A Review on Duo Mining Techniques

R. Marutha Veni[1], M. Praveena[2], V. GanaPriya[3]

[1]Dr. SNS Rajalakshmi College of Arts and Science,
Saravanampatti, Coimbatore-49, India
maruthaveni12@gmail.com

[2]Dr. SNS Rajalakshmi College of Arts and Science,
Saravanampatti, Coimbatore-49, India
praveenamarannan@gmail.com

[3]Dr. SNS Rajalakshmi College of Arts and Science,
Saravanampatti, Coimbatore-49, India
visitpriya.n@gmail.com

Abstract: ***The combination of data and text mining is referred to as "Duo-mining". Text and data mining are fast growing areas and are believed to have high commercial potential value in knowledge discovery and information filtering areas of application. Although text mining manages unstructured data, most of knowledge discovery and information filtering can be done using data mining. Despite that, both technologies do not actively predict and prevent problems, instead they leave the work to the experts to manually interpret the data, anticipate future events and make the final decision. Also, the paper highlights the benefits of combining duo mining and multi-agents in prediction.***

Keywords: Association, Clustering, Data Mining knowledge, Pattern mining, Regression, Text mining

1. Introduction

Data mining (sometimes called data or knowledge discovery) is the process of analyzing data from different perspectives and summarizing it into useful information. Text mining is similar to data mining, except that data mining tools are designed to handle structured data from databases or XML files, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents, HTML files, etc. As a result, text mining is a much better solution for companies, where large volumes of diverse types of information must be merged and managed. The combination of data and text mining is referred to as "duo-mining". Duo- mining gives companies the edge on consolidated information for better decision making.

2. Data Mining

Data mining, also known as knowledge-discovery in databases (KDD), is the practice of automatically searching large stores of data for patterns. Data mining software is one of a number of analytical tools for analyzing data. It allows users to analyze data from many different dimensions or angles, categorize it, and summarize the relationships identified. Technically, data mining is the process of finding correlations or patterns among dozens of fields in large relational databases.

2.1 Data, Information, and Knowledge

Data are any facts, numbers, or text that can be processed by a computer. Today, organizations are accumulating vast and growing amounts of data in different formats and different databases. This includes:

- Operational or transactional data such as, sales, cost, inventory, payroll, and accounting.
- Nonoperational data, such as industry sales, forecast data, and macro economic data.
- Meta data - data about the data itself, such as logical database design or data dictionary definitions.

2.1.1 Information

The patterns, associations, or relationships among all this data can provide information. For example, analysis of retail point of sale transaction data can yield information on which products are selling and when.

2.2 Knowledge

Information can be converted into knowledge about historical patterns and future trends. For example, summary information on retail supermarket sales can be analyzed in light of promotional efforts to provide knowledge of consumer buying behavior. Thus, a manufacturer or retailer could determine which items are most susceptible to promotional efforts.

2.2.1 Data Warehouses

Data warehousing is defined as a process of centralized data management and retrieval. Data warehousing, like data mining, is a relatively new term although the concept itself has been around for years. Data warehousing represents an ideal vision of maintaining a central repository of all organizational data. Centralization of data is needed to maximize user access and analysis. Dramatic technological advances are making this vision a reality for many companies. And, equally dramatic advances in data analysis software are allowing users to access this data freely. The data analysis software is what supports data mining.

2.3 What can data mining do?

Data mining is primarily used today by companies with a strong consumer focus - retail, financial, communication, and marketing organizations. It enables these companies to

determine relationships among "internal" factors such as price, product positioning, or staff skills, and "external" factors such as economic indicators, competition, and customer demographics. And, it enables them to determine the impact on sales, customer satisfaction, and corporate profits. Finally, it enables them to "drill down" into summary information to view detail transactional data.

2.3.1 Example

For example, Blockbuster Entertainment mines its video rental history database to recommend rentals to individual customers. American Express can suggest products to its cardholders based on analysis of their monthly expenditures.

2.4 How does data mining work?

While large-scale information technology has been evolving separate transaction and analytical systems, data mining provides the link between the two. Data mining software analyzes relationships and patterns in stored transaction data based on open-ended user queries. Several types of analytical software are available: statistical, machine learning, and neural networks.

Generally, any of four types of relationships are sought:

- Classes: Stored data is used to locate data in predetermined groups. For example, a restaurant chain could mine customer purchase data to determine when customers visit and what they typically order. This information could be used to increase traffic by having daily specials.

- Clusters: Data items are grouped according to logical relationships or consumer preferences. For example, data can be mined to identify market segments or consumer affinities.

- Associations: Data can be mined to identify associations. The beer-diaper example is an example of associative mining.

- Sequential patterns: Data is mined to anticipate behavior patterns and trends. For example, an outdoor equipment retailer could predict the likelihood of a backpack being purchased based on a consumer's purchase of sleeping bags and hiking shoes.

Data mining consists of five major elements:

- Extract, transform, and load transaction data onto the data warehouse system.
- Store and manage the data in a multidimensional database system.
- Provide data access to business analysts and information technology professionals.
- Analyze the data by application software.
- Present the data in a useful format, such as a graph or table.

Different levels of analysis are available:

- Artificial neural networks: Non-linear predictive models that learn through training and resemble biological neural networks in structure.
- Genetic algorithms: Optimization techniques that use processes such as genetic combination, mutation, and natural selection in a design based on the concepts of natural evolution.
- Decision trees: Tree-shaped structures that represent sets of decisions. These decisions generate rules for the classification of a dataset. Specific decision tree methods include Classification and Regression Trees (CART) and Chi Square Automatic Interaction Detection (CHAID). CART and CHAID are decision tree techniques used for classification of a dataset. They provide a set of rules that you can apply to a new (unclassified) dataset to predict which records will have a given outcome. CART segments a dataset by creating 2-way splits while CHAID segments using chi square tests to create multi-way splits. CART typically requires less data preparation than CHAID.
- Nearest neighbor method: A technique that classifies each record in a dataset based on a combination of the classes of the k record(s) most similar to it in a historical dataset (where k 1). Sometimes called the k-nearest neighbor technique.
- Rule induction: The extraction of useful if-then rules from data based on statistical significance.
- Data visualization: The visual interpretation of complex relationships in multidimensional data. Graphics tools are used to illustrate data relationships.

2.5 What technological infrastructure is required?

There are two critical technological drivers:

- Size of the database: the more data being processed and maintained, the more powerful the system required.
- Query complexity: the more complex the queries and the greater the number of queries being processed, the more powerful the system required.

Relational database storage and management technology is adequate for many data mining applications less than 50 gigabytes. However, this infrastructure needs to be significantly enhanced to support larger applications. Some vendors have added extensive indexing capabilities to improve .query performance. Others use new hardware architectures such as Massively Parallel Processors (MPP) to achieve order-of-magnitude improvements in query time. For example, MPP systems from NCR link hundreds of high-speed Pentium processors to achieve performance levels exceeding those of the largest supercomputers.

2.5 Data mining classes

Data mining commonly involves four classes of tasks:

- 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 email program might attempt to classify an email as legitimate

or spam. Common algorithms include decision tree learning, nearest neighbor, naive Bayesian classification, neural networks and support vector machines.
- Regression - Attempts to find a function which models the data with the least error.
- Association rule learning - 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 sometimes referred to as market basket analysis.

3. Applications of Data Mining

Games: The availability of oracles for certain combinatorial games, also called table bases (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 up. This is the extraction of human-usable strategies from these oracles.

Business: 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 Data clustering can also be used to automatically discover the segments or groups within a customer data set. Data mining can also be helpful to human-resources departments in identifying the characteristics of their most successful employees.

Science and engineering: Data mining has been widely used in area of science and engineering, such as bioinformatics, genetics, medicine, education and electrical power engineering.

Spatial data mining: Spatial data mining is the application of data mining techniques to spatial data. Spatial data mining follows along the same functions in data mining, with the end objective to find patterns in 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.

3.1 Pattern mining

"Pattern mining" is a data mining technique 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.

3.2 Subject-based data mining

"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.

4. Data Mining Technology

It is a process of extracting knowledge hidden in large volumes of raw data.

- Competitive advantage requires abilities.
- Abilities are built through knowledge.
- Knowledge comes from data.
- The process of extracting knowledge from data is called Data Mining.

Typical tasks addressed by data mining include:

- Rate customers by their propensity to respond to an offer
- Identify cross-sell opportunities
- Detect fraud and abuse in insurance and finance
- Estimate probability of an illness re-occurrence or hospital re-admission
- Isolate root causes of an outcome in clinical studies
- Determine optimal sets of parameters for a production line operation
- Predict peak load of a network

4.1 Academic Resources

- A collection of various academic publications from Megaputer.
- Anomaly Localization - a clustering algorithm with automated attribute selection and invariant to functional coordinate transformations
- Generating Non-linear Functions - PolyAnalyst's data analysis techniques
- Inferring Functional Programs with Machine Learning
- Discovering Numeric Dependencies in the form of Rational Expressions
- Symbolic Knowledge Discovery - introducing PolyAnalyst 2.0, a combination of statistical data preprocessing and symbolic KDD techniques.

5. Text mining

Text mining, also known as intelligent text analysis, text data mining or knowledge-discovery in text (KDT), refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining has been defined as “the discovery by computer of new, previously unknown information, by automatically extracting information from different written resources.

5.1 Purpose of text mining

To discover and use knowledge that is contained in a document collection as a whole, extracting essential information from document collections and from a variety of different sources. Text mining lets executives ask questions of their text-based resources quickly extract information and find answers they never imagined.

5.2 Steps to Text Mining

- Preprocessing the text to distill the documents into a structured format.
- Reducing the results into a more practical size.

- Mining the reduced data with traditional data mining techniques.

Text preprocessing transforms text into an information-rich, term-by-document matrix. This large grid indicates the frequency of every term within the document collection. During this stage, feature extraction is also used to locate specific bits of information, such as customer names, organizations and addresses.

Clustering, classification and predictive methods are applied to the reduced data, using traditional data mining techniques. Conventional structured data sources can also be included in the analysis to enrich the discovery of underlying trends and patterns within the data.

5.3 Examples of Text Mining

a. Sales and marketing executives can count on text mining tools to analyze company descriptions in their prospect database. The results help executives target customers for new sales and marketing campaigns.

b. Linguists at a university in Belgium use text mining to analyze summaries of ancient and modern texts. Researchers mine textual information in several languages and use the results to address philological and psychological questions.

c. A new text mining project at a university medical center will let doctors make better use of medical databases such as Medline, PsychInfo and Toxline for evidence-based medicine. Search results of these medical databases can often yield 2,000 matches, but advanced modeling with text mining technologies can reduce the results to 100 highly relevant documents and sort those 100 documents into smaller subgroups or categories.

6. Text Mining Technology

It is a process of making sense from unstructured data.

Organizations use natural language to communicate with employees, customers, partners and general public, as well as to organize information internally for future reference.

Manual analysis is inefficient for processing large volumes of text, as it is slow, biased and prone to human errors. Statistical techniques fall short of achieving this goal because text documents contain hidden linguistic and semantic relationships that have to be taken into account. Eliciting knowledge from unstructured text represents a major technological challenge.

6.1 Solutions

The solution is offered by Text Mining – the technology for automated knowledge discovery in large volumes of text based on a combination of linguistic, semantic, statistical and machine learning techniques.

Many important tasks can be solved with text mining:

- Automate and increase quality of the analysis of survey responses
- Identify main repair issues and generate reports by analyzing call center transcripts
- Determine root causes of problems from the analysis of incident reports
- Detect and visualize correlations in the usage of biomarkers in research articles
- Predict the best subrogation potential cases from the analysis of insurance claims notes

7. Duo Mining

The combination of data and text mining is referred to as “duo-mining”. Duo-mining gives companies the edge on consolidated information for better decision making. This process combination has proven to be especially useful to banking and credit card companies. Instead of only being able to analyze the structured data they collect from transactions, they can add call logs from customer services and further analyze customers and spending patterns from the text mining side. These new developments in text mining technology that go beyond simple searching methods are the key to information discovery.

7.1 Put Them Together and You Get High Value

Recently, vendors such as Intelligent Results, SAS and SPSS have started to recommend to their customers that they combine data and text mining. And the results have been interesting, to say the least.

This is not surprising, for two reasons. First, the enterprise has vastly expanded the universe in which to find patterns - always a good thing. Secondly, a pattern in data or text can amplify or clarify patterns in its counterpart. In both cases, there is a multiplier effect going on.

But rather than being theoretical, let's be specific. Collections and recovery departments in banks and credit card companies have used duo-mining to good effect. Using data mining to look at repayment trends, these enterprises have a good idea on who is going to default on a loan, for example. When logs from the collection agents are added to the mix, the understanding gets even better. For example, text mining can understand the difference in intent between, "I will pay," "I won't pay," "I paid" and generate a propensity to pay score - which, in turn, can be data mined. To take another example, if a customer says, "I can't pay because a tree fell on my house;" all of a sudden it is clear that it's not a "bad" delinquency - but rather a sales opportunity for a home loan.

By using data mining and text mining in tandem, enterprises have been able to improve average "lift" over using just one technology to around 20 percent, with the range being from 5 to 50 percent. Other areas where duo-mining has paid off include analyzing product wish lists, open-ended survey questions and customer attrition patterns at cell phone companies.

7.2 Some Practical Hints

Companies looking to do duo-mining in such applications need to be wary of several things, especially in regards to

text mining. First, some text mining technologies need large amounts of text to analyze - several page memos, for example - while call logs are sometimes just snippets in comparison. Second, "stemming," a popular technique in text analysis in which various forms of a word are distilled into one word - "pay," "paid," "will pay," "won't pay" = "pay" - may need to be turned off. To take the collections example, stemming would prevent the enterprise from understanding the customer's intent. Therefore, companies need to ensure that the technology they're using is tuned to the problem at hand.

In addition, some companies' solutions are more toolkit-oriented (SAS and SPSS) while others are more application-oriented (Intelligent Results). Which is more appropriate depends on what the company wants to do and the level of in-house expertise.
With those caveats in mind, enterprises should investigate duo-mining. It's a combination of two time-tested technologies that can lead to big payoffs.

8. Conclusion

As the amount of structured and unstructured data in our world continues to increase, duo mining tools that allow us to sift through this information with ease will become more and more valuable. Duo mining tools are beginning to be readily applied in the biomedical field, where the volume of information on a particular topic makes it impossible for a researcher to cover all the material, much less explore related texts. Duo mining methods can also be used by the government’s intelligence and security agencies to try to piece together terrorist warnings and other security threats before they occur. Another area that is already benefiting from duo mining tools is education. Students and educators can find more information relating to their topics at faster speeds than they can use traditional ad hoc searching.

References

[1] http://www.information-management.com/news/1010449-1.html
[2] http://filebox.vt.edu/users/wfan/paper/text_mining_final_preprint.pdf
[3] Gordon, M.D, Lindsay, R and Fan, W. Literature-based discovery on the WWW, ACM Transactions on Internet Technology (TOIT), 2, 4, (2002), 262-275.
[4] Han, J, Altman, R.B, Kumar, V, Mannila, H, and Pregibon, D. Emerging Scientific Applications in Data Mining. CACM, 45, 8, (2002), 54-58.
[5] Hearst,M.What is text mining. http://www.sims.berkeley.edu/~hearst/textmining.html, (2004)
[6] Informatik http://www-i5.informatik.rwthaachen.de/lehrstuhl/projects/DocMINER/DocMINER.html, 2004
[7] http://dml.cs.byu.edu/~cgc/docs/mldm_tools/Reading/Power%20of%20Text%20Mining.pdf
[8] http://library.witpress.com/pages/listpapers.asp?q_bid=429
[9] http://www.aaai.org/aitopics/pmwiki/pmwiki.php/AITopics/DataMining