A Concise Study on Text Mining for Business Intelligence

C.Kanakalakshmi[1], Dr.R.Manicka chezian[2]

Research Scholar, Dept. of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[1]

Associate Professor, Dept. of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi, India[2]

Abstract: Text Mining is the process of extracting interesting information or knowledge or patterns from the unstructured text that are from different sources. The amount of data stored has been enormously increasing day by day in many business organizations, so discovering patterns and trends out of massive data is a great challenge. Many organizations have large amounts of data in the form of unstructured text. For the purpose of knowledge discovery and information retrieval from such textual data text mining is used. Text mining is a kind of data mining technique responsible for retrieving valuable information from collection of text. The Business intelligence systems combine operational and historical data with analytical tools to present valuable and competitive information to business planners and decision makers. This large amount of information can lead to the development of new opportunities for the organization. The use of Text mining in extracting business intelligence is discussed in this paper.

Key Words: Text Mining, Business Intelligence, Data Warehouse, Document Warehouse, ETL.

I.   INTRODUCTION

The amount of business data that is generated has risen steadily every year and more and more types of information are being stored in unstructured or semi-structured formats. Traditional data mining has no power to deal with the huge amount of unstructured and semi-structured written materials based on natural languages. Text mining is the process of furnishing and extracting information from such unstructured data. Enterprises [2] today rely on a set of automated tools for knowledge discovery to gain business insight and intelligence.

Text Mining is an interdisciplinary field encompassing computational linguistics, statistics, and machine learning. Text Mining [8] uses complex Natural Language Processing (NLP) techniques. Business intelligence (BI) [1] is the set of techniques and tools for the transformation of raw data into meaningful and useful information for business analysis purposes. In modern businesses [3], increasing standards, automation, and technologies have led to vast amounts of data becoming available.

Data warehouse technologies have set up repositories to store this data. Improved Extract, Transform, Load (ETL)[1] tools have increased the speedy collection of data. BI technologies [1] 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.

II.   LITERATURE STUDY

Navita Kumari et al [1] proposed the work of recent articles in business intelligence and about the data that affects business and its process and the characteristics of data quality that are used in Business intelligence and the tools that are used for Business like Data mining, OLAP,

Data Warehousing . Abdul-Aziz Rashid Al-Azmi et al [2] proposed the work on how Data mining and Text mining tools are used to achieve business intelligence by finding hidden relations, and predicting future events from vast amounts of data. This uncovered knowledge helps in gaining completive advantages, better customers' relationships, and even fraud detection.

Jayanthi Ranjan et al [3] had given a effective way of integrating enterprise applications in real time by adopting Business Intelligence (BI) tools and systems and about various Business intelligence components and techniques. Vishal Gupta et al [4] proposed the technology foundations that are used in Text mining and various applications of Text mining in different business intelligence perspectives. Muhammad Obeidat et al [6] proposed a work on Business Intelligence domain and provides a few stimulating and innovate theories and practices related to the future trends and challenges of Business Intelligence as well as the surrounding technologies, such as data warehousing and cloud computing.

 K.L.Sumathy et al [8] presented the general framework for Text mining and the different areas that incorporates the importance of text mining. Palak Gupta et al [10] outline the recent ideas about how to pursue exploratory data analysis over text and the real text data mining efforts and the Business intelligence trends. Li Gao et al [11] proposed the different ways in extracting Business Intelligence from huge amount of textual information sources within business systems and how the Text mining is applied in each stage of Business Intelligence system.

III.   TEXT MINING IN BUSINESS INTELLIGENCE

Text mining [4] is the discovery by computer of previously unknown knowledge in text, by automatically extracting information from different written resources.

Text mining [5] can represent flexible approaches to information management, research and analysis. Text mining can expand the fists of data mining to the ability to deal with textual materials. Figure.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 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 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. The basic technologies for solving the problems of extracting business intelligence from text are Information Retrieval, Computational Linguistics, and Pattern Recognition [7]. Enterprises are paying more attention to Business Intelligence (BI)[6] theories and technologies to leverage their continually growing pools of data, enabling their workers and managers to make better business decisions. Particularly as most valuable business information is encoded in the unstructured text documents, including Internet web pages, specialized Text On-Line Analytical Processing (OLAP) solutions are needed to perform multi-dimensional analysis on text documents in the same way as on structured relational data. Since text mining and information retrieval are major technologies for handling text data, It is expected that the proposed architecture, which integrates information retrieval, text mining, and information extraction technologies alongside relational OLAP technologies, would make an effective platform toward total Business Intelligence.
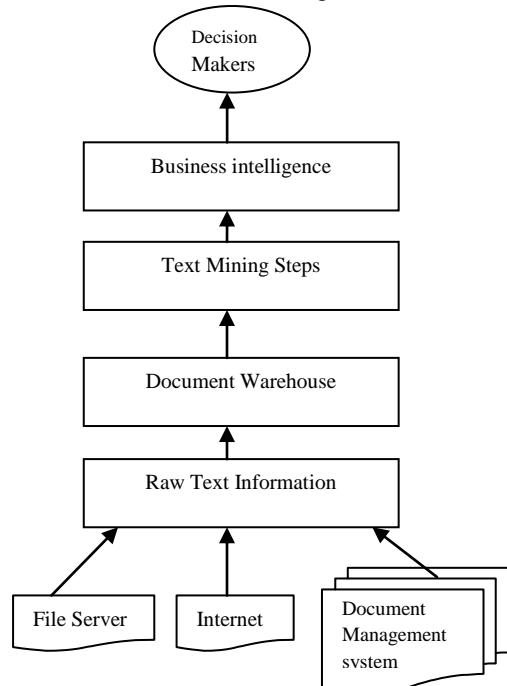

Figure 1: Text Mining in Business Intelligence

IV.   EXTRACTION TRANSFORMATION LOADING (ETL)

Extraction Transformation Loading [9] is aimed at filing unstructured textual material into categories and structured fields. In the legal documents sector the document filing and information management operations deal with the particular features of language, in which the identification and tagging of relevant elements for juridical purposes is necessary. The data can come from any source i.e., a mainframe application, an Enterprise resource Planning (ERP) application, a Customer Relationship management (CRM) tool, a flat file, and an Excel spreadsheet. All these types of data must be transformed into a single suitable format and stored in large repository called Data warehouse. To make a Data warehouse , a process known as Extraction, transformation, and Loading (ETL) [9] is followed which involves Extracting data from various outside sources, Transforming it to fit business needs, and ultimately Loading it into the data warehouse.
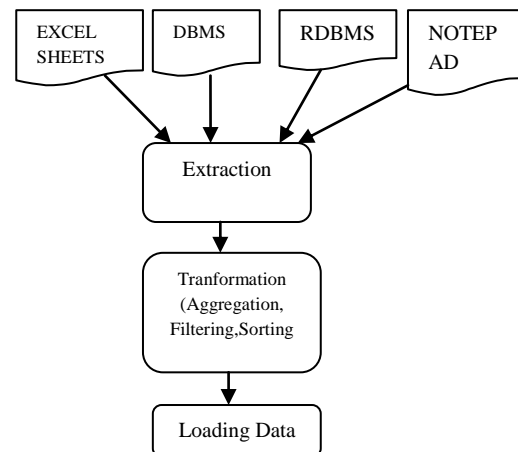

Figure 2: ETL Framework

V.   DOCUMENT WAREHOUSES

Data Warehouse [10] is actually a repository of business or enterprise databases which gives a picture of historical and current organization's operations. A data warehouse [11] 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. Document warehouse is a repository for Business Intelligence keeping variety of document types from different sources to automatically extract and store the salient features of documents. The four defining characters for document warehouse are multiple document types, multiple document sources, to automatically draw and explicitly store the essential features of document in the document warehouse and to semantically related documents.

As shown in Figure 2, the first part of an ETL process is to extract the data from various source systems. Data warehouse consolidate data from different source systems. These sources may have different formats of data. Data source formats can be relational databases and flat files,

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). So Extraction of these different format data which uses different internal representation is difficult process. Extraction tool must understand all different data storage formats. The transformation phase applies a number of rules to the extracted data so as to convert different data formats into single format. These transformation rules will be applied by transformation tool as per the requirements. In transformation selection, translation and summarization of data is done. The loading phase loads the transformed data into the data warehouse so that it can be used for various analytical purposes. Various reporting and analytical tools can be applied to data warehouse. Once data is loaded into data warehouse it cannot be updated. Loading is time consuming process so it is being done very few times.A good ETL tool should be able to communicate with 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.
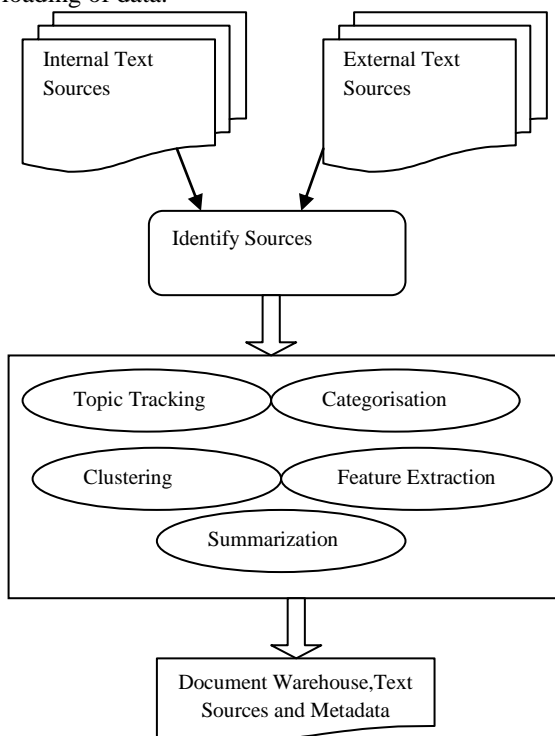

Figure 3: Document Warehouse Construction

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. Document warehouses [10] 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. Data Warehouse provides information related to what, when, who, where and how aspects but document warehouse answers the main query of users that is "why". As shown in Figure 3 Document Warehouse gathers information from both internal and external sources to enable long-term strategic management. All data collected from various sources undergo summarization, categorization, feature extraction, clustering and topic tracking to give textual data for storage in document warehouse. This textual data can be in the form of normal documents, summaries, language translations or metadata.

VI. STAGES IN TEXT MINING FOR BUSINESS INTELLIGENCE

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. Text Mining modules [10] for Business Intelligence include the following as shown in Figure 4.

1. Metadata [12]- It manages enterprise databases in form of data warehouse and document warehouse which are indexed to MS Word reports.

2. Data Extraction- It pulls reports, manually or as a scheduled task from data warehouse repositories and stores them using the scripting language as Perl that enables changes by end users.

3. Text Mining Solution- this module contains data mining reports with likelihood of frauds and thus gives the client a competitive advantage and technical details that are kept as secret in the corporate for current and future trading. Text mining [11] and related document warehouse can be used to expand the functions of Business intelligence. In the first stage, data collection is done. The data collecting stage is improved by text mining .The next analysis of data is done Using powerful tools provided by related technologies. 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.
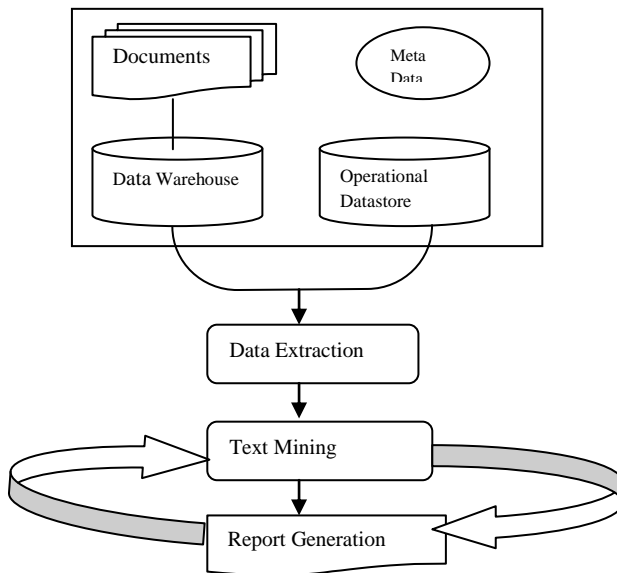

Figure 4: Text Mining Stages for BI

VII.   CONCLUSION

The most information   is stored as text; text mining is believed to have a high commercial potential value. Knowledge may be discovered from many sources of information; yet, unstructured texts remain the largest readily available source of knowledge. Text Mining technique is used to extract interesting information or knowledge from the text documents which are usually in the unstructured form. Text mining is widely used in extracting business intelligence from huge amount of textual information sources within business systems. The combined evaluation of text mining with each stage of Business Intelligence systems shows that Text Mining has the dramatic power to expand the scope of business intelligence.

REFERENCES

[1] Navita Kumari , "Business Intelligence In A Nutshell", International Journal of   Innovative Research in Computer and Communication Engineering Volume 1, Issue 4, June 2013, pp. 969-975.
[2] Abdul-Aziz Rashid Al-Azmi, "Data,   Text, And Web Mining for Business   Intelligence: A Survey", International Journal of Data Mining & Knowledge Management   Process Volume 3, Issue.2, March 2013, pp.1-21.
[3] Jayanthi Ranjan, "Business Intelligence: Concepts, Components, Techniques and Benefits", Journal of   Theoretical   and   Applied Information   Volume 9, Issue 1, November 2009, pp. 60 - 70.
[4] Vishal Gupta, Gurpreet S. Lehal, "A Survey of text mining techniques and   applications", Journal Of Emerging Technologies In Web Intelligence,   Volume 1, Issue 1, August 2009, pp. 60-76.
[5] Divya Nasa , " Text Mining Techniques -   A Survey   ", International Journal of   Advanced   Research in   Computer Science   and Software Engineering , Volume 2,   Issue 4, April 2012, pp.50-54.
[6] Muhammad Obeidat, Sarah North,   Vebol Rattanak, "Business Intelligence   Domain and Beyond", Universal Journal   of Industrial and Business Management, Volume 2, Issue 6, August 2014, pp. 127-134.
[7] S. T. Dumais, "Latent Semantic   Analysis"   ,   Annual Review of   Information Science and Technology, Volume 38, Issue 4, 2004, pp. 189-230.
[8] K.L.Sumathy, M.Chidambaram , "Text   Mining: Concepts, Applications, Tools   and Issues – An Overview", International   Journal of Computer Applications, Volume 80, Issue 4, October 2013, pp.29-32.
[9] Rajender Singh Chhillar "Extraction Transformation Loading – A Road to   Data   warehouse", 2nd National Conference Mathematical Techniques: Emerging Paradigms for Electronics and IT Industries, India, pp. 384-388.
[10] Palak Gupta1, Barkha Narang, "Role of Text Mining in Business Intelligence", Gian Jyoti E-Journal, Volume 1, Issue 2, March 2012.
[11] Li Gao, Elizabeth Chang, and Song   Han, "Powerful Tool to Expand   Business Intelligence: Text Mining", Proceedings Of World Academy Of Science, Engineering And Technology, Volume 8, October 2005.
[12] Rafi Ahmad Khan, Dr. S.M. K. Quadri, "Business Intelligence: An Integrated   Approach", Business Intelligence Journal, Volume 5, Issue 1, January 2012, pp.64-70.

BIOGRAPHIES

Dr. R.Manickachezian received his M.Sc., degree in Applied Science from P.S.G College of Technology, Coimbatore, India in 1987. He completed his M.S. degree in Software Systems from Birla Institute of Technology and Science, Pilani, Rajasthan, India and Ph D degree in Computer Science from School of Computer Science and Engineering, Bharathiar University, Coimbatore, India. He served as a Faculty of Maths and Computer Applications at P.S.G College of Technology, Coimbatore from 1987 to 1989. Presently, he has been working as an Associate Professor of Computer Science in N G M College (Autonomous), Pollachi under Bharathiar University, Coimbatore, India since 1989. He has published thirty papers in international/national journal and conferences. He is a recipient of many awards like Desha Mithra Award and Best Paper Award. His research focuses on Network Databases, Data Mining, Distributed Computing, Data Compression, Mobile Computing, Real Time Systems and Bio-Informatics.

C.Kanakalakshmi is a Research Scholar in Department of Computer Science, Nallamuthu Gounder Mahalingam College, Pollachi. She received her Master of Computer Applications (M.C.A) in 2011 from Nallamuthu Gounder Mahalingam College, Pollachi under Bharathiar University, Coimbatore. She has presented papers in International/National conferences and attended Workshop, Seminars. Her research focuses on Data Mining.