International Journal of Information, Control and Computer Sciences
ISSN: 2517-9942
Vol:1, No:8, 2007

Powerful Tool to Expand Business Intelligence: Text Mining

Li Gao, Elizabeth Chang, and Song Han

***Abstract*—**With the extensive inclusion of document, especially text, in the business systems, data mining does not cover the full scope of Business Intelligence. Data mining cannot deliver its impact on extracting useful details from the large collection of unstructured and semi-structured written materials based on natural languages. The most pressing issue is to draw the potential business intelligence from text. In order to gain competitive advantages for the business, it is necessary to develop the new powerful tool, text mining, to expand the scope of business intelligence.

In this paper, we will work out the strong points of text mining in extracting business intelligence from huge amount of textual information sources within business systems. We will apply text mining to each stage of Business Intelligence systems to prove that text mining is the powerful tool to expand the scope of BI. After reviewing basic definitions and some related technologies, we will discuss the relationship and the benefits of these to text mining. Some examples and applications of text mining will also be given. The motivation behind is to develop new approach to effective and efficient textual information analysis. Thus we can expand the scope of Business Intelligence using the powerful tool, text mining.

***Keywords*—**Business intelligence, document warehouse, text mining.

I. Introduction

IT is imperative to expand the scope of business intelligence to include knowledge discovery from textual information which is widespread involved in business systems in order to gain competitive advantages for the business. Traditional data mining has no power to deal with the huge amount of unstructured and semi-structured written materials based on natural languages: text. New techniques must be developed to deal with this pressing issue. It is the right time for text mining to obtain more attention [1], [2], [6], [7], [9], [10], [11].

Manuscript received 2005. This work is supported by ARC APAI Scholarship, Curtin Research Fellowship and ARC Funding within the Centre CEEBI and School of Information Systems, Curtin Business School, Curtin University of Technology, Australia. Australian Research Council (ARC) is the commonwealth research foundation of Australia.

Li Gao is with the School of Information Systems, Curtin Business School, Curtin University of Technology, GPO Box U1987, Perth, WA 6845, Australia.

Prof. Elizabeth Chang is with the School of Information Systems, Curtin Business School, Curtin University of Technology, GPO Box U1987, Perth, WA 6845, Australia.

Dr. Song Han is with the School of Information Systems, Curtin Business School, Curtin University of Technology, GPO Box U1987, Perth, WA 6845, Australia. (phone: +61-8-9266-4488; fax: +61-8-9266-3076; E-mail: song.han@cbs.curtin.edu.au).

Business Intelligence (BI) is a process for increasing the competitive advantages of a business by intelligent use of available information collection for users to make wise decision [1], [2]. It is well known that some techniques and resources such as data warehouses, multidimensional models, and ad hoc reports are related to Business Intelligence [3]. Although these techniques and resources have served us well, they do not totally cover the full scope of business intelligence [3].

Data in major business are complex in nature and often, poorly organized and exist in deferent formats [4]. Business executives collect consumer responses from telephone, mail, online survey and e-mail. They scan data from retail sales, marketing plans, credit cards records and competitor's press. They store business transactions, status memo, and legal briefs and log text form focus groups, online bulletin boards, and user groups. Business data collections grow larger every day. To date, data mining performs well analyzing numeric and short character string information within the business systems. However, this so-called structured data excludes the most widespread format for expressing information and knowledge: text. Some texts are critical for the business systems, such as project status information, marketing reports, details of industry regulations, competitors' advertising strategies, and descriptions of new technologies in patent application. There is just too much information. The most pressing issue is to draw the potential business intelligence from text. Furthermore, it is impossible to ignore the business intelligence value of text with the bread and deep development of information [1], [2].

Data mining is well known that it has performed well in drawing business intelligence out from various collected structured internal data by discovering useful information and knowledge and predicting the overall trends to help users make wise decision [5], [6]. Yet, it does not seem to have obvious impact on analyzing textual information within the business systems. Text Mining is the expanding technology to discover useful knowledge from huge amount of information in forms of text [7]. Data mining has already provided deep foundation and many powerful techniques for developing new technology, text mining. Data mining and text mining are something similar, but they have some notable differences.

In this paper, we will describe the power of using text mining to expand the scope of the business intelligence and discuss the processes of text mining dealing with textual information.

The organization of the rest of the paper is as follows: In Section 2, we will work out the sketch of the strong points of

text mining in expanding the scope of Business Intelligence. Section 3 will discuss the document warehouses that offer text mining the efficient repository. Metadata that drives the document warehouse will be discussed in Section 4. After some basic technologies for text mining discussed in Section 5, Section 6 will address the potential and strong points of text mining that can be used in business intelligence systems. Section 7 concludes this paper.

II. Powerful Tool to Expand BI - Text Mining

With the widespread inclusion of document, especially text, in the business systems, business executives can not get useful details from the large collection of unstructured and semi-structured written materials based on natural languages within our traditional business intelligence systems. It is the right time to develop the powerful tool to expand the scope of business intelligence to gain more competitive advantages for the business.

Data mining has been touted to be the solution for the business intelligence. We can learn its good performance form the classical example that data mining can scan a large amount of retail sales to find the money-making purchasing patterns of the consumers to decide which products would be placed close together on shelves. For example, if a consumer buys a digital camera, he must want to buy the memory card, photo printer or photo papers along with it. A related application is automatic detection of fraud, such as in credit card usage. The bank scanner looks through huge number of credit card records to find deviations form normal spending patterns. It has very possibility to be a fraud if a credit card was used to buy a bottle of beer followed quickly by a huge amount of cash withdrawn. The bank scanner can conclude that this credit card would not be used by its lawful owner since the first purchase want to test the card if it is active.

Text mining is a variation of data mining and is a relatively new discipline. Like many new research areas, it is hard to give a generally agree-upon definition. Commonly, text mining is the discovery by computer of previously unknown knowledge in text, by automatically extracting information from different written resources. Noticeably, not like finding the ore from huge amount of rocks, the goal of text mining is to extract new, never-before encountered information, such as finding overall trends in textual data and detecting potential frauds.

Text mining can represent flexible approaches to information management, research and analysis. Thus text mining can expand the fists of data mining to the ability to deal with textual materials. The following Fig. 1 addresses the process of using text mining and related methods and techniques to extract business intelligence from multi sources of raw text information. Although there seems something like that of data mining, this process of text mining gains the extra power to extract expanding business intelligence. The details of "Steps of Text Mining" in the Fig. 1 that include some key related technologies will be provided in Section 5.

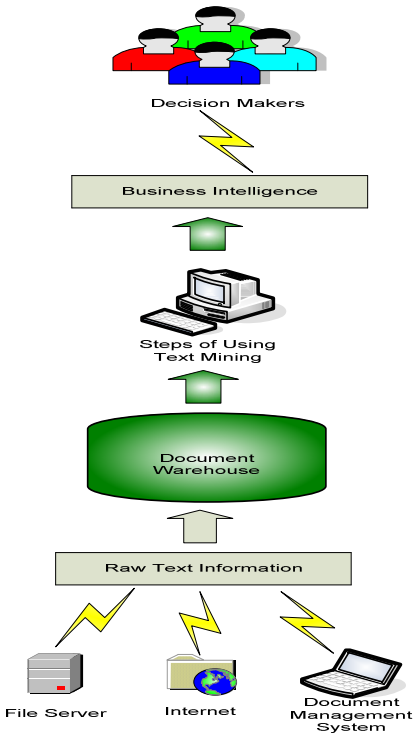

Fig. 1 The process of using text mining to extract business intelligence from multi sources of raw text information. Although there seems something like that of data mining, this process of text mining gains the extra power to extract expanding business intelligence

Text mining did not emerge from an academic vacuum but grew from a number of related technologies. The underlying technologies are based on probability theory, statistics and artificial intelligence. Cooperating with the data warehouse in data mining, document warehouse offers text mining the efficient repository that will be discussed in the next section. Moreover, we will discuss three of these related basic technologies that can offer fundamental tools for solving the problems of extracting business intelligence from text, namely Information Retrieval, Computational Linguistics, and Pattern Recognition [8], [9], [10], [11].

III. Efficient Repository for BI - Document Warehouses

The benefits of data warehousing are well understood [12]. A data warehouse provides a historical, integrated view of an organization's operations. For the most part, data warehouse focuses on internal sources. That is, most of the information is internally generated, and it describes internal processed such as sales, manufacturing, inventory management, and quality control.

Comparing with the definitions of data warehouse accepted widely within the community of data mining, we can get the four defining characters for document warehouse. They are 1) multiple document types; 2) multiple document sources; 3) to automatically draw and explicitly store the essential features of document in the document warehouse; 4) to integrate

semantically related documents. The key element in document warehousing is that document warehouse can make the information entailed in the raw text easily accessible in order to restructure transaction text to meet the needs for query and analysis.

Similar with data warehouses support data mining to analysis large volumes of numeric data, document warehouses provide text mining with efficient repositories to extract business intelligence and support decision making operations. Document warehouses are designed to store a large amount of unstructured or semi-structured written sources based on natural language, such as emails, full-text documents, HTML files, etc. The exact nature of this textual information can include complete documents, automatically generated summaries of documents, translations of documents in several languages, metadata about documents, such as author's names, publication dates, and subject keywords, automatically extracted key features, clustering information about similar documents, thematic or topical indexes. From above, we can get the core operations performed on text during document warehousing: summarization, feature extraction, clustering, categorization, and topic tracking. The following figure denotes that the basic steps in document warehouse construction.

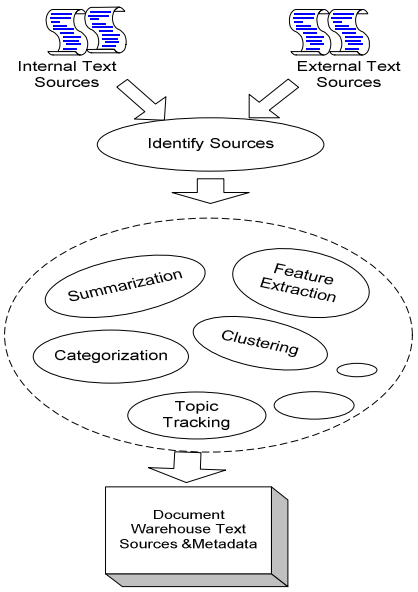

Fig. 2 The basic steps to construct the document warehouse. The core operations performed on text during document warehousing are summarization, feature extraction, clustering, categorization and topic tracking

Document warehouses distinguish themselves from data warehouses by the types of questions they are designed to answer. Data warehouses are excellent tools for answering who, what, when, where, and how much questions, but they lose their power dealing with why questions which just are the strong point of document warehouses.

The most deferent characteristic between data warehouse and document warehouse is that, in practice, data warehouse are often internally focused. We use them to better analyze the operational information of our organizations and rarely include externally sources of information. Document warehouses, though, can gather and process text from any source, internal or external, and this is the key to the ability of document warehouses to support strategic management that looks beyond the internal operation to the external factors that influence an organization. Of course, we would use external sources for data warehousing as well but the work involved in finding and acquiring relevant data in appropriate formats is many times outweighed by the marginal benefit of having the additional information.

IV. Managing Document Warehouse Metadata

Metadata drives the document warehouse. Moreover, metadata has important relations with operations in text mining. Information about document in the warehouse, technical operations, and business requirements are all embedded in metadata. The three types of metadata are content metadata, technical metadata and business metadata. The following gives the details.

A. Content Metadata

Content metadata concisely describes the key subjects of a document. Content metadata will come from at least two sources, including some documents and document management system. In both cases, the predefined metadata may not be as comprehensive as is needed in the document warehouse. In these cases, feature extraction, categorization, and summarization techniques can augment existing metadata to round out our metadata requirements.

B. Technical Metadata

Some key operations, including document searching, retrieval, and processing, are chiefly in the charge of technical metadata. Document warehouse technical metadata control both the documents to extract from document management systems and topics to search on the Internet by determining the files to copy from internal file servers. It can also control preprocessing operations, such as machine translation and format conversion, although sometimes these requirements are determined on the fly.

Moreover, technical metadata can be used to determine which operations should be applied to which documents, since not all techniques supported by document warehouses are required for every document. For example, e-mails do not require summarization. Newsgroup messages may only be useful in a broader context, such as a cluster of related documents. Controlling which text analysis operations are applied to which documents is the province of technical metadata.

C. Business Metadata

While content metadata is about what is inside of documents, and technical metadata describes how to process those documents, business metadata is less fixed. Controlling quality and user access to documents is the province of business metadata.

A new challenge is to develop excellent access control system to maintain the security and privacy of the document warehouse. Thus, it can help text mining reduce the potential risks and frauds. Access control in transactional systems is easily implemented. The need is so pervasive that relational database management systems regularly provide a robust collection of security control mechanisms to manage the creating, reading, updating, and deleting of transactional records. Data warehouse security is more complex. Users might be restricted to viewing detail data from their own regions or product lines but have access to aggregate-level measures crossing report-centers. Document warehouse access control can be more complex still. In some cases, access is based upon the source of the document. For example, notes on pending labor negotiations or legal briefs from ongoing litigation should not be publicly available in the document warehouse.

The benefit of effectively employing these types of metadata is that it can significantly ease the administration of the document warehouse. Metadata is essential to controlling the complex array of options in the document warehouse and it can improve the precision of searches.

V. Some Basic Technologies for Text Mining

Text mining is the art and technology of extracting information and knowledge from text collections stored in the structured repository, document warehouse, for conducting text mining and related business intelligence operations. The practice of text mining builds upon a number of related disciplines, in particular, Information Retrieval (IR), Computational Linguistics and Pattern Recognition. Details are given as follows.

A. Information Retrieval

Information Retrieval (IR) is the first step in text mining. Something like looking for ore from rock, the goal of Information Retrieval is to help users find documents that satisfy their information needs [9]. Something is different that the desired information is not known and coexists with many other valid pieces of information. Information Retrieval operations and document warehousing search are like a funnel, channeling documents into the warehouse without much concern for their content. An IR system acts like a first-round filter on these documents, providing them to users researching a particular problem. But it does not answer problems with volume of text that an end user must read to extract information from these documents.

Information Retrieval is a broad field with many subject areas and has developed models for representing large collections of text such that users can find documents in particular topics. The problem is up to what is currently of interest to the user and how to represent and identify documents about a particular set of topics.

The two basic representation schemes used in many IR technologies are the vector space model and latent semantic indexing. The vector space model can minimize the cost of representing the documents and the queries. It can efficiently find documents that meet the criteria of a specific query by calculating the Euclidean distance between two vectors representing the possible documents and the specific query respectively. Latent semantic indexing has been developed to balance some of the limitations of the vector space model, especially the problems of synonymy and polysemy.

Some of the greatest potential value of text mining techniques lies in spanning multiple clusters. For example, the hotel industry may tend to collect information in one cluster, while the agile industry tends to concern another, and neither is aware of the common point between them. With the aid of a text mining process model and text mining algorithms for extracting common features, readers can expand the sweep and effectiveness of text-based research.

B. Computational Linguistics

Since text mining deals with the textual information based on natural language, we can easily get the critical conflict between natural language and the limited ability of computer to understand natural language. Computers lack the human's ability to easily distinguish and apply linguistic patterns to text and overcome obstacles handling such as slang, spelling variations and contextual meaning. However, human lack the computer's ability to process text in large volumes or at high speeds. Herein lays the key to text mining: creating technology that combines a human's linguistic capabilities with the speed and accuracy of a computer.

Fortunately, there have been technological advances that have begun to close the obvious gap between natural languages and the ability of computer to process languages. The field of Computational Linguistics (also known as Natural Language Processing) has produced technologies that teach computers natural languages so that they may analyze, understand, and even generate text. Empirical computational linguistics computes statistics over large text collections in order to discover useful patterns. Within natural language processing, these patterns can be used to inform algorithms for various sub-problems, including part-speech tagging, word meaning disambiguation and bilingual dictionary creation. Some of the technologies that have been developed and can be used in the text mining processing are topic tracking, summarization, categorization, clustering, concept linkage, information visualization, and question answering.

IR information retrieval can give documents of interest to users so large that they cannot read all of them, and thus they run the risk of missing essential information. Fortunately, work in computational linguistics has provided a set of fundamental analysis tools. With these tools, we can delve into the structure of texts to extract even more targeted information.

C. Pattern Recognition

Pattern Recognition is the process of searching for predefined sequences in text. Unlike pattern matching with regular expressions in programming languages, this type of pattern recognition works with words as well as morphological and syntactic properties. Two different levels of pattern recognition are word and term matching and relevancy signatures. Word and term matching has been used successfully in the migraine research example. It is known that word and term matching is easier to implement than the other

approaches but require significant manual intervention. The approach of relevancy signatures was developed by Ellen Riloff and Wendy lehnert [13]. It builds upon both morphological and syntactic information provided by a part of speech tagger.

Based on the above technologies and document warehouse, text mining can use computer to deal with huge amount of collected raw textual information. To draw business intelligence from this text information is not a simple task. Several key steps described in the following Fig. 3 are required to fulfill the goal of text mining.

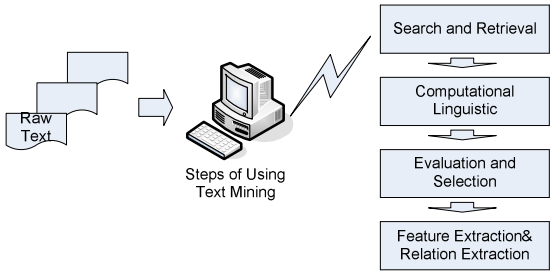

Fig. 3 To draw business intelligence from these text is not a simple task. Several key steps are required to fulfill the goal of text mining

The first step is the compilation of a document collection. The following steps are the linguistic processing of the documents in the collection, the evaluation and exploration of those documents, and then the extraction of features and relationship from a target set of documents relative to a particular business problem.

A well-known example of this is Dan Swanson's research in the 1980's that identified magnesium deficiency as a contributor to migraine headaches [14]. Swanson looked at articles with titles containing the keyword "migraine", then from those identified keywords that appeared often within the documents. One such term was "spreading depression." He then looked for titles containing "spreading depression" and repeated the process with those documents. Then, he identified "magnesium deficiency" as a key term, and hypothesized that magnesium deficiency was a factor contributing to migraine headaches. There were no direct links between the two, and no previous research had been done suggested the two were related. The hypothesis was only made from linking related documents from migraine, to spreading depression, to magnesium deficiency. The direct link between migraine headaches and magnesium deficiency was later proved accurate by actual scientific experiments, showing that Swanson's linkage methods could be a valuable process in future medical research.

VI. Using Text Mining in Each Stage of BI

Business Intelligence system is about using information wisely to gain competitive advantages in the business. It is fundamentally concerned with improving the effectiveness and efficiency of knowledge management and decision support. We can learn the key stages of Business Intelligence systems described in the following Fig. 4.

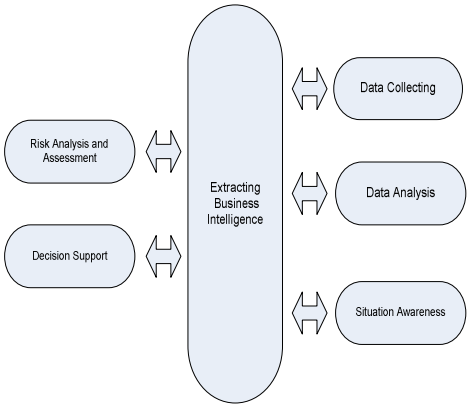

Fig. 4 The key stages of Business Intelligence systems. In each stage, text mining and related document warehouse can be used to expand the functions of Business Intelligence

In each stage, text mining and related document warehouse can be used to expand the functions of Business intelligence. In the first stage, data collecting, traditional BI system collects data within the business systems. Since text mining and related document warehouse can deal with internal as well as external text, BI system will gain the expanding scope of objective information. The data collecting stage improved by text mining is something like the collection step in document warehouse construction.

Text mining can deliver the impact to the second data analysis stage. Using powerful tools provided by related technologies described in Section 5, text mining can make the abilities of BI stronger to analyze and synthesis help useful knowledge from collected documents.

In the stage of situation awareness, text mining can link the useful facts and inferences and filter out irrelevant information which operations are just the strong points of text mining.

As to the risk analysis and assessment stage, text mining can identify reasonable decisions or courses of action based on the expectation of risk and reward. Text mining will perform well to discover what plausible actions might be taken, or decisions made, at different times. Text mining also devotes to weighing up the current and future risk, cost or benefit of taking one action over another, or making one decision versus another. It is about inferring and summarizing your best options or choices.

Just like traditional data mining technology can provide the decision support, text mining can employ semi-interactive software to identify good decisions and strategies in the above decision support step. Moreover, text mining can predict the future overall trends. Text mining can help use information wisely and provide warning users within BI systems of important events, such as takeovers, market changes, and poor staff performance, so that users can take preventative steps.

Text mining is designed to help users analyze and make better business decisions, to improve sales or customer satisfaction or staff morale and gain competitive advantages in

the business. In brief, text mining presents the information you need, when you need it.

VII. Conclusion

In this paper, after having reviewed basic definitions and some related technologies, we have discussed the relationship and the benefits of these to text mining. Then we have worked out the strong points of text mining in extracting business intelligence from huge amount of textual information sources within business systems. We have combined text mining with each stage of Business Intelligence systems to prove that text mining has the dramatic power to expand the scope of business intelligence.

Acknowledgment

The authors would present theirs thanks to the anonymous reviewers of the International Program Committee within this international conference.

References

[1] B. de Ville, "Microsoft Data Mining: Integrated Business Intelligence for e-Commerce and Knowledge Management", Boston: Digital Press, 2001.
[2] P. Bergeron, C. A. Hiller, "Competitive intelligence", in B. Cronin, Annual Review of Information Science and Technology, Medford, N.J.: Information Today, vol. 36, chapter 8, 2002.
[3] M. J. A. Berry, G. Linoff, "Data Mining Techniques: For Marketing, Sales, and Customer Relationship Management", Wiley Computer Publishing, 2nd edition, 2004.
[4] X. Wu, P. S. Yu, G. Piatetsky-Shapiro, "Data Mining: How Research Meets Practical Development?" Knowledge and Information Systems, vol. 5(2):248-261, 2003.
[5] D. Pyle, "Business Modeling and Data Mining", Morgan Kaufmann, San Francisco, CA, 2003.
[6] M. H. Dunham, "Data Mining-Introductory and Advanced Topics", Prentice Hall, 2005.
[7] R. P. Hart, "The Text Analysis Program", DICTION 5.0, Thousand Oaks, Calif.: Sage.
[8] H. Liu, H. Motoda, L. Yu, "Feature Extraction, Selection, and Construction", in N. Ye, editor, The Handbook of Data Mining, pp. 22-41. Lawrence Erlbaum Associates, Inc., Mahwah, NJ, 2003.
[9] R. Baeza-Yates, B. Ribeiro-Neto, "Modern Information Retrieval", Addison- Wesley Longman Publishing Company, 1999.
[10] Y. Yang, J. O. Pederson, "A comparative study on feature selection in text categorization", Morgan Kaufmann, 1997, 412-420.
[11] S. T. Dumais, "Latent Semantic Analysis", in B.Cronin (ed.), Annual Review of Information Science and Technology, vol.38, chapter 4, Medford, N.J.: Information Today, 2004, pp. 189-230.
[12] C. Date, "Introduction to Database Systems", 8th ed., Upper Saddle River, N.J.: Pearson Addison Wesley, 2003.
[13] E. Riloff, W. Lehnert, "Automatically Constructing a Dictionary for InformationExtraction Tasks," Proceedings of the Eleventh Annual Conference of Machine Learning, 25-32, 1994.
[14] D. R. Swanson, "Two medical literatures that are logically but not bibliographically connected", JASIS, vol. 38(4), 1987, 228-223.