Issues in Data Mining and Information Retrieval

Ammar Yassir and Smitha Nayak,
alfayumi@gmail.com smithank@gmail.com
Department of Computing, Muscat College, Sultanate of Oman

Abstract— Data mining, as we use the term, is the exploration and analysis of large quantities of data in order to discover meaningful patterns and rules. For the purposes of this book, we assume that the goal of data mining is to allow a corporation to improve its marketing, sales, and customer support operations through a better understanding of its customers. Keep in mind, however, that the data mining techniques and tools described here are equally applicable in fields ranging from law enforcement to radio astronomy, medicine, and industrial process control.

In fact, hardly any of the data mining algorithms were first invented with commercial applications in mind. The commercial data miner employs a grab bag of techniques borrowed from statistics, computer science, and machine learning research. The choice of a particular combination of techniques to apply in a particular situation depends on the nature of the data mining task, the nature of the available data, and the skills and preferences of the data miner. Data mining is largely concerned with building models. A model is simply an algorithm or set of rules that connects a collection of inputs (often in the form of fields in a corporate database) to a particular target or outcome.

Keywords: Data Mining, Information Retrieval and Data Warehousing.

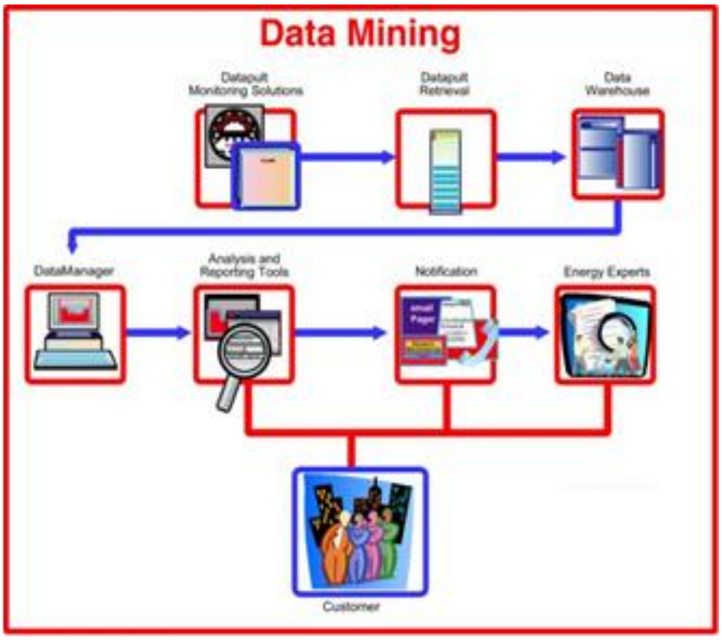
I. INTRODUCTION

Data mining is the drilling down for lost data that has lain dormant, sometimes for years. Often a company has not been aware it possessed this data—usually because of decentralized database management, lack of relational database systems, or the existence of legacy systems with old and forgotten databases. The real value of the data lies in analyzing it to reveal or create relationships that have been previously undiscovered. Having huge banks of data is of no value whatsoever if you don't bother to evaluate it. Evaluation can relate to anything from sales records to seasonal correlations; it can be applied to any supplier–customer relationship, whether in the private or public sector or in industrial, commercial, or consumer markets. (Tukey, 50)

The results of data mining can be grouped as follows:

- Association of events that can be correlated. A computer purchase, for example, is likely to involve the simultaneous purchase of a printer.
- Sequences as one event leads to another. Computer and printer purchase may be followed by the purchase of a scanner.
- Classification through the recognition of patterns. These can be based on any relevant data—income, sales, location, or even average summer rainfall! It all depends on how you see the data benefiting your business.
- Forecasting. This is a natural extrapolation from the other results and can facilitate more accurate projections. Projected beer consumption, for example, could also be related to future consumption of peanuts or potato chips.

In reality, successful data mining starts with data integration. The integration of disparate legacy systems and databases reduces operating costs by cutting out duplicate administration. Further, it provides closer control in forming the basis for more comprehensive customer information and tighter, more focused marketing strategies. (Berry, 12-15)



Nowhere is data integration more necessary than in established industries like insurance, telcos, and utilities. Frequently the established players have legacy systems dating back two decades or more. Their dominance of the market has been secure for a lot longer than that, yet it is now under threat from well-funded competitors with tailored systems that make them faster on their feet when it comes to accessing—and manipulating—data to benefit their business and customers. The customer, of course, is king, and technology is now making a reality of customer-serving trends like e-commerce, customer relationship management (CRM), and enterprise application integration (EAI). For most organizations these areas are linked, because there is a need to understand, and make the best use of, the data that contains the customer information.

II. Aims Of The Study

The aim of this paper is to study research related issues in data mining and information retrieval.

III. Materials and Methods

In order to study research related issues in data mining and information retrieval, a mixed method research strategy (combining qualitative and quantitative methods) was employed with secondary data and analysis. The most popular research approaches used to make informed scientific decisions include qualitative and quantitative methods. Differences between quantitative and qualitative research lie in their approach to identification of the research problems and reviews of the literature. The two approaches have different strategies in specifying the purpose, data collection, data analysis, reporting, and evaluating research. In identifying a research problem, Creswell (2005) claimed that it is descriptive and explanatory for quantitative but exploratory for qualitative research.

Quantitative research uses scientific methods to investigate phenomena and address issues and problems. These methods utilize an objective manner that enhances the reliability of the information and reduces biases. Qualitative research answers questions and explores new knowledge in a natural environment. This approach attempts to understand all aspects of people's behaviors, attitudes, and experiences. To address the research questions, the qualitative approach depends on four main data collections strategies: participation, observation, interviews, and analysis.

Qualitative research explores a given phenomenon in order to provide further understanding and enhanced knowledge. Qualitative research questions are generally broad and the numbers of subjects in the study can be small (Burns & Grove, 2005). In qualitative methods, the researcher depends on the observations or experiences of the participants.

IV. RESEARCH DESIGN

This dissertation utilized a combination of primary and secondary research methods. This paper foresees that the primary research would be external focused as we would collect the opinion and views of experts in the field and practicing users through an e-survey with key stakeholders to understand their views about research related issues in data mining and information retrieval. (Berry, 20)

This research paper followed the method of both primary and secondary collection of data. A number of sources have been utilized for the sake of data extraction like books, internet publications, journals, articles, surveys, interviews and questionnaire. Mixed method of research is aimed at gathering of information with the help of various mediums like focus group interviews, organizational case studies, literature, broadcast media, publications and other kind of primary and secondary sources. This kind of research involved both human and non-human subjects for studying the concept and factors in details.

V. Research Participants

There were 50 participants interviewed and given questionnaires to study research related issues in data mining and information retrieval.

VI. Informed Consent

Informed consent is an important component of research and is an integral part of the research process. For the proposed study, the researcher implemented practical steps to ensure that all participants are educated about the proposed study in order to make an informed decision. Participation were voluntary and individuals having the right to choose not to participate or to withdraw from either phase of the proposed study at any time.

VII. Data Analysis

As the study employed mixed methodology therefore both Qualitative Data Analysis and Quantitative Data Analysis techniques were used. The research analyzed the qualitative part of the study using content analysis. This type of analysis provides an image of the participants' perceptions, feelings, experiences, ideas, concerns, and attitudes. For the study, the content analysis process involved these a few steps. Within research, there are different statistical processes for designing a study as quantitative data analysis. Statistical analysis for example, gives meaning to the numbers collected within a particular study. The categories of statistical procedures include descriptive, associative, and inferential. (Cabena, 96)

VIII. Results

The survey and respondents views regarding the research related issues in data mining and information retrieval showed that the following research related issues are there in data mining and information retrieval but there are some other issues as well.

IX. MISSING VALUES

Some data mining algorithms are capable of treating "missing" as a value and incorporating it into rules. Others cannot handle missing values, unfortunately. None of the obvious solutions preserve the true distribution of the variable. Throwing out all records with missing values introduces bias because it is unlikely that such records are distributed randomly. Replacing the missing value with some likely value such as the mean or the most common value adds spurious information. Replacing the missing value with an unlikely value is even worse since the data mining algorithms will not recognize that –999, say, is an unlikely value for age. The algorithms will go ahead and use it. When missing values must be replaced, the best approach is to impute them by creating a model that has the missing value as its target variable. (Han, 78-80)

X. VALUES WITH MEANINGS THAT CHANGE OVER TIME

When data comes from several different points in history, it is not uncommon for the same value in the same field to have changed its meaning over time. Credit class "A" may always be the best, but the exact range of credit scores that get classed as an "A" may change from time to time. Dealing with this properly requires a well-designed data warehouse where such changes in meaning are recorded so a new variable can be defined that has a constant meaning over time. (Johnson, 82)

XI. INCONSISTENT DATA ENCODING

When information on the same topic is collected from multiple sources, the various sources often represent the same data different ways. If these differences are not caught, they add spurious distinctions that can lead to erroneous conclusions. In one call-detail analysis project, each of the markets studied had a different way of indicating a call to check one's own voice mail. In one city, a call to voice mail from the phone line associated with that mailbox was recorded as having the same origin and destination numbers. In another city, the same situation was represented by the presence of a specific nonexistent number as the call destination. In yet another city, the actual number dialed to reach voice mail was recorded. Understanding apparent differences in voice mail habits between cities required putting the data in a common form. (Mannila, 259–289)

The same data set contained multiple abbreviations for some states and, in some cases, a particular city was counted separately from the rest of the state. If issues like this are not resolved, you may find yourself building a model of calling patterns to California based on data that excludes calls to Los Angeles. (Masi, 99)

XII. ETHICAL ISSUES

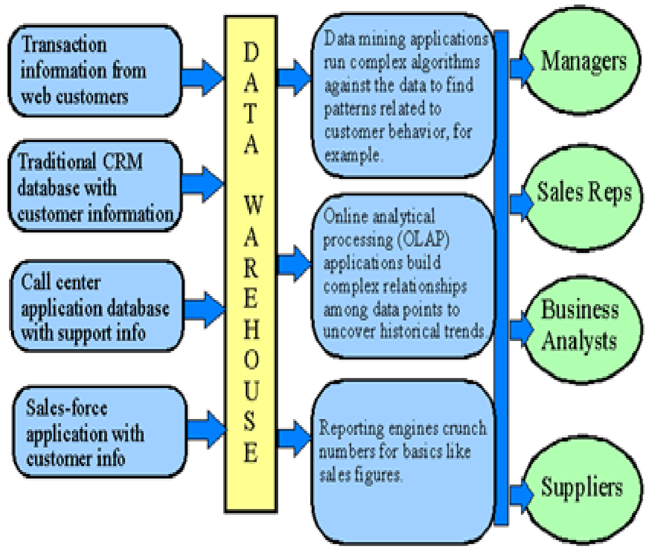
Human systems have proven over and over to be highly nonlinear, with many unintended consequences. Data mining technology is no different in that respect. Massive data storage facilities and software (statistical and artificially intelligent) that are capable of visualizing relationships and identifying patterns from large scale data sets give us many useful tools. However, as with almost everything complex, these technologies can create problems as well as solve them. There has long been a concern with the risks of concentrating too much data into one location. (Murphy, 68)

While data mining offers the ability to learn many useful things, it is a powerful tool which can be used for bad as well as good. In fact, what is good for General Motors is not necessarily good for the consumers of General Motors products. Better information can lead to the design of products more useful to consumers, but which can also be used to extract more money for the same product. Yield management is designed to do just that for American Airlines and other carriers. Yield management is very similar to data mining—identifying the probability of last-minute cancellations to allow overbooking, and develop price schedules that maximize revenue obtained from customers. For most people, this is good business, and is to be commended. In fact, some would like to see similar tools available to consumers to help in competing with the airlines.

XIII. DISCUSSION

The data mining process can fail in many ways. Failure can take several forms, including simply failing to answer the questions you set out to answer, as well as "discovering" things you already know. An especially pernicious form of failure is learning things that aren't true. This can happen in many ways: when the data used for mining is not representative; or when it contains accidental patterns that fail to generalize; or when it has been summarized in a way that destroys information; or when it mixes information from time periods that should be kept separate. (Ruquest, 86)

Data Mining: Dig, Discover, & Share*


Transaction information from web customers
Traditional CRM database with customer information
Call center application database with support info
Sales-force application with customer info
DATA WAREHOUSE
Data mining applications run complex algorithms against the data to find patterns related to customer behavior, for example.
Online analytical processing (OLAP) applications build complex relationships among data points to uncover historical trends.
Reporting engines crunch numbers for basics like sales figures.
Managers
Sales Reps
Business Analysts
Suppliers

* Adapted from Roberts-Witt's illustration, *PC Magazine*, November 19, 2002, p. ubiz 5.

Data useful to business come in many forms. For instance, an automobile insurance company, faced with millions of accident claims, realizes that not all claims are legitimate. If they are extremely tough and investigate each claim thoroughly, they would spend more money on investigation than they would pay out in claims. They would also find that they're unable to sell any new policies. If they were as understanding and trusting as their television ads imply, they would reduce their investigation costs to zero, and leave themselves vulnerable to fraudulent claims. Insurance firms have developed ways to profile claims, considering many variables along the way, and so provide an early indication of cases that probably merit expending funds for investigation. This has the effect of reducing overall policy expenses, because it discourages fraud, while minimizing the imposition on valid claims. The same approach is used by the Internal Revenue Service in processing individual tax returns. Fraud detection has become a viable data mining industry, with a large number of software vendors. This is typical of many applications of data mining.

XIV. Achieving the Essentials

To ensure that customers are closely connected to them, organizations are now striving to establish multichannel distribution strategies, supply chains, e-procurement, knowledge management, and CRM initiatives.
Business development, whether along traditional lines or as e-business via PCs, television, or mobile communications, requires:

1. Interactivity
2. Personalization
3. Secure transactional capability

Refined data mining is essential to achieve the first two; by using it, business can stimulate the customer's desire to interact. (Sullivan, 73)

XV. Digging for the Benefits

Suppliers as well as customers can benefit from data mining and analysis. Every business is a supplier of something, whether it's a service or a product.
From the data mined the supplier, should be able to analyze internal trends that can be capitalized to benefit its own long-term CRM activity. Look to answer some very telling questions:

- Are all the company's service offerings relevant?
- If they are, are they reusable?
- Is customer service proactive? responsive? consistent?
- Is there evidence of steady improvement?
- What is customers' perception of the company?
- Is there a need for internal training?
- Is there evidence of persistent product failings?

Embarking on such a search prompts suppliers to take a long, hard look at their business. And such an exercise doesn't apply only to long-established businesses, although they may benefit the most at the beginning. Every business needs to mine its own data to make sure the rigor mortis of maturity doesn't set in.

XVI. Reducing Customer Churn

In a sense we have arrived at a best-case scenario, with sophisticated integration leading to efficient and effective use of data, both internally and externally. Evidence shows that in a free market companies can lose a very high percentage of their customers every five years—maybe as many as 45–50%. Other research indicates that a rise of just 5 percent in customer retention can result in an 80 percent rise in profit.
An overriding priority for any supplier is therefore how to retain its customers. Good CRM is central to customer retention, and CRM relies on a detailed understanding of customer profiles, together with the direct selling and cross-selling opportunities they reveal. Good customer profiles within the existing base can also be used to identify areas for successful expansion into new ventures, new geographic areas, and new product launches.
In other words, your data provides you with knowledge—and in business, knowledge is power. Sound data analysis allows you to translate that knowledge into proactive marketing to existing customers and accurately-profiled prospects. (Thuraisingham, 45-46)
Ironically, companies often pay big money for information about their existing and prospective markets while the knowledge hidden in their own IT systems is overlooked.

Frequently they don’t realize what they have because one set of data is held separately from another—on different systems, at different locations—and there are no means to collate and mine it. Once data is combined, the knowledge extracted is usually greater than the sum of its parts.

As an example, imagine a scenario involving half-a-dozen companies operating under a group umbrella.

- Transaction 1: A customer has a store card and buys goods over the counter.
- Transaction 2: The same customer buys some Christmas presents through a seasonal mail-order catalog.
- Transaction 3: The customer decides to have a new central heating system installed … and so it goes on.
- A whole series of transactions, all for the same address, but all through different companies and logged on different databases.

With the tools to tie this information together and mine it intelligently, incredibly detailed customer profiles emerge that can put untold knowledge into the hands of marketers, sales executives, designers, engineers—looked at from different angles, the uses are legion.

XVII. Data Warehousing

With data mining the concept of data warehousing is fast gaining acceptance, and some research suggests that 50 percent of all companies are either using a data warehouse or planning to build one.

A data warehouse isn’t simply for data mining, it’s a resource susceptible to a variety of analytical processes. Data is first extracted from operational databases, cleaned up to remove redundant data and fill in blank and missing fields, and then organized into consistent formats.

Analysts can then drill down into the data using data access and data-mining tools as well as online reporting software, including online analytical processing (OLAP), statistical modeling tools, and geographic information systems (GIS).

With a data warehouse you have access to the information that fuels your company’s growth. A profitable future depends heavily on extracting only data that has the potential to become useful, and with data mining you have the ability to extract data you didn’t even know existed.

This will usually be information on customers, partners, and key business trends; the process can also be used for fraud protection, enhancement of customer satisfaction, analysis of product repositioning, discovery of profit centers, or corporate asset management. You can highlight loyal customers, then discover what it is that has kept them loyal—tailoring subsequent offers and benefits to retain them. Similarly you can spot reasons for churn, backtrack to discover the indicators leading to those reasons (frequently missed), and take the necessary steps to reduce it significantly. (Thuraisingham, 45-46)

XVIII. Conclusions

In conclusion, it can be said that this research has found some very important issues in data mining and information retrieval. Many tools are available for data mining and can accomplish a number of functions. The tools come from areas of statistics, operations research, and artificial intelligence, and provide techniques useful in accomplishing a variety of analytic functions, such as cluster identification, discriminant analysis, and development of association rules. Data mining software provides a powerful means of applying these tools to large sets of data, giving organizational management a means of coping with an overwhelming glut of data and converting some of it into useful knowledge.

The point of data mining is to have a variety of tools available to assist the analyst and user in better understanding what the data consists of. Each method does something different, and usually this implies that a specific problem is best treated with a particular algorithm type. However, sometimes different algorithm types can be used for the same problem. Most involve setting parameters, which can be important when it comes to the effectiveness of the method. Further, output needs to be interpreted.

References

[1] Berry, Michael; Gordon, Linoff. Data Mining Techniques: For Marketing, Sales, and Customer Support. New York: Wiley, 2007. Pp. 12-15.

[2] Berry, Michael; Gordon, Linoff. Mastering Data Mining: The Art and Science of Customer Relationship Management. New York: Wiley, 2000. Pp. 20.

[3] Cabena, Peter. Discovering Data Mining: From Concept to Implementation. Upper Saddle River, N.J.: Prentice Hall PTR, 2008. Pp. 96.

[4] Han, Jiawei; Micheline, Kamber. Data Mining: Concepts and Techniques. San Francisco: Morgan Kaufmann Publishers, 2001. Pp. 78-80.

[5] Johnson, R. Colin. “Protocol Aimed at Data Mining.” Electronic Engineering Times September 18, 2000, Pp. 82.

[6] Mannila, H.; Toivonen, H.; Verkamo, A. I.. “Discovery of Frequent Episodes in Event Sequence”. Data Mining and Knowledge Discovery. 1, 3, (2007) Pp. 259–289.

[7] Masi, C.G. "Data Mining Can Tame Mountains of Information." Research & Development, November, 2000. Pp. 99.

[8] Murphy, Victoria. "You've Got Expertise." Forbes, February 5, 2001. Pp. 68.

[9] Ruquest, Mark E. "Planning is Key to Exploiting Technical Data." National Underwriter, November 27, 2000. Pp. 86.

[10] Sullivan, Tom. "Picture This: Data Analysis Becomes More Graphic." InfoWorld, October 16, 2000. Pp. 73.

[11] Thuraisingham, Bhavani. Data Mining: Technologies, Techniques, Tools, and Trends. Boca Raton, Fla.: CRC Press, 2009. Pp. 45-46.

[12] Tukey, J.. Exploratory Data Analysis.. Reading. MA, Addison-Wesley. (2005) Pp. 50.

Ammar Yassir received the B.Sc. degree with Honors in Computer Science in the year 2002 from Future University, Sudan and Master in Business Administration and Information Technology degree from Sikkim Manipal University, India in 2006 and currently a Ph.D. candidate in Information Technology, CMJ University, Shillong, India. He is now lecturer at Muscat College, Sultanate of Oman. He has published an international paper in IJCSNS.

Smitha Nayak received the B.Sc. degree in Physics in 1998 from Mumbai University, India and Master of Computer Application degree from Visweshwaraiya Technological University, India in 2001 and currently a Ph.D. candidate in Information Technology, CMJ University, Shillong, India. She is now lecturer at Muscat College, Sultanate of Oman. She has published an international paper in IJCSNS.