subscription databases and news aggregation sources which have simplified the secondary source collection process. Social media sources are also becoming important—providing potential interviewee names, as well as opinions and attitudes, and sometimes breaking news (e.g., via Twitter).

Organizations must be careful not to spend too much time and effort on old competitors without realizing the existence of any new competitors. Knowing more about your competitors will allow your business to grow and succeed. The practice of competitive intelligence is growing every year, and most companies and business students now realize the importance of knowing their competitors.

According to Arjan Singh and Andrew Beurschgens in their 2006 article in the *Competitive Intelligence Review*, there are four stages of development of a competitive intelligence capability with a firm. It starts with "stick fetching", where a CI department is very reactive, up to "world class", where it is completely integrated in the decision-making process.

Recent Trends

The technical advances in massive parallel processing offered by the Hadoop "big data" architecture has allowed the creation of multiple platforms for named-entity recognition such as the Apache Projects OpenNLP and Apache Stanbol. The former includes pre-trained statistical parsers that can discern elements key to establishing trends and evaluating competitive position and responding appropriately. Public information mining from *SEC.gov*, Federal Contract Awards, social media (Twitter, Reddit, Facebook, and others), vendors, and competitor websites now permit real-time counterintelligence as a strategy for horizontal and vertical market expansion and product positioning. This occurs in an automated fashion on massive marketplaces such as Amazon.com and their classification and prediction of product associations and purchase probability.

Similar Fields

Competitive intelligence has been influenced by national strategic intelligence. Although national intelligence was researched 50 years ago, competitive intelligence was introduced during the 1990s. Competitive-intelligence professionals can learn from national-intelligence experts, especially in the analysis of complex situations. Competitive intelligence may be confused with (or seen to overlap) environmental scanning, business intelligence and market research. Craig Fleisher questions the appropriateness of the term, comparing it to business intelligence, competitor intelligence, knowledge management, market intelligence, marketing research and strategic intelligence.

Fleisher suggests that business intelligence has two forms. Its narrow (contemporary) form is more focused on information technology and internal focus than CI, while its broader (historical) definition is more inclusive than CI. Knowledge management (KM), when improperly achieved, is seen as an information-technology driven organizational practice relying on data mining, corporate intranets and mapping organizational assets to make it accessible to organization members for decision-making. CI shares some aspects of KM; they are human-intelligence- and experience-based for a more-sophisticated qualitative analysis. KM is essential for effective change. A key effective factor is a powerful, dedicated IT system executing the full intelligence cycle.

Market intelligence (MI) is industry-targeted intelligence developed in real-time aspects of competitive events taking place among the four Ps of the marketing mix (pricing, place, promotion and product) in the product (or service) marketplace to better understand the market's attractiveness. A time-based competitive tactic, MI is used by marketing and sales managers to respond to consumers more quickly in the marketplace. Fleisher suggests it is not distributed as widely as some forms of CI, which are also distributed to non-marketing decision-makers. Market intelligence has a shorter time horizon than other intelligence areas, and is measured in days, weeks, or (in slower-moving industries) months.

Market research is a tactical, method-driven field consisting of neutral, primary research of customer data (beliefs and perceptions) gathered in surveys or focus groups, and is analyzed with statistical-research techniques. CI draws on a wider variety (primary and secondary) of sources from a wider range of stakeholders (suppliers, competitors, distributors, substitutes and media) to answer existing questions, raise new ones and guide action.

Ben Gilad and Jan Herring lay down a set of prerequisites defining CI, distinguishing it from other information-rich disciplines such as market research or business development. They show that a common body of knowledge and a unique set of tools (key intelligence topics, business war games and blindspots analysis) distinguish CI; while other sensory activities in a commercial firm focus on *one* segment of the market (customers, suppliers or acquisition targets), CI synthesizes data from all high-impact players (HIP).

Gilad later focused his delineation of CI on the difference between information and intelligence. According to him, the common denominator among organizational sensory functions (whether they are called market research, business intelligence or market intelligence) is that they deliver information rather than intelligence. Intelligence, says Gilad, is a perspective on facts rather than the facts themselves. Unique among corporate functions, competitive intelligence has a perspective of risks and opportunities for a firm's performance; as such, it (not information activities) is part of an organization's risk-management activity.

Ethics

Ethics has been a long-held issue of discussion among CI practitioners. Essentially, the questions revolve around what is and is not allowable in terms of CI practitioners' activity. A number of excellent scholarly treatments have been generated on this topic, most prominently addressed through Society of Competitive Intelligence Professionals publications. The book *Competitive Intelligence Ethics: Navigating the Gray Zone* provides nearly twenty separate views about ethics in CI, as well as another 10 codes used by various individuals or organizations. Combining that with the over two dozen scholarly articles or studies found within the various CI bibliographic entries, it is clear that no shortage of study has gone into better classifying, understanding and addressing CI ethics.

Competitive information may be obtained from public or subscription sources, from networking with competitor staff or customers, disassembly of competitor products or from field research interviews. Competitive intelligence research is distinguishable from industrial espionage, as CI practitioners generally abide by local legal guidelines and ethical business norms.

Outsourcing

Outsourcing has become a big business for competitive intelligence professionals. Companies like Aperio Intelligence, Cast Intelligence, Fuld & Co, Black Cube, Securitas and Emissary Investigative Services offer corporate intelligence services.

References

- Ward, J. & Peppard, J. (2002). Situation Analysis. In Strategic Planning for information systems. (pp. 82 - 83). England: John Wiley & Sons. ISBN 978-0-470-84147-1
- D'Atri A., De Marco M., Casalino N. (2008). "Interdisciplinary Aspects of Information Systems Studies", Physica-Verlag, Springer, Germany, pp. 1–416, doi:10.1007/978-3-7908-2010-2 ISBN 978-3-7908-2009-6
- Verna Allee, The Knowledge Evolution: Expanding Organizational Intelligence, Butterworth-Heinemann (1997) ISBN 0-7506-9842-X
- David Perkins, King Arthur's Round Table: How Collaborative Conversations Create Smart Organizations, Wiley (2002) ISBN 0-4712-3772-8
- Tufte, Edward (1983). The Visual Display of Quantitative Information. Cheshire, Connecticut: Graphics Press. ISBN 0-9613921-4-2.
- [David Loshin (2003), "Business Intelligence: The Savvy Manager's Guide, Getting Onboard with Emerging IT", Morgan Kaufmann Publishers, ISBN 9781558609167].
- Kimball, R., Ross, M., Thornthwaite, W., Mundy, J., Becker, B. The Data Warehouse Lifecycle Toolkit, Wiley Publishing, Inc., 2008. ISBN 978-0-470-14977-5
- McGonagle, John J. and Carolyn M. Vella (2003). The Manager's Guide to Competitive Intelligence. Westport CT: Greenwood Publishing Group. p. 184. ISBN 1567205712.

7

Operational Intelligence: Technological Components

Operational intelligence has a number of aspects that have been elucidated in this chapter. Some of these features are complex event processing, business process management, metadata and root cause analysis. The components discussed in this text are of great importance to broaden the existing knowledge on operational intelligence.

Operational Intelligence

Operational intelligence (OI) is a category of real-time dynamic, business analytics that delivers visibility and insight into data, streaming events and business operations. Operational Intelligence solutions run queries against streaming data feeds and event data to deliver real-time analytic results as operational instructions. Operational Intelligence provides organizations the ability to make decisions and immediately act on these analytic insights, through manual or automated actions.

Purpose

The purpose of OI is to monitor business activities and identify and detect situations relating to inefficiencies, opportunities, and threats and provide operational solutions. Some definitions define operational intelligence an event-centric approach to delivering information that empowers people to make better decisions.

In addition, these metrics act as the starting point for further analysis (drilling down into details, performing root cause analysis — tying anomalies to specific transactions and of the business activity).

Sophisticated OI systems also provide the ability to associate metadata with metrics, process steps, channels, etc. With this, it becomes easy to get related information, e.g., “retrieve the contact information of the person that manages the application that executed the step in the business transaction that took 60% more time than the norm,” or “view the acceptance/rejection trend for the customer who was denied approval in this transaction,” or “Launch the application that this process step interacted with.”

Features

Different operational intelligence solutions may use many different technologies and be implemented in different ways. This section lists the common features of an operational intelligence solution:

- Real-time monitoring
- Real-time situation detection
- Real-time dashboards for different user roles
- Correlation of events
- Industry-specific dashboards
- Multidimensional analysis
    - Root cause analysis
    - Time Series and trend analysis
- Big Data Analytics: Operational Intelligence is well suited to address the inherent challenges of Big Data. Operational Intelligence continuously monitors and analyzes the variety of high velocity, high volume Big Data sources. Often performed in memory, OI platforms and solutions then present the incremental calculations and changes, in real-time, to the end-user.

Technology Components

Operational intelligence solutions share many features, and therefore many also share technology components. This is a list of some of the commonly found technology components, and the features they enable:

- Business activity monitoring (BAM) - Dashboard customization and personalization
- Complex event processing (CEP) - Advanced, continuous analysis of real-time information and historical data
- Business process management (BPM) - To perform model-driven execution of policies and processes defined as Business Process Model and Notation (BPMN) models
- Metadata framework to model and link events to resources
- Multi-channel publishing and notification
- Dimensional database
- Root cause analysis
- Multi-protocol event collection

Operational intelligence is a relatively new market segment (compared to the more mature business intelligence and business process management segments). In addition to companies that produce dedicated and focussed products in this area, there are numerous companies in adjacent areas that provide solutions with some OI components.

Operational intelligence places complete information at one’s fingertips, enabling one to make smarter decisions in time to maximize impact. By correlating a wide variety of events and data from both streaming feeds and historical data silos, operational intelligence helps organizations gain real-time visibility of information, in context, through advanced dashboards, real-time insight into business performance, health and status so that immediate action based on business policies and processes can be taken. Operational intelligence applies the benefits of real-time analytics, alerts, and actions to a broad spectrum of use cases across and beyond the enterprise.

One specific technology segment is AIDC (Automatic Identification and Data Capture) represented by barcodes, RFID and voice recognition.

Comparison with Other Technologies or Solutions

Business Intelligence

OI is often linked to or compared with business intelligence (BI) or real time business intelligence, in the sense that both help make sense out of large amounts of information. But there are some basic differences: OI is primarily activity-centric, whereas BI is primarily data-centric. As with most technologies, each of these could be sub-optimally coerced to perform the other's task. OI is, by definition, real-time, unlike BI or "On-Demand" BI, which are traditionally after-the-fact and report-based approaches to identifying patterns. Real-time BI (i.e., On-Demand BI) relies on the database as the sole source of events.

OI provides continuous, real-time analytics on data at rest and data in-flight, whereas BI typically looks only at historical data at rest. OI and BI can be complementary. OI is best used for short-term planning, such as deciding on the "next best action," while BI is best used for longer-term planning (over the next days to weeks). BI requires a more reactive approach, often reacting to events that have already taken place.

If all that is needed is a glimpse at historical performance over a very specific period of time, existing BI solutions should meet the requirement. However, historical data needs to be analyzed with events that are happening now, or to reduce the time between when intelligence is received and when action is taken, then Operational Intelligence is the more appropriate approach.

Systems Management

System Management mainly refers to the availability and capability monitoring of IT infrastructure. Availability monitoring refers to monitoring the status of IT infrastructure components such as servers, routers, networks, etc. This usually entails pinging or polling the component and waiting to receive a response. Capability monitoring usually refers to synthetic transactions where user activity is mimicked by a special software program, and the responses received are checked for correctness.

Complex Event Processing

There is a strong relationship between complex event processing companies and operational intelligence, especially since CEP is regarded by many OI companies as a core component of their OI solutions. CEP companies tend to focus solely on development of a CEP framework for other companies to use within their organisations as a pure CEP engine.

Business Activity Monitoring

Business activity monitoring (BAM) is software that aids in monitoring of business processes, as those processes are implemented in computer systems. BAM is an enterprise solution primarily intended to provide a real-time summary of business processes to operations managers and upper management. The main difference between BAM and OI appears to be in the implementation details — real-time situation detection appears in BAM and OI and is often implemented using CEP. Furthermore, BAM focuses on high-level process models whereas OI instead relies on correlation to infer a relationship between different events.

Business Process Management

A business process management suite is the runtime environment where one can perform model-driven execution of policies and processes defined as BPMN models. As part of an operational intelligence suite, a BPM suite can provide the capability to define and manage policies across the enterprise, apply the policies to events, and then take action according to the predefined policies. A BPM suite also provides the capability to define policies as if/then statements and apply them to events.

Business Activity Monitoring

Business activity monitoring (BAM) is software that aids in monitoring of business activities, as those activities are implemented in computer systems.

The term was originally coined by analysts at Gartner, Inc. and refers to the aggregation, analysis, and presentation of real-time information about activities inside organizations and involving customers and partners. A business activity can either be a business process that is orchestrated by business process management (BPM) software, or a business process that is a series of activities spanning multiple systems and applications. BAM is an enterprise solution primarily intended to provide a real-time summary of business activities to operations managers and upper management.

Goals and Benefits

The goals of business activity monitoring is to provide real time information about the status and results of various operations, processes, and transactions. The main benefits of BAM are to enable an enterprise to make better informed business decisions, quickly address problem areas, and re-position organizations to take full advantage of emerging opportunities.

Key Features

One of the most visible features of BAM solutions is the presentation of information on dashboards containing the key performance indicators (KPIs) used to provide assurance and visibility of activity and performance. This information is used by technical and business operations to provide visibility, measurement, and assurance of key business activities. It is also exploited by event correlation to detect and warn of impending problems.

Although BAM systems usually use a computer dashboard display to present data, BAM is distinct from the dashboards used by business intelligence (BI) insofar as events are processed in real-time or near real-time and pushed to the dashboard in BAM systems, whereas BI dashboards refresh at predetermined intervals by polling or querying databases. Depending on the refresh interval selected, BAM and BI dashboards can be similar or vary considerably.

Some BAM solutions additionally provide trouble notification functions, which allows them to interact automatically with the issue tracking system. For example, whole groups of people can be sent e-mails, voice or text messages, according to the nature of the problem. Automated problem solving, where feasible, can correct and restart failed processes.

Deployment Effort

In nearly all BAM deployments extensive tailoring to specific enterprises is required. Many BAM solutions seek to reduce extensive customization and may offer templates that are written to solve common problems in specific sectors, for example banking, manufacturing, and stockbroking. Due to the high degree of system integration required for initial deployment, many enterprises use experts that specialize in BAM to implement solutions.

BAM is now considered a critical component of Operational Intelligence (OI) solutions to deliver visibility into business operations. Multiple sources of data can be combined from different organizational silos to provide a common operating picture that uses current information. Wherever real-time insight has the greatest value, OI solutions can be applied to deliver the needed information.

Processing Events

All BAM solutions process events. While most of the first BAM solutions were closely linked to BPM solutions and therefore processed events emitted as the process was being orchestrated, this had the disadvantage of requiring enterprises to invest in BPM before being able to acquire and use BAM. The newer generation of BAM solutions are based on complex event processing (CEP) technology, and can process high volumes of underlying technical events to derive higher level business events, therefore reducing the dependency on BPM, and providing BAM to a wider audience of customers.

Examples

A bank might be interested in minimizing the amount of money it borrows overnight from a central bank. Interbank transfers must be communicated and arranged through automation by a set time each business day. The failure of any vital communication could cost the bank large sums in interest charged by the central bank. A BAM solution would be programmed to become aware of each message and await confirmation. Failure to receive confirmation within a reasonable amount of time would be grounds for the BAM solution to raise an alarm that would set in motion manual intervention to investigate the cause of the delay and to push the problem toward resolution before it becomes costly.

Another example involves a mobile telecommunications company interested in detecting a situation whereby new customers are not set up promptly and correctly on their network and within the various CRM and billing solutions. Low-level technical events such as messages passing from one application to another over a middleware system, or transactions detected within a database logfile, are processed by a CEP engine. All events relating to an individual customer are correlated in order to detect an anomalous situation whereby an individual customer has not been promptly or correctly provisioned, or set up. An alert can be generated to notify technical operations or to notify business operations, and the failed provisioning step may be restarted automatically.

Complex Event Processing

Event processing is a method of tracking and analyzing (processing) streams of information (data) about things that happen (events), and deriving a conclusion from them. Complex event processing, or CEP, is event processing that combines data from multiple sources to infer events or patterns that suggest more complicated circumstances. The goal of complex event processing is to identify meaningful events (such as opportunities or threats) and respond to them as quickly as possible.

These events may be happening across the various layers of an organization as sales leads, orders or customer service calls. Or, they may be news items, text messages, social media posts, stock market feeds, traffic reports, weather reports, or other kinds of data. An event may also be defined as a “change of state,” when a measurement exceeds a predefined threshold of time, temperature, or other value. Analysts suggest that CEP will give organizations a new way to analyze patterns in real-time and help the business side communicate better with IT and service departments.

The vast amount of information available about events is sometimes referred to as the event cloud.

Conceptual Description

Among thousands of incoming events, a monitoring system may for instance receive the following three from the same source:

- church bells ringing.
- the appearance of a man in a tuxedo with a woman in a flowing white gown.
- rice flying through the air.

From these events the monitoring system may infer a *complex event*: a wedding. CEP as a technique helps discover complex events by analyzing and correlating other events: the bells, the man and woman in wedding attire and the rice flying through the air.

CEP relies on a number of techniques, including:

- Event-pattern detection
- Event abstraction
- Event filtering
- Event aggregation and transformation

- Modeling event hierarchies
- Detecting relationships (such as causality, membership or timing) between events
- Abstracting event-driven processes

Commercial applications of CEP exist in variety of industries and include algorithmic stock-trading, the detection of credit-card fraud, business activity monitoring, and security monitoring.

History

The CEP area has roots in discrete event simulation, the active database area and some programming languages. The activity in the industry was preceded by a wave of research projects in the 1990s. According to the first project that paved the way to a generic CEP language and execution model was the Rapide project in Stanford University, directed by David Luckham. In parallel there have been two other research projects: Infospheres in California Institute of Technology, directed by K. Mani Chandy, and Apama in University of Cambridge directed by John Bates. The commercial products were dependents of the concepts developed in these and some later research projects. Community efforts started in a series of event processing symposiums organized by the Event Processing Technical Society, and later by the ACM DEBS conference series. One of the community efforts was to produce the event processing manifesto

Related Concepts

CEP is used in Operational Intelligence (OI) solutions to provide insight into business operations by running query analysis against live feeds and event data. OI solutions collect real-time data and correlate against historical data to provide insight into and analysis of the current situation. Multiple sources of data can be combined from different organizational silos to provide a common operating picture that uses current information. Wherever real-time insight has the greatest value, OI solutions can be applied to deliver the information needed.

In network management, systems management, application management and service management, people usually refer instead to event correlation. As CEP engines, event correlation engines (*event correlators*) analyze a mass of events, pinpoint the most significant ones, and trigger actions. However, most of them do not produce new inferred events. Instead, they relate high-level events with low-level events.

Inference engines, e.g. rule-based reasoning engines typically produce inferred information in artificial intelligence. However, they do not usually produce new information in the form of complex (i.e., inferred) events.

Example

A more systemic example of CEP involves a car, some sensors and various events and reactions. Imagine that a car has several sensors—one that measures tire pressure, one that measures speed, and one that detects if someone sits on a seat or leaves a seat.

In the first situation, the car is moving and the pressure of one of the tires moves from 45 psi to 41 psi over 15 minutes. As the pressure in the tire is decreasing, a series of events containing the tire pressure is generated. In addition, a series of events containing the speed of the car is generated. The car's Event Processor may detect a situation whereby a loss of tire pressure over a relatively long period of time results in the creation of the "lossOfTirePressure" event. This new event may trigger a reaction process to note the pressure loss into the car's maintenance log, and alert the driver via the car's portal that the tire pressure has reduced.

In the second situation, the car is moving and the pressure of one of the tires drops from 45 psi to 20 psi in 5 seconds. A different situation is detected—perhaps because the loss of pressure occurred over a shorter period of time, or perhaps because the difference in values between each event were larger than a predefined limit. The different situation results in a new event "blowOutTire" being generated. This new event triggers a different reaction process to immediately alert the driver and to initiate onboard computer routines to assist the driver in bringing the car to a stop without losing control through skidding.

In addition, events that represent detected situations can also be combined with other events in order to detect more complex situations. For example, in the final situation the car is moving normally and suffers a blown tire which results in the car leaving the road and striking a tree, and the driver is thrown from the car. A series of different situations are rapidly detected. The combination of "blowOutTire", "zeroSpeed" and "driverLeftSeat" within a very short period of time results in a new situation being detected: "occupantThrownAccident". Even though there is no direct measurement that can determine conclusively that the driver was thrown, or that there was an accident, the combination of events allows the situation to be detected and a new event to be created to signify the detected situation. This is the essence of a complex (or composite) event. It is complex because one cannot directly detect the situation; one has to infer or deduce that the situation has occurred from a combination of other events.

Types

Most CEP solutions and concepts can be classified into two main categories:

- Aggregation-oriented CEP
- Detection-oriented CEP

An aggregation-oriented CEP solution is focused on executing on-line algorithms as a response to event data entering the system. A simple example is to continuously calculate an average based on data in the inbound events.

Detection-oriented CEP is focused on detecting combinations of events called events patterns or situations. A simple example of detecting a situation is to look for a specific sequence of events.

There also exist hybrid approaches.

Integration with Business Process Management

A natural fit for CEP has been with Business Process Management, or BPM. BPM focuses on end-to-end business processes, in order to continuously optimize and align for its operational environment.

However, the optimization of a business does not rely solely upon its individual, end-to-end processes. Seemingly disparate processes can affect each other significantly. Consider this scenario: In the aerospace industry, it is good practice to monitor breakdowns of vehicles to look for trends (determine potential weaknesses in manufacturing processes, material, etc.). Another separate process monitors current operational vehicles' life cycles and decommissions them when appropriate. One use for CEP is to link these separate processes, so that in the case of the initial process (breakdown monitoring) discovering a malfunction based on metal fatigue (a significant event), an action can be created to exploit the second process (life cycle) to issue a recall on vehicles using the same batch of metal discovered as faulty in the initial process.

The integration of CEP and BPM must exist at two levels, both at the business awareness level (users must understand the potential holistic benefits of their individual processes) and also at the technological level (there needs to be a method by which CEP can interact with BPM implementation). For a recent state of the art review on the integration of CEP with BPM, which is frequently labeled as Event-Driven Business Process Management, refer to.

Computation-oriented CEP's role can arguably be seen to overlap with Business Rule technology.

For example, customer service centers are using CEP for click-stream analysis and customer experience management. CEP software can factor real-time information about millions of events (clicks or other interactions) per second into business intelligence and other decision-support applications. These "recommendation applications" help agents provide personalized service based on each customer's experience. The CEP application may collect data about what customers on the phone are currently doing, or how they have recently interacted with the company in other various channels, including in-branch, or on the Web via self-service features, instant messaging and email. The application then analyzes the total customer experience and recommends scripts or next steps that guide the agent on the phone, and hopefully keep the customer happy.

In Financial Services

The financial services industry was an early adopter of CEP technology, using complex event processing to structure and contextualize available data so that it could inform trading behavior, specifically algorithmic trading, by identifying opportunities or threats that indicate traders (or automatic trading systems) should buy or sell. For example, if a trader wants to track stocks that have five up movements followed by four down movements, CEP technology can track such an event. CEP technology can also track drastic rise and fall in number of trades. Algorithmic trading is already a practice in stock trading. It is estimated that around 60% of Equity trading in the United States is by way of algorithmic trades. CEP is expected to continue to help financial institutions improve their algorithms and be more efficient.

Recent improvements in CEP technologies have made it more affordable, helping smaller firms to create trading algorithms of their own and compete with larger firms. CEP has evolved from an emerging technology to an essential platform of many capital markets. The technology's most consistent growth has been in banking, serving fraud detection, online banking, and multichannel marketing initiatives.

Today, a wide variety of financial applications use CEP, including profit, loss, and risk management systems, order and liquidity analysis, quantitative trading and signal generation systems, and others.

Integration with time series databases

A time series database is a software system that is optimized for the handling of data organized by time. Time series are finite or infinite sequences of data items, where each item has an associated timestamp and the sequence of timestamps is non-decreasing. Elements of a time series are often called ticks. The timestamps are not required to be ascending (merely non-decreasing) because in practice the time resolution of some systems such as financial data sources can be quite low (milliseconds, microseconds or even nanoseconds), so consecutive events may carry equal timestamps.

Time series data provides a historical context to the analysis typically associated with complex event processing. This can apply to any vertical industry such as finance and cooperatively with other technologies such as BPM.

Consider the scenario in finance where there is a need to understand historic price volatility to determine statistical thresholds of future price movements. This is helpful for both trade models and transaction cost analysis.

The ideal case for CEP analysis is to view historical time series and real-time streaming data as a single time continuum. What happened yesterday, last week or last month is simply an extension of what is occurring today and what may occur in the future. An example may involve comparing current market volumes to historic volumes, prices and volatility for trade execution logic. Or the need to act upon live market prices may involve comparisons to benchmarks that include sector and index movements, whose intra-day and historic trends gauge volatility and smooth outliers.

Business Process Management

Business process management (BPM) is a field in operations management that focuses on improving corporate performance by managing and optimizing a company's business processes. It can therefore be described as a "process optimization process." It is argued that BPM enables organizations to be more efficient, more effective and more capable of change than a functionally focused, traditional hierarchical management approach. These processes can impact the cost and revenue generation of an organization.

As a policy-making approach, BPM sees processes as important assets of an organization that must be understood, managed, and developed to announce value-added products and services to clients or customers. This approach closely resembles other total quality management or continual improvement process methodologies and BPM proponents also claim that this approach can be supported, or enabled, through technology. As such, many BPM articles and scholars frequently discuss BPM from one of two viewpoints: people and/or technology.

Definitions

BPMInstitute.org defines Business Process Management as:

> the definition, improvement and management of a firm's end-to-end enterprise business processes in order to achieve three outcomes crucial to a performance-based, customer-driven firm: 1) clarity on strategic direction, 2) alignment of the firm's resources, and 3) increased discipline in daily operations.

The Workflow Management Coalition, BPM.com and several other sources have come to agreement on the following definition:

> Business Process Management (BPM) is a discipline involving any combination of modeling, automation, execution, control, measurement and optimization of business activity flows, in support of enterprise goals, spanning systems, employees, customers and partners within and beyond the enterprise boundaries.

The Association Of Business Process Management Professionals defines BPM as:

> Business Process Management (BPM) is a disciplined approach to identify, design, execute, document, measure, monitor, and control both automated and non-automated business processes to achieve consistent, targeted results aligned with an organization's strategic goals. BPM involves the deliberate, collaborative and increasingly technology-aided definition, improvement, innovation, and management of end-to-end business processes that drive business results, create value, and enable an organization to meet its business objectives with more agility. BPM enables an enterprise to align its business processes to its business strategy, leading to effective overall company performance through improvements of specific work activities either within a specific department, across the enterprise, or between organizations.

Gartner defines Business process management (BPM) as:

> "the discipline of managing processes (rather than tasks) as the means for improving business performance outcomes and operational agility. Processes span organizational boundaries, linking together people, information flows, systems and other assets to create and deliver value to customers and constituents."

It is common to confuse BPM with a BPM Suite (BPMS). BPM is a professional discipline done by people, whereas a BPMS is a technological suite of tools designed to help the BPM professionals accomplish their goals. BPM should also not be confused with an application or solution developed to support a particular process. Suites and solutions represent ways of automating business processes, but automation is only one aspect of BPM.

Changes in Business Process Management

The concept of business process may be as traditional as concepts of tasks, department, production, and outputs, arising from job shop scheduling problems in the early 20th Century. The management and improvement approach as of 2010, with formal definitions and technical modeling, has been around since the early 1990s. Note that the term “business process” is sometimes used by IT practitioners as synonymous with the management of middleware processes or with integrating application software tasks.

Although BPM initially focused on the automation of business processes with the use of information technology, it has since been extended to integrate human-driven processes in which human interaction takes place in series or parallel with the use of technology. For example, workflow management systems can assign individual steps requiring deploying human intuition or judgment to relevant humans and other tasks in a workflow to a relevant automated system.

More recent variations such as “human interaction management” are concerned with the interaction between human workers performing a task.

As of 2010 technology has allowed the coupling of BPM with other methodologies, such as Six Sigma. Some BPM tools such as SIPOCs, process flows, RACIs, CTQs and histograms allow users to:

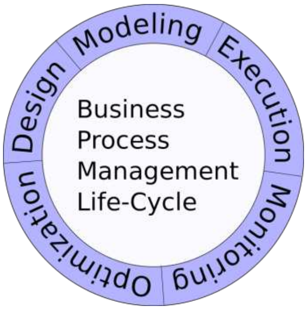
- visualize - functions and processes
- measure - determine the appropriate measure to determine success
- analyze - compare the various simulations to determine an optimal improvement
- improve - select and implement the improvement
- control - deploy this implementation and by use of user-defined dashboards monitor the improvement in real time and feed the performance information back into the simulation model in preparation for the next improvement iteration
- re-engineer - revamp the processes from scratch for better results

This brings with it the benefit of being able to simulate changes to business processes based on real-world data (not just on assumed knowledge). Also, the coupling of BPM to industry methodologies allows users to continually streamline and optimize the process to ensure that it is tuned to its market need.

As of 2012 research on BPM has paid increasing attention to the compliance of business processes. Although a key aspect of business processes is flexibility, as business processes continuously need to adapt to changes in the environment, compliance with business strategy, policies and government regulations should also be ensured. The compliance aspect in BPM is highly important for governmental organizations. As of 2010 BPM approaches in a governmental context largely focus on operational processes and knowledge representation. Although there have been many technical studies on operational business processes in both the public and private sectors, researchers have rarely taken legal compliance activities into account, for instance the legal implementation processes in public-administration bodies.

BPM Life-cycle

Business process management activities can be arbitrarily grouped into categories such as design, modeling, execution, monitoring, and optimization.



Design

Process design encompasses both the identification of existing processes and the design of "to-be" processes. Areas of focus include representation of the process flow, the factors within it, alerts and notifications, escalations, standard operating procedures, service level agreements, and task hand-over mechanisms.

Whether or not existing processes are considered, the aim of this step is to ensure that a correct and efficient theoretical design is prepared.

The proposed improvement could be in human-to-human, human-to-system or system-to-system workflows, and might target regulatory, market, or competitive challenges faced by the businesses.

The existing process and the design of new process for various applications will have to synchronise and not cause major outage or process interruption.

Modeling

Modeling takes the theoretical design and introduces combinations of variables (e.g., changes in rent or materials costs, which determine how the process might operate under different circumstances).

It may also involve running "what-if analysis"(Conditions-when, if, else) on the processes: *"What if I have 75% of resources to do the same task?" "What if I want to do the same job for 80% of the current cost?"*.

Execution

One of the ways to automate processes is to develop or purchase an application that executes the required steps of the process; however, in practice, these applications rarely execute all the steps of the process accurately or completely. Another approach is to use a combination of software and human intervention; however this approach is more complex, making the documentation process difficult.

As a response to these problems, software has been developed that enables the full business process (as developed in the process design activity) to be defined in a computer language which can be directly executed by the computer. The process models can be run through execution engines that automate the processes directly from the model *(e.g.* calculating a repayment plan for a loan) or, when a step is too complex to automate, Business Process Modeling Notation (BPMN) provides front-end capability for human input. Compared to either of the previous approaches, directly executing a process definition can be more straightforward and therefore easier to improve. However, automating a process definition requires flexible and comprehensive infrastructure, which typically rules out implementing these systems in a legacy IT environment.

Business rules have been used by systems to provide definitions for governing behavior, and a business rule engine can be used to drive process execution and resolution.

Monitoring

Monitoring encompasses the tracking of individual processes, so that information on their state can be easily seen, and statistics on the performance of one or more processes can be provided. An example of this tracking is being able to determine the state of a customer order *(e.g.* order arrived, awaiting delivery, invoice paid) so that problems in its operation can be identified and corrected.

In addition, this information can be used to work with customers and suppliers to improve their connected processes. Examples are the generation of measures on how quickly a customer order is processed or how many orders were processed in the last month. These measures tend to fit into three categories: cycle time, defect rate and productivity.

The degree of monitoring depends on what information the business wants to evaluate and analyze and how business wants it to be monitored, in real-time, near real-time or ad hoc. Here, business activity monitoring (BAM) extends and expands the monitoring tools generally provided by BPMS.

Process mining is a collection of methods and tools related to process monitoring. The aim of process mining is to analyze event logs extracted through process monitoring and to compare them with an *a priori* process model. Process mining allows process analysts to detect discrepancies between the actual process execution and the *a priori* model as well as to analyze bottlenecks.

Optimization

Process optimization includes retrieving process performance information from modeling or monitoring phase; identifying the potential or actual bottlenecks and the potential opportunities for cost savings or other improvements; and then, applying those enhancements in the design of the process. Process mining tools are able to discover critical activities and bottlenecks, creating greater business value.

Re-engineering

When the process becomes too complex or inefficient, and optimization is not fetching the desired output, it is usually recommended by a company steering committee chaired by the president / CEO to re-engineer the entire process cycle. Business process reengineering (BPR) has been used by organizations to attempt to achieve efficiency and productivity at work.

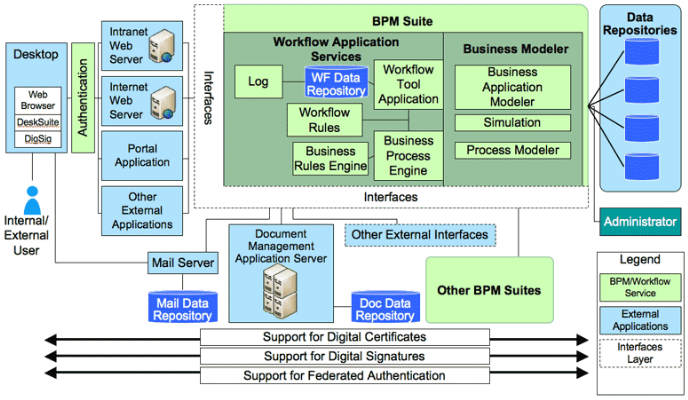
BPM Suites

A market has developed for Enterprise software leveraging the Business Process Management concepts to organize and automate processes. The recent convergence of these software from distinct pieces such as Business rules engine, Business Process Modelling, Business Activity Monitoring and Human Workflow has given birth to integrated Business Process Management Suites. Forrester Research, Inc recognize the BPM suite space through three different lenses:

- human-centric BPM
- integration-centric BPM (Enterprise Service Bus)
- document-centric BPM (Dynamic Case Management)

However, standalone integration-centric and document-centric offerings have matured into separate, standalone markets.

Practice


BPM Suite
Workflow Application Services
Log
WF Data Repository
Workflow Tool Application
Workflow Rules
Business Rules Engine
Business Process Engine
Business Modeler
Business Application Modeler
Simulation
Process Modeler
Data Repositories
Administrator
Interfaces
Desktop
Web Browser
DeskSuite
DigSig
Authentication
Intranet Web Server
Internet Web Server
Portal Application
Other External Applications
Internal/ External User
Mail Server
Mail Data Repository
Document Management Application Server
Doc Data Repository
Other External Interfaces
Other BPM Suites
Support for Digital Certificates
Support for Digital Signatures
Support for Federated Authentication
Legend
BPM/Workflow Service
External Applications
Interfaces Layer

Example of Business Process Management (BPM) Service Pattern: This pattern shows how business process management (BPM) tools can be used to implement business processes through the orchestration of activities between people and systems.

While the steps can be viewed as a cycle, economic or time constraints are likely to limit the process to only a few iterations. This is often the case when an organization uses the approach for short to medium term objectives rather than trying to transform the organizational culture. True iterations are only possible through the collaborative efforts of process participants. In a majority of organizations, complexity will require enabling technology to support the process participants in these daily process management challenges.

To date, many organizations often start a BPM project or program with the objective of optimizing an area that has been identified as an area for improvement.

Currently, the international standards for the task have limited BPM to the application in the IT sector, and ISO/IEC 15944 covers the operational aspects of the business. However, some corporations with the culture of best practices do use standard operating procedures to regulate their operational process. Other standards are currently being worked upon to assist in BPM implementation (BPMN, Enterprise Architecture, Business Motivation Model).

BPM Technology

BPM is now considered a critical component of operational intelligence (OI) solutions to deliver real-time, actionable information. This real-time information can be acted upon in a variety of ways - alerts can be sent or executive decisions can be made using real-time dashboards. OI solutions use real-time information to take automated action based on pre-defined rules so that security measures and or exception management processes can be initiated.

As such, some people view BPM as "the bridge between Information Technology (IT) and Business.". In fact, an argument can be made that this "holistic approach" bridges organizational and technological silos.

There are four critical components of a BPM Suite:

- Process engine — a robust platform for modeling and executing process-based applications, including business rules
- Business analytics — enable managers to identify business issues, trends, and opportunities with reports and dashboards and react accordingly
- Content management — provides a system for storing and securing electronic documents, images, and other files
- Collaboration tools — remove intra- and interdepartmental communication barriers through discussion forums, dynamic workspaces, and message boards

BPM also addresses many of the critical IT issues underpinning these business drivers, including:

- Managing end-to-end, customer-facing processes
- Consolidating data and increasing visibility into and access to associated data and information
- Increasing the flexibility and functionality of current infrastructure and data
- Integrating with existing systems and leveraging service oriented architecture (SOA)
- Establishing a common language for business-IT alignment

Validation of BPMS is another technical issue that vendors and users need to be aware of, if regulatory compliance is mandatory. The validation task could be performed either by an authenticated third party or by the users themselves. Either way, validation documentation will need to be generated. The validation document usually can either be published officially or retained by users.

Cloud Computing BPM

Cloud computing business process management is the use of (BPM) tools that are delivered as software services (SaaS) over a network. Cloud BPM business logic is deployed on an application server and the business data resides in cloud storage.

Market

According to Gartner, 20% of all the "shadow business processes" will be supported by BPM cloud platforms. Gartner refers to all the hidden organizational processes that are supported by IT departments as part of legacy business processes such as Excel spreadsheets, routing of emails using rules, phone calls routing, etc. These can, of course also be replaced by other technologies such as workflow software.

Benefits

The benefits of using cloud BPM services include removing the need and cost of maintaining specialized technical skill sets in-house and reducing distractions from an enterprise's main focus. It offers controlled IT budgeting and enables geographical mobility..

Internet of Things

The emerging Internet of Things poses a significant challenge to control and manage the flow of information through large numbers of devices. To cope with this, a new direction known as BPM Everywhere shows promise as way of blending traditional process techniques, with additional capabilities to automate the handling of all the independent devices.

Metadata

In the 2010s, metadata typically refers to digital forms; however, even traditional card catalogues from the 1960s and 1970s are an example of metadata, as the cards contain information about the books in the library (author, title, subject, etc.).

Metadata is "data [information] that provides information about other data". Three distinct types of metadata exist: structural metadata, descriptive metadata, and administrative metadata.

Structural metadata is data about the containers of data. For instance a "book" contains data, and data about the book is metadata about that container of data.

Descriptive metadata uses individual instances of application data or the data content.

In many countries, the metadata relating to emails, telephone calls, web pages, video traffic, IP connections and cell phone locations are routinely stored by government organizations.

History

Metadata was traditionally used in the card catalogs of libraries until the 1980s, when libraries converted their catalog data to digital databases. In the 2000s, as digital formats are becoming the prevalent way of storing data and information, metadata is also used to describe digital data using metadata standards.

There are different metadata standards for each different discipline (e.g., museum collections, digital audio files, websites, etc.). Describing the contents and context of data or data files increases its usefulness. For example, a web page may include metadata specifying what software language the page is written in (e.g., HTML), what tools were used to create it, what subjects the page is about, and where to find more information about the subject. This metadata can automatically improve the reader's experience and make it easier for users to find the web page online. A CD may include metadata providing information about the musicians, singers and songwriters whose work appears on the disc.

A principal purpose of metadata is to help users find relevant information and discover resources. Metadata also helps to organize electronic resources, provide digital identification, and support the archiving and preservation of resources. Metadata assists users in resource discovery by "allowing resources to be found by relevant criteria, identifying resources, bringing similar resources together, distinguishing dissimilar resources, and giving location information." Metadata of telecommunication activities including Internet traffic is very widely collected by various national governmental organizations. This data is used for the purposes of traffic analysis and can be used for mass surveillance.

Definition

Metadata means "data about data". Although the "meta" prefix means «after» or «beyond», it is used to mean «about» in epistemology. Metadata is defined as the data providing information about one or more aspects of the data; it is used to summarize basic information about data which can make tracking and working with specific data easier. Some examples include:

- Means of creation of the data
- Purpose of the data
- Time and date of creation
- Creator or author of the data
- Location on a computer network where the data was created
- Standards used
- File size

For example, a digital image may include metadata that describes how large the picture is, the color depth, the image resolution, when the image was created, the shutter speed, and other data. A text document's metadata may contain information about how long the document is, who the author is, when the document was written, and a short summary of the document. Metadata within web pages can also contain descriptions of page content, as well as key words linked to the content. These links are often called "Metatags", which were used as the primary factor in determining order for a web search until the late 1990s. The reliance of metatags in web searches was decreased in the late 1990s because of "keyword stuffing". Metatags were being largely misused to trick search engines into thinking some websites had more relevance in the search than they really did.

Metadata can be stored and managed in a database, often called a metadata registry or metadata repository. However, without context and a point of reference, it might be impossible to identify metadata just by looking at it. For example: by itself, a database containing several numbers, all 13 digits long could be the results of calculations or a list of numbers to plug into an equation - without any other context, the numbers themselves can be perceived as the data. But if given the context that this database is a log of a book collection, those 13-digit numbers may now be identified as ISBNs - information that refers to the book, but is not itself the information within the book. The term "metadata" was coined in 1968 by Philip Bagley, in his book "Extension of Programming Language Concepts" where it is clear that he uses the term in the ISO 11179 "traditional" sense, which is "structural metadata" i.e. "data about the containers of data"; rather than the alternate sense "content about individual instances of data content" or metacontent, the type of data usually found in library catalogues. Since then the fields of information management, information science, information technology, librarianship, and GIS have widely adopted the term. In these fields the word *metadata* is defined as "data about data". While this is the generally accepted definition, various disciplines have adopted their own more specific explanation and uses of the term.

Types

While the metadata application is manifold, covering a large variety of fields, there are specialized and well-accepted models to specify types of metadata. Bretherton & Singley (1994) distinguish between two distinct classes: structural/control metadata and guide metadata. *Structural metadata* describes the structure of database objects such as tables, columns, keys and indexes. *Guide metadata* helps humans find specific items and are usually expressed as a set of keywords in a natural language. According to Ralph Kimball metadata can be divided into 2 similar categories: technical metadata and business metadata. *Technical metadata* corresponds to internal metadata, and *business metadata* corresponds to external metadata. Kimball adds a third category, *process metadata*. On the other hand, NISO distinguishes among three types of metadata: descriptive, structural, and administrative.

Descriptive metadata is typically used for discovery and identification, as information to search and locate an object, such as title, author, subjects, keywords, publisher. *Structural metadata* describes how the components of an object are organized. An example of structural metadata would be how pages are ordered to form chapters of a book. Finally, *administrative metadata* gives information to help manage the source. Administrative metadata refers to the technical information, including file type, or when and how the file was created. Two sub-types of administrative metadata are rights management metadata and preservation metadata. *Rights management metadata* explains intellectual property rights, while *preservation metadata* contains information to preserve and save a resource.

Structures

Metadata (metacontent) or, more correctly, the vocabularies used to assemble metadata (metacontent) statements, is typically structured according to a standardized concept using a well-defined metadata scheme, including: metadata standards and metadata models. Tools such as controlled vocabularies, taxonomies, thesauri, data dictionaries, and metadata registries can be used to apply further standardization to the metadata. Structural metadata commonality is also of paramount importance in data model development and in database design.

Syntax

Metadata (metacontent) syntax refers to the rules created to structure the fields or elements of metadata (metacontent). A single metadata scheme may be expressed in a number of different markup or programming languages, each of which requires a different syntax. For example, Dublin Core may be expressed in plain text, HTML, XML, and RDF.

A common example of (guide) metacontent is the bibliographic classification, the subject, the Dewey Decimal class number. There is always an implied statement in any "classification" of some object. To classify an object as, for example, Dewey class number 514 (Topology) (i.e. books having the number 514 on their spine) the implied statement is: "<book><subject heading><514>. This is a subject-predicate-object triple, or more importantly, a class-attribute-value triple. The first two elements of the triple (class, attribute) are pieces of some structural metadata having a defined semantic. The third element is a value, preferably from some controlled vocabulary, some reference (master) data. The combination of the metadata and master data elements results in a statement which is a metacontent statement i.e. "metacontent = metadata + master data". All of these elements can be thought of as "vocabulary". Both metadata and master data are vocabularies which can be assembled into metacontent statements. There are many sources of these vocabularies, both meta and master data: UML, EDIFACT, XSD, Dewey/UDC/LoC, SKOS, ISO-25964, Pantone, Linnaean Binomial Nomenclature, etc. Using controlled vocabularies for the components of metacontent statements, whether for indexing or finding, is endorsed by ISO 25964: "If both the indexer and the searcher are guided to choose the same term for the same concept, then relevant documents will be retrieved." This is particularly relevant when considering search engines of the internet, such as Google. The process indexes pages then matches text strings using its complex algorithm; there is no intelligence or "inferencing" occurring, just the illusion thereof.

Hierarchical, Linear and Planar Schemata

Metadata schemata can be hierarchical in nature where relationships exist between metadata elements and elements are nested so that parent-child relationships exist between the elements. An example of a hierarchical metadata schema is the IEEE LOM schema, in which metadata elements may belong to a parent metadata element. Metadata schemata can also be one-dimensional, or linear, where each element is completely discrete from other elements and classified according to one dimension only. An example of a linear metadata schema is the Dublin Core schema, which is one dimensional. Metadata schemata are often two dimensional, or planar, where each element is completely discrete from other elements but classified according to two orthogonal dimensions.

Hypermapping

In all cases where the metadata schemata exceed the planar depiction, some type of hypermapping is required to enable display and view of metadata according to chosen aspect and to serve special views. Hypermapping frequently applies to layering of geographical and geological information overlays.

Granularity

The degree to which the data or metadata is structured is referred to as its “granularity”. “Granularity” refers to how much detail is provided. Metadata with a high granularity allows for deeper, more detailed, and more structured information and enables greater levels of technical manipulation. A lower level of granularity means that metadata can be created for considerably lower costs but will not provide as detailed information. The major impact of granularity is not only on creation and capture, but moreover on maintenance costs. As soon as the metadata structures become outdated, so too is the access to the referred data. Hence granularity must take into account the effort to create the metadata as well as the effort to maintain it.

Standards

International standards apply to metadata. Much work is being accomplished in the national and international standards communities, especially ANSI (American National Standards Institute) and ISO (International Organization for Standardization) to reach consensus on standardizing metadata and registries. The core metadata registry standard is ISO/IEC 11179 Metadata Registries (MDR), the framework for the standard is described in ISO/IEC 11179-1:2004. A new edition of Part 1 is in its final stage for publication in 2015 or early 2016. It has been revised to align with the current edition of Part 3, ISO/IEC 11179-3:2013 which extends the MDR to support registration of Concept Systems. This standard specifies a schema for recording both the meaning and technical structure of the data for unambiguous usage by humans and computers. ISO/IEC 11179 standard refers to metadata as information objects about data, or “data about data”. In ISO/IEC 11179 Part-3, the information objects are data about Data Elements, Value Domains, and other reusable semantic and representational information objects that describe the meaning and technical details of a data item. This standard also prescribes the details for a metadata registry, and for registering and administering the information objects within a Metadata Registry. ISO/IEC 11179 Part 3 also has provisions for describing compound structures that are derivations of other data elements, for example through calculations, collections of one or more data elements, or other forms of derived data. While this standard describes itself originally as a “data element” registry, its purpose is to support describing and registering metadata content independently of any particular application, lending the descriptions to being discovered and reused by humans or computers in developing new applications, databases, or for analysis of data collected in accordance with the registered metadata content. This standard has become the general basis for other kinds of metadata registries, reusing and extending the registration and administration portion of the standard.

The Dublin Core metadata terms are a set of vocabulary terms which can be used to describe resources for the purposes of discovery. The original set of 15 classic metadata terms, known as the Dublin Core Metadata Element Set are endorsed in the following standards documents:

- IETF RFC 5013
- ISO Standard 15836-2009
- NISO Standard Z39.85.

Although not a standard, Microformat (also mentioned in the section metadata on the internet below) is a web-based approach to semantic markup which seeks to re-use existing HTML/XHTML tags to convey metadata. Microformat follows XHTML and HTML standards but is not a standard in itself. One advocate of microformats, Tantek Çelik, characterized a problem with alternative approaches:

> “Here’s a new language we want you to learn, and now you need to output these additional files on your server. It’s a hassle. (Microformats) lower the barrier to entry.”

Use

Photographs

Metadata may be written into a digital photo file that will identify who owns it, copyright and contact information, what brand or model of camera created the file, along with exposure information (shutter speed, f-stop, etc.) and descriptive information, such as keywords about the photo, making the file or image searchable on a computer and/or the Internet. Some metadata is created by the camera and some is input by the photographer and/or software after downloading to a computer. Most digital cameras write metadata about model number, shutter speed, etc., and some enable you to edit it; this functionality has been available on most Nikon DSLRs since the Nikon D3, on most new Canon cameras since the Canon EOS 7D, and on most Pentax DSLRs since the Pentax K-3. Metadata can be used to make organizing in post-production easier with the use of key-wording. Filters can be used to analyze a specific set of photographs and create selections on criteria like rating or capture time.

Photographic Metadata Standards are governed by organizations that develop the following standards. They include, but are not limited to:

- IPTC Information Interchange Model IIM (International Press Telecommunications Council),
- IPTC Core Schema for XMP
- XMP – Extensible Metadata Platform (an ISO standard)
- Exif – Exchangeable image file format, Maintained by CIPA (Camera & Imaging Products Association) and published by JEITA (Japan Electronics and Information Technology Industries Association)

- Dublin Core (Dublin Core Metadata Initiative – DCMI)
- PLUS (Picture Licensing Universal System).
- VRA Core (Visual Resource Association)

Telecommunications

Information on the times, origins and destinations of phone calls, electronic messages, instant messages and other modes of telecommunication, as opposed to message content, is another form of metadata. Bulk collection of this call detail record metadata by intelligence agencies has proven controversial after disclosures by Edward Snowden Intelligence agencies such as the NSA are keeping online metadata of millions of internet user for up to a year, regardless of whether or not they are persons of interest to the agency.

Video

Metadata is particularly useful in video, where information about its contents (such as transcripts of conversations and text descriptions of its scenes) is not directly understandable by a computer, but where efficient search of the content is desirable. There are two sources in which video metadata is derived: (1) operational gathered metadata, that is information about the content produced, such as the type of equipment, software, date, and location; (2) human-authored metadata, to improve search engine visibility, discoverability, audience engagement, and providing advertising opportunities to video publishers. In today's society most professional video editing software has access to metadata. Avid's MetaSync and Adobe's Bridge are two prime examples of this.

Web Pages

Web pages often include metadata in the form of meta tags. Description and keywords in meta tags are commonly used to describe the Web page's content. Meta elements also specify page description, key words, authors of the document, and when the document was last modified. Web page metadata helps search engines and users to find the types of web pages they are looking for.

Creation

Metadata can be created either by automated information processing or by manual work. Elementary metadata captured by computers can include information about when an object was created, who created it, when it was last updated, file size, and file extension. In this context an *object* refers to any of the following:

- A physical item such as a book, CD, DVD, a paper map, chair, table, flower pot, etc.
- An electronic file such as a digital image, digital photo, electronic document, program file, database table, etc.

Data virtualization has emerged in the 2000s as the new software technology to complete the virtualization "stack" in the enterprise. Metadata is used in data virtualization servers which are enterprise infrastructure components, alongside database and application servers. Metadata in these servers is saved as persistent repository and describe business objects in various enterprise systems and applications. Structural metadata commonality is also important to support data virtualization.

Statistics and Census Services

Standardization work has had a large impact on efforts to build metadata systems in the statistical community. Several metadata standards are described, and their importance to statistical agencies is discussed. Applications of the standards at the Census Bureau, Environmental Protection Agency, Bureau of Labor Statistics, Statistics Canada, and many others are described. Emphasis is on the impact a metadata registry can have in a statistical agency.

Library and Information Science

Metadata has been used in various ways as a means of cataloging items in libraries in both digital and analog format. Such data helps classify, aggregate, identify, and locate a particular book, DVD, magazine or any object a library might hold in its collection. Until the 1980s, many library catalogues used 3x5 inch cards in file drawers to display a book's title, author, subject matter, and an abbreviated alpha-numeric string (call number) which indicated the physical location of the book within the library's shelves. The Dewey Decimal System employed by libraries for the classification of library materials by subject is an early example of metadata usage. Beginning in the 1980s and 1990s, many libraries replaced these paper file cards with computer databases. These computer databases make it much easier and faster for users to do keyword searches. Another form of older metadata collection is the use by US Census Bureau of what is known as the "Long Form." The Long Form asks questions that are used to create demographic data to find patterns of distribution. Libraries employ metadata in library catalogues, most commonly as part of an Integrated Library Management System. Metadata is obtained by cataloguing resources such as books, periodicals, DVDs, web pages or digital images. This data is stored in the integrated library management system, ILMS, using the MARC metadata standard. The purpose is to direct patrons to the physical or electronic location of items or areas they seek as well as to provide a description of the item/s in question.

More recent and specialized instances of library metadata include the establishment of digital libraries including e-print repositories and digital image libraries. While often based on library principles, the focus on non-librarian use, especially in providing metadata, means they do not follow traditional or common cataloging approaches. Given the custom nature of included materials, metadata fields are often specially created e.g. taxonomic classification fields, location fields, keywords or copyright statement. Standard file information such as file size and format are usually automatically included. Library operation has for decades been a key topic in efforts toward international standardization. Standards for metadata in digital libraries include Dublin Core, METS, MODS, DDI, DOI, URN, PREMIS schema, EML, and OAI-PMH. Leading libraries in the world give hints on their metadata standards strategies.

In Museums

Metadata in a museum context is the information that trained cultural documentation specialists, such as archivists, librarians, museum registrars and curators, create to index, structure, describe, identify, or otherwise specify works of art, architecture, cultural objects and their images. Descriptive metadata is most commonly used in museum contexts for object identification and resource recovery purposes.

Usage

Metadata is developed and applied within collecting institutions and museums in order to:

- Facilitate resource discovery and execute search queries.
- Create digital archives that store information relating to various aspects of museum collections and cultural objects, and serves for archival and managerial purposes.
- Provide public audiences access to cultural objects through publishing digital content online.

Standards

Many museums and cultural heritage centers recognize that given the diversity of art works and cultural objects, no single model or standard suffices to describe and catalogue cultural works. For example, a sculpted Indigenous artifact could be classified as an artwork, an archaeological artifact, or an Indigenous heritage item. The early stages of standardization in archiving, description and cataloging within the museum community began in the late 1990s with the development of standards such as Categories for the Description of Works of Art (CDWA), Spectrum, the Conceptual Reference Model (CIDOC), Cataloging Cultural Objects (CCO) and the CDWA Lite XML schema. These standards use HTML and XML markup languages for machine processing, publication and implementation. The Anglo-American Cataloguing Rules (AACR), originally developed for characterizing books, have also been applied to cultural objects, works of art and architecture. Standards, such as the CCO, are integrated within a Museum's Collection Management System (CMS), a database through which museums are able to manage their collections, acquisitions, loans and conservation. Scholars and professionals in the field note that the "quickly evolving landscape of standards and technologies" create challenges for cultural documentarians, specifically non-technically trained professionals. Most collecting institutions and museums use a relational database to categorize cultural works and their images. Relational databases and metadata work to document and describe the complex relationships amongst cultural objects and multi-faceted works of art, as well as between objects and places, people and artistic movements. Relational database structures are also beneficial within collecting institutions and museums because they allow for archivists to make a clear distinction between cultural objects and their images; an unclear distinction could lead to confusing and inaccurate searches.

Cultural Objects and Art Works

An object's materiality, function and purpose, as well as the size (e.g., measurements, such as height, width, weight), storage requirements (e.g., climate-controlled environment) and focus of the museum and collection, influence the descriptive depth of the data attributed to the object by cultural documentarians. The established institutional cataloging practices, goals and expertise of cultural documentarians and database structure also influence the information ascribed to cultural objects, and the ways in which cultural objects are categorized. Additionally, museums often employ standardized commercial collection management software that prescribes and limits the ways in which archivists can describe artworks and cultural objects. As well, collecting institutions and museums use Controlled Vocabularies to describe cultural objects and artworks in their collections. Getty Vocabularies and the Library of Congress Controlled Vocabularies are reputable within the museum community and are recommended by CCO standards. Museums are encouraged to use controlled vocabularies that are contextual and relevant to their collections and enhance the functionality of their digital information systems. Controlled Vocabularies are beneficial within databases because they provide a high level of consistency, improving resource retrieval. Metadata structures, including controlled vocabularies, reflect the ontologies of the systems from which they were created. Often the processes through which cultural objects are described and categorized through metadata in museums do not reflect the perspectives of the maker communities.

Museums and The Internet

Metadata has been instrumental in the creation of digital information systems and archives within museums, and has made it easier for museums to publish digital content online. This has enabled audiences who might not have had access to cultural objects due to geographic or economic barriers to have access to them. In the 2000s, as more museums have adopted archival standards and created intricate databases, discussions about Linked Data between museum databases have come up in the museum, archival and library science communities. Collection Management Systems (CMS) and Digital Asset Management tools can be local or shared systems. Digital Humanities scholars note many benefits of interoperability between museum databases and collections, while also acknowledging the difficulties achieving such interoperability.

Law

United States of America

Problems involving metadata in litigation in the United States are becoming widespread. Courts have looked at various questions involving metadata, including the discoverability of metadata by parties. Although the Federal Rules of Civil Procedure have only specified rules about electronic documents, subsequent case law has elaborated on the requirement of parties to reveal metadata. In October 2009, the Arizona Supreme Court has ruled that metadata records are public record. Document metadata have proven particularly important in legal environments in which litigation has requested metadata, which can include sensitive information detrimental to a certain party in court. Using metadata removal tools to “clean” or redact documents can mitigate the risks of unwittingly sending sensitive data. This process partially protects law firms from potentially damaging leaking of sensitive data through electronic discovery.

Australia

In Australia the need to strengthen national security has resulted in the introduction of a new metadata storage law. This new law means that both security and policing agencies will be allowed to access up to two years of an individual’s metadata, supposedly to make it easier to stop any terrorist attacks and serious crimes from happening. In the 2000s, the law does not allow access to content of people's messages, phone calls or email and web-browsing history, but these provisions could be changed by the government.

In Healthcare

Australian medical research pioneered the definition of metadata for applications in health care. That approach offers the first recognized attempt to adhere to international standards in medical sciences instead of defining a proprietary standard under the World Health Organization (WHO) umbrella. The medical community yet did not approve the need to follow metadata standards despite research that supported these standards.

Data Warehousing

Data warehouse (DW) is a repository of an organization's electronically stored data. Data warehouses are designed to manage and store the data. Data warehouses differ from business intelligence (BI) systems, because BI systems are designed to use data to create reports and analyze the information, to provide strategic guidance to management. Metadata is an important tool in how data is stored in data warehouses. The purpose of a data warehouse is to house standardized, structured, consistent, integrated, correct, "cleaned" and timely data, extracted from various operational systems in an organization. The extracted data are integrated in the data warehouse environment to provide an enterprise-wide perspective. Data are structured in a way to serve the reporting and analytic requirements. The design of structural metadata commonality using a data modeling method such as entity relationship model diagramming is important in any data warehouse development effort. They detail metadata on each piece of data in the data warehouse. An essential component of a data warehouse/business intelligence system is the metadata and tools to manage and retrieve the metadata. Ralph Kimball describes metadata as the DNA of the data warehouse as metadata defines the elements of the data warehouse and how they work together.

Kimball et al. refers to three main categories of metadata: Technical metadata, business metadata and process metadata. Technical metadata is primarily definitional, while business metadata and process metadata is primarily descriptive. The categories sometimes overlap.

- Technical metadata defines the objects and processes in a DW/BI system, as seen from a technical point of view. The technical metadata includes the system metadata, which defines the data structures such as tables, fields, data types, indexes and partitions in the relational engine, as well as databases, dimensions, measures, and data mining models. Technical metadata defines the data model and the way it is displayed for the users, with the reports, schedules, distribution lists, and user security rights.
- Business metadata is content from the data warehouse described in more user-friendly terms. The business metadata tells you what data you have, where they come from, what they mean and what their relationship is to other data in the data warehouse. Business metadata may also serve as a documentation for the DW/BI system. Users who browse the data warehouse are primarily viewing the business metadata.
- Process metadata is used to describe the results of various operations in the data ware-

house. Within the ETL process, all key data from tasks is logged on execution. This includes start time, end time, CPU seconds used, disk reads, disk writes, and rows processed. When troubleshooting the ETL or query process, this sort of data becomes valuable. Process metadata is the fact measurement when building and using a DW/BI system. Some organizations make a living out of collecting and selling this sort of data to companies - in that case the process metadata becomes the business metadata for the fact and dimension tables. Collecting process metadata is in the interest of business people who can use the data to identify the users of their products, which products they are using, and what level of service they are receiving.

On The Internet

The HTML format used to define web pages allows for the inclusion of a variety of types of metadata, from basic descriptive text, dates and keywords to further advanced metadata schemes such as the Dublin Core, e-GMS, and AGLS standards. Pages can also be geotagged with coordinates. Metadata may be included in the page's header or in a separate file. Microformats allow metadata to be added to on-page data in a way that regular web users do not see, but computers, web crawlers and search engines can readily access. Many search engines are cautious about using metadata in their ranking algorithms due to exploitation of metadata and the practice of search engine optimization, SEO, to improve rankings. See Meta element article for further discussion. This cautious attitude may be justified as people, according to Doctorow, are not executing care and diligence when creating their own metadata and that metadata is part of a competitive environment where the metadata is used to promote the metadata creators own purposes. Studies show that search engines respond to web pages with metadata implementations, and Google has an announcement on its site showing the meta tags that its search engine understands. Enterprise search startup Swiftype recognizes metadata as a relevance signal that webmasters can implement for their website-specific search engine, even releasing their own extension, known as Meta Tags 2.

In Broadcast Industry

In broadcast industry, metadata is linked to audio and video broadcast media to:

- *identify* the media: clip or playlist names, duration, timecode, etc.
- *describe* the content: notes regarding the quality of video content, rating, description (for example, during a sport event, keywords like *goal*, *red card* will be associated to some clips)
- *classify* media: metadata allows to sort the media or to easily and quickly find a video content (a TV news could urgently need some archive content for a subject). For example, the BBC have a large subject classification system, Lonclass, a customized version of the more general-purpose Universal Decimal Classification.

This metadata can be linked to the video media thanks to the video servers. Most major broadcast sport events like FIFA World Cup or the Olympic Games use this metadata to distribute their video content to TV stations through keywords. It is often the host broadcaster who is in charge of organizing metadata through its *International Broadcast Centre* and its video servers. This metadata is recorded with the images and are entered by metadata operators (*loggers*) who associate in live metadata available in *metadata grids* through software (such as Multicam(LSM) or IPDirector used during the FIFA World Cup or Olympic Games).

Geospatial

Metadata that describes geographic objects in electronic storage or format (such as datasets, maps, features, or documents with a geospatial component) has a history dating back to at least 1994 (refer MIT Library page on FGDC Metadata). This class of metadata is described more fully on the geospatial metadata article.

Ecological and Environmental

Ecological and environmental metadata is intended to document the "who, what, when, where, why, and how" of data collection for a particular study. This typically means which organization or institution collected the data, what type of data, which date(s) the data was collected, the rationale for the data collection, and the methodology used for the data collection. Metadata should be generated in a format commonly used by the most relevant science community, such as Darwin Core, Ecological Metadata Language, or Dublin Core. Metadata editing tools exist to facilitate metadata generation (e.g. Metavist, Mercury: Metadata Search System, Morpho). Metadata should describe provenance of the data (where they originated, as well as any transformations the data underwent) and how to give credit for (cite) the data products.

Digital Music

When first released in 1982, Compact Discs only contained a Table Of Contents (TOC) with the number of tracks on the disc and their length in samples. Fourteen years later in 1996, a revision of the CD Red Book standard added CD-Text to carry additional metadata. But CD-Text was not widely adopted. Shortly thereafter, it became common for personal computers to retrieve metadata from external sources (e.g. CDDB, Gracenote) based on the TOC.

Digital audio formats such as digital audio files superseded music formats such as cassette tapes and CDs in the 2000s. Digital audio files could be labelled with more information than could be contained in just the file name. That descriptive information is called the audio tag or audio metadata in general. Computer programs specializing in adding or modifying this information are called tag editors. Metadata can be used to name, describe, catalogue and indicate ownership or copyright for a digital audio file, and its presence makes it much easier to locate a specific audio file within a group, typically through use of a search engine that accesses the metadata. As different digital audio formats were developed, attempts were made to standardize a specific location within the digital files where this information could be stored.

As a result, almost all digital audio formats, including mp3, broadcast wav and AIFF files, have similar standardized locations that can be populated with metadata. The metadata for compressed and uncompressed digital music is often encoded in the ID3 tag. Common editors such as TagLib support MP3, Ogg Vorbis, FLAC, MPC, Speex, WavPack TrueAudio, WAV, AIFF, MP4, and ASF file formats.

Cloud Applications

With the availability of Cloud applications, which include those to add metadata to content, metadata is increasingly available over the Internet.

Administration and Management

Storage

Metadata can be stored either *internally*, in the same file or structure as the data (this is also called *embedded metadata*), or *externally*, in a separate file or field from the described data. A data repository typically stores the metadata *detached* from the data, but can be designed to support embedded metadata approaches. Each option has advantages and disadvantages:

- Internal storage means metadata always travels as part of the data they describe; thus, metadata is always available with the data, and can be manipulated locally. This method creates redundancy (precluding normalization), and does not allow managing all of a system's metadata in one place. It arguably increases consistency, since the metadata is readily changed whenever the data is changed.
- External storage allows collocating metadata for all the contents, for example in a database, for more efficient searching and management. Redundancy can be avoided by normalizing the metadata's organization. In this approach, metadata can be united with the content when information is transferred, for example in Streaming media; or can be referenced (for example, as a web link) from the transferred content. On the down side, the division of the metadata from the data content, especially in standalone files that refer to their source metadata elsewhere, increases the opportunity for misalignments between the two, as changes to either may not be reflected in the other.

Metadata can be stored in either human-readable or binary form. Storing metadata in a human-readable format such as XML can be useful because users can understand and edit it without specialized tools. On the other hand, these formats are rarely optimized for storage capacity, communication time, and processing speed. A binary metadata format enables efficiency in all these respects, but requires special libraries to convert the binary information into human-readable content.

Database Management

Each relational database system has its own mechanisms for storing metadata. Examples of relational-database metadata include:

- Tables of all tables in a database, their names, sizes, and number of rows in each table.
- Tables of columns in each database, what tables they are used in, and the type of data stored in each column.

In database terminology, this set of metadata is referred to as the catalog. The SQL standard specifies a uniform means to access the catalog, called the information schema, but not all databases implement it, even if they implement other aspects of the SQL standard. For an example of database-specific metadata access methods, see Oracle metadata. Programmatic access to metadata is possible using APIs such as JDBC, or SchemaCrawler.

Root Cause Analysis

Root cause analysis (RCA) is a method of problem solving used for identifying the root causes of faults or problems. A factor is considered a root cause if removal thereof from the problem-fault-sequence prevents the final undesirable event from recurring; whereas a causal factor is one that affects an event's outcome, but is not a root cause. Though removing a causal factor can benefit an outcome, it does not prevent its recurrence with certainty.

For example, imagine a fictional segment of students who received poor testing scores. After initial investigation, it was verified that students taking tests in the final period of the school day got lower scores. Further investigation revealed that late in the day, the students lacked ability to focus. Even further investigation revealed that the reason for the lack of focus was hunger. So, the root cause of the poor testing scores was hunger, remedied by moving the testing time to soon after lunch.

As another example, imagine an investigation into a machine that stopped because it overloaded and the fuse blew. Investigation shows that the machine overloaded because it had a bearing that wasn't being sufficiently lubricated. The investigation proceeds further and finds that the automatic lubrication mechanism had a pump which was not pumping sufficiently, hence the lack of lubrication. Investigation of the pump shows that it has a worn shaft. Investigation of why the shaft was worn discovers that there isn't an adequate mechanism to prevent metal scrap getting into the pump. This enabled scrap to get into the pump, and damage it. The root cause of the problem is therefore that metal scrap can contaminate the lubrication system. Fixing this problem ought to prevent the whole sequence of events recurring. Compare this with an investigation that does not find the root cause: replacing the fuse, the bearing, or the lubrication pump will probably allow the machine to go back into operation for a while. But there is a risk that the problem will simply recur, until the root cause is dealt with.

Following the introduction of Kepner–Tregoe analysis—which had limitations in the highly complex arena of rocket design, development and launch—RCA arose in the 1950s as a formal study by the National Aeronautics and Space Administration (NASA) in the United States. New methods of problem analysis developed by NASA included a high level assessment practice called MORT (Management Oversight Risk Tree). MORT differed from RCA by assigning causes to common classes of cause shortcomings that could be summarized into a short list. These included work practice, procedures, management, fatigue, time pressure, along with several others. For example: if an aircraft accident occurred as a result of adverse weather conditions augmented by pressure to leave on time; failure to observe weather precautions could indicate a management or training problem; and lack of appropriate weather concern might indict work practices. Because several measures (methods) may effectively address the root causes of a problem, RCA is an iterative process and a tool of continuous improvement.

RCA is applied to methodically identify and correct the root causes of events, rather than to simply address the symptomatic result. Focusing correction on root causes has the goal of entirely preventing problem recurrence. Conversely, RCFA (Root Cause Failure Analysis) recognizes that complete prevention of recurrence by one corrective action is not always possible.

RCA is typically used as a reactive method of identifying event(s) causes, revealing problems and solving them. Analysis is done *after* an event has occurred. Insights in RCA make it potentially useful as a preemptive method. In that event, RCA can be used to *forecast* or predict probable events even *before* they occur. While one follows the other, RCA is a completely separate process to incident management.

Rather than one sharply defined methodology, RCA comprises many different tools, processes, and philosophies. However, several very-broadly defined approaches or "schools" can be identified by their basic approach or field of origin: safety-based, production-based, assembly-based, process-based, failure-based, and systems-based.

- Safety-based RCA arose from the fields of accident analysis and occupational safety and health.
- Production-based RCA has roots in the field of quality control for industrial manufacturing.
- Process-based RCA, a follow-on to production-based RCA, broadens the scope of RCA to include business processes.
- Failure-based RCA originates in the practice of failure analysis as employed in engineering and maintenance.
- Systems-based RCA has emerged as an amalgam of the preceding schools, incorporating elements from other fields such as change management, risk management and systems analysis.

Despite the different approaches among the various schools of root cause analysis, all share some common principles. Several general processes for performing RCA can also be defined.

General Principles

1. The primary aim of root cause analysis is: to identify the factors that resulted in the nature, the magnitude, the location, and the timing of the harmful outcomes (consequences) of one or more past events; to determine what behaviors, actions, inactions, or conditions need to be changed; to prevent recurrence of similar harmful outcomes; and to identify lessons that may promote the achievement of better consequences. ("Success" is defined as the near-certain prevention of recurrence.)
2. To be effective, root cause analysis must be performed systematically, usually as part of an investigation, with conclusions and root causes that are identified backed up by documented evidence. A team effort is typically required.
3. There may be more than one root cause for an event or a problem, wherefore the difficult part is demonstrating the persistence and sustaining the effort required to determine them.

4. The purpose of identifying all solutions to a problem is to prevent recurrence at lowest cost in the simplest way. If there are alternatives that are equally effective, then the simplest or lowest cost approach is preferred.
5. The root causes identified will depend on the way in which the problem or event is defined. Effective problem statements and event descriptions (as failures, for example) are helpful and usually required to ensure the execution of appropriate analyses.
6. One logical way to trace down root causes is by utilizing hierarchical clustering data-mining solutions (such as graph-theory-based data mining). A root cause is defined in that context as "the conditions that enable one or more causes". Root causes can be deductively sorted out from upper groups of which the groups include a specific cause.
7. To be effective, the analysis should establish a sequence of events or timeline for understanding the relationships between contributory (causal) factors, root cause(s) and the defined problem or event to be prevented.
8. Root cause analysis can help transform a reactive culture (one that reacts to problems) into a forward-looking culture (one that solves problems before they occur or escalate). More importantly, RCA reduces the frequency of problems occurring over time within the environment where the process is used.
9. Root cause analysis as a force for change is a threat to many cultures and environments. Threats to cultures are often met with resistance. Other forms of management support may be required to achieve effectiveness and success with root cause analysis. For example, a "non-punitive" policy toward problem identifiers may be required.

General Process for Performing and Documenting an RCA-based Corrective Action

RCA (in steps 3, 4 and 5) forms the most critical part of successful corrective action, directing the corrective action at the true root cause of the problem. Knowing the root cause is secondary to the goal of prevention, as it is not possible to determine an absolutely effective corrective action for the defined problem without knowing the root cause.

1. Define the problem or describe the event to prevent in the future. Include the qualitative and quantitative attributes (properties) of the undesirable outcomes. Usually this includes specifying the natures, the magnitudes, the locations, and the timing of events. In some cases, "lowering the risks of reoccurrences" may be a reasonable target. For example, "lowering the risks" of future automobile accidents is certainly a more economically attainable goal than "preventing all" future automobile accidents.
2. Gather data and evidence, classifying it along a timeline of events to the final failure or crisis. For every behavior, condition, action and inaction, specify in the "timeline" what should have been done when it differs from what was done.
3. In data mining Hierarchical Clustering models, use the clustering groups instead of classifying: (a) peak the groups that exhibit the specific cause; (b) find their upper-groups; (c) find group characteristics that are consistent; (d) check with experts and validate.

4. Ask “why” and identify the causes associated with each sequential step towards the defined problem or event. “Why” is taken to mean “What were the factors that directly resulted in the effect?”
5. Classify causes into two categories: causal factors that relate to an event in the sequence; and root causes that interrupted that step of the sequence chain when eliminated.
6. Identify all other harmful factors that have equal or better claim to be called “root causes.” If there are multiple root causes, which is often the case, reveal those clearly for later optimum selection.
7. Identify corrective action(s) that will, with certainty, prevent recurrence of each harmful effect and related outcomes or factors. Check that each corrective action would, if pre-implemented before the event, have reduced or prevented specific harmful effects.
8. Identify solutions that, when effective and with consensus agreement of the group: prevent recurrence with reasonable certainty; are within the institution’s control; meet its goals and objectives; and do not cause or introduce other new, unforeseen problems.
9. Implement the recommended root cause correction(s).
10. Ensure effectiveness by observing the implemented solutions in operation.
11. Identify other possibly useful methodologies for problem solving and problem avoidance.
12. Identify and address the other instances of each harmful outcome and harmful factor.

References

- Luckham, David C. (2012). Event Processing for Business: Organizing the Real-Time Enterprise. Hoboken, New Jersey: John Wiley & Sons, Inc.,. p. 3. ISBN 978-0-470-53485-4.
- National Information Standards Organization; Rebecca Guenther; Jaqueline Radebaugh (2004). Understanding Metadata (PDF). Bethesda, MD: NISO Press. ISBN 1-880124-62-9. Retrieved 2 April 2014.
- Kimball, Ralph (2008). The Data Warehouse Lifecycle Toolkit (Second ed.). New York: Wiley. pp. 10, 115–117, 131–132, 140, 154–155. ISBN 978-0-470-14977-5.
- Wilson, Paul F.; Dell, Larry D.; Anderson, Gaylord F. (1993). Root Cause Analysis: A Tool for Total Quality Management. Milwaukee, Wisconsin: ASQ Quality Press. pp. 8–17. ISBN 0-87389-163-5.
- Taiichi Ohno (1988). Toyota Production System: Beyond Large-Scale Production. Portland, Oregon: Productivity Press. p. 17. ISBN 0-915299-14-3.
- “VRA Core Support Pages”. Visual Resource Association Foundation. Visual Resource Association Foundation. Retrieved 27 February 2016.
- Hooland, Seth Van; Verborgh, Ruben (2014). Linked Data for Libraries, Archives and Museums: How to Clean, Link and Publish Your Metadata. London: Facet.
- “ISO 15836:2009 - Information and documentation - The Dublin Core metadata element set”. Iso.org. 2009-02-18. Retrieved 2013-08-17.
- Bates, John, John Bates of Progress explains how complex event processing works and how it can simplify the use of algorithms for finding and capturing trading opportunities, Fix Global Trading, retrieved May 14, 2012
- Library of Congress Network Development and MARC Standards Office (2005-09-08). “Library of Congress Washington DC on metadata”. Loc.gov. Retrieved 2011-12-23.

Permissions

All chapters in this book are published with permission under the Creative Commons Attribution Share Alike License or equivalent. Every chapter published in this book has been scrutinized by our experts. Their significance has been extensively debated. The topics covered herein carry significant information for a comprehensive understanding. They may even be implemented as practical applications or may be referred to as a beginning point for further studies.

We would like to thank the editorial team for lending their expertise to make the book truly unique. They have played a crucial role in the development of this book. Without their invaluable contributions this book wouldn't have been possible. They have made vital efforts to compile up to date information on the varied aspects of this subject to make this book a valuable addition to the collection of many professionals and students.

This book was conceptualized with the vision of imparting up-to-date and integrated information in this field. To ensure the same, a matchless editorial board was set up. Every individual on the board went through rigorous rounds of assessment to prove their worth. After which they invested a large part of their time researching and compiling the most relevant data for our readers.

The editorial board has been involved in producing this book since its inception. They have spent rigorous hours researching and exploring the diverse topics which have resulted in the successful publishing of this book. They have passed on their knowledge of decades through this book. To expedite this challenging task, the publisher supported the team at every step. A small team of assistant editors was also appointed to further simplify the editing procedure and attain best results for the readers.

Apart from the editorial board, the designing team has also invested a significant amount of their time in understanding the subject and creating the most relevant covers. They scrutinized every image to scout for the most suitable representation of the subject and create an appropriate cover for the book.

The publishing team has been an ardent support to the editorial, designing and production team. Their endless efforts to recruit the best for this project, has resulted in the accomplishment of this book. They are a veteran in the field of academics and their pool of knowledge is as vast as their experience in printing. Their expertise and guidance has proved useful at every step. Their uncompromising quality standards have made this book an exceptional effort. Their encouragement from time to time has been an inspiration for everyone.

The publisher and the editorial board hope that this book will prove to be a valuable piece of knowledge for students, practitioners and scholars across the globe.

Index

A

Abstraction-based Summarization, 110

Acquiring Information, 242

Aided Summarization, 111

Analytics, 1-4, 6, 8, 10-12, 14, 16, 18, 20-66, 68-74, 76, 78, 80, 82, 84, 86, 88, 90, 92, 94, 96, 98, 100, 102, 104, 106, 108, 110, 112, 114, 116, 118, 120, 122, 124, 126, 128, 130, 132, 134, 136, 138, 140, 142, 144, 146, 148, 150, 152, 154, 156, 158, 160, 162, 164, 166, 168, 170-172, 174, 176, 178, 180, 182, 184, 186, 188, 190, 192, 194, 196, 198, 200, 202, 204, 206, 208-210, 212-214, 216, 218, 222, 224, 226, 228-232, 234, 236, 238, 240, 242, 244, 246, 248, 250, 252, 254, 256, 258, 260, 262, 264, 266, 268, 270, 272-274, 276, 278, 280, 282, 284, 286-288, 290, 292, 294, 296, 298, 300, 302, 304

Anomaly Detection, 64, 66, 73-74, 96

Apriori Algorithm, 79

Association Rule Learning, 64, 66, 75, 81-82

Automatic Summarization, 64, 110-111, 118, 120

B

Behavioral Analytics, 23, 57-60

Behavioral Segmentation, 186-187

Bottom-up Design, 135

Business Activity Monitoring, 20, 264, 273, 275, 278, 285 286

Business Analytics, 3, 22-25, 51, 57, 61, 231, 242, 272, 287

Business Performance Management, 1, 220, 225-229

Business Process Discovery, 220, 230-233, 264-265

Business Process Management, 3, 225, 264, 272-273, 275, 279-283, 286-287

Business Sponsorship, 5

C

Centroid-based Clustering, 86

Cluster Analysis, 64-65, 71, 73-74, 82-83, 85, 94-97, 128, 185, 194

Comparison With Business Analytics, 3

Competence Analysis, 222, 225

Competitive Intelligence, 3, 204, 266-271

Complex Event Processing, 1, 3, 19-20, 27, 272-274, 276 277, 280-281, 305

Conformed Dimension, 142

Connectivity-based Clustering (hierarchical Clustering), 84

Context Analysis, 220, 222-223

Custom-coded Mobile Bi Apps, 15

D

Data Access Via A Mobile Browser, 12

Data Cleansing, 6, 258-260, 262-263

Data Mart, 2, 66, 129-130, 132, 136-139, 162, 169

Data Mining, 1, 3, 28-29, 31-34, 37, 63-75, 78-79, 81-83, 85, 87, 89, 91, 93, 95-97, 99-101, 103, 105, 107, 109-111, 113, 115, 117, 119-128, 130-131, 141, 156, 229, 232, 238, 265, 270, 298, 304

Data Profiling, 5-6, 165, 167-168, 256-258, 262

Data Vault Modeling, 136, 154-155

Data Visualization, 4, 10, 20, 24, 30, 62, 141, 229, 245-247, 249-251, 253-255

Data Warehouse, 2, 6-7, 18-20, 66, 121, 129-139, 141-142, 145, 147, 149, 154-156, 158, 160, 162-164, 168, 172, 227, 257, 259, 262, 271, 298, 305

Degenerate Dimension, 144

Demographic Segmentation, 180, 183-185, 192

Density-based Clustering, 88-90

Digital Analytics, 26

Dimension (data Warehouse), 142

Distribution-based Clustering, 87

E

Embedded Analytics, 22, 30

Examples Of Data Mining, 64, 69, 120

External Evaluation, 92

Extract, Transform, Load, 140, 161-162, 262

Extraction-based Summarization, 110

Extrapolation, 108-109

F

Fixed-form Mobile Bi Applications, 16

Free Open-source Data Mining Software, 71

G

Geographic Segmentation, 183-184, 219

Graphical Tool-developed Mobile Bi Apps, 16

H

Hybrid Design, 136

I

Influence From The Internet, 175

Information Delivery To Mobile Devices, 12

Information System, 4, 133, 220, 231, 234-239, 241, 264

Internal Analysis, 222-223, 225
Internal Evaluation, 90-91
Interpolation, 48, 108

J
Junk Dimension, 143

L
Learning Analytics, 22, 31-37

M
Market Research, 73, 95, 131, 173-179, 181, 183, 185-187, 189, 191, 193, 195, 197, 199, 201, 203-205, 207-213, 215-217, 219, 270
Market Segmentation, 95, 173-174, 178-182, 191, 208, 219
Market Trend, 173, 195-197, 199
Marketing Optimization, 24-25
Marketing Research, 173, 175-176, 204, 207-219, 270
Master Data Management, 6, 129, 136, 139-141, 256
Medical Data Mining, 124
Metadata, 9-10, 131-132, 141, 144, 156-158, 167-168, 256-257, 272-273, 288-302, 305
Mobile Business Intelligence, 11-12, 16, 20
Mobile Client Application, 12
Multi-variable Account Segmentation, 192
Music Data Mining, 126

N
Node-set-based Algorithms, 80
Nonlinear Regression, 109

O
Operational Intelligence, 20, 229, 242, 272-279, 281, 283, 285, 287, 289, 291, 293, 295, 297, 299, 301, 303, 305
Organizational Intelligence, 241-242, 245, 267, 271

P
Portfolio Analytics, 25
Pre-processing, 56, 64, 66-67
Predictive Analytics, 1, 3, 10, 23-24, 27, 37-43, 46-47, 49 52, 63-65, 68-69, 73
Prescriptive Analytics, 1, 3, 23-24, 38, 51-56
Process Mining, 1, 3, 220, 230-231, 233, 264-266, 285
Processing Information, 242
Psychographic Segmentation, 186
Purpose-built Mobile Bi Apps, 13

R
Real-life Etl Cycle, 164
Real-time Business Intelligence, 18-20, 242
Regression Analysis, 46, 64-65, 101-104, 108-109
Results Validation, 66-67
Risk Analytics, 26-27
Role-playing Dimension, 144
Root Cause Analysis, 272-273, 302-305

S
Security Analytics, 26
Sensor Data Mining, 126
Server-less Technology, 19
Slowly Changing Dimension, 129, 146-147
Social Media Analytics, 22, 56
Software Analytics, 22, 26, 28-30
Spatial Data Mining, 125-126
Star Schema, 129, 131, 134, 136, 138, 155, 169-171
Static Data Push, 12
Statistical Classification, 73, 97-98
Swot Analysis, 173-174, 200-202, 204-207, 216, 219-220, 222, 225
Swot Landscape Analysis, 203

T
Temporal Data Mining, 126
Top-down Design, 135
Transmission Of Master Data, 141
Trend Analysis, 220-224, 273

U
Utilization Of Information, 243

V
Versus Operational System, 136
Visual Data Mining, 72, 126