Text Mining – A Requisite for Developing Business Intelligence

Jitendra Singh Tomar

Amity University, Sector 125, Gautam Buddha Nagar, UP, Pin-201313, India

Abstract

Extracting relevant information from data repositories in the context of an event and disclosing hidden patterns from it is termed as Mining. The data repositories are enormous, capable of storing huge transacted data to support business organizations. Discovering the hidden patterns in the sea of information is getting tedious and complex in nature and possesses a challenge to the organization in planning procedures. Since business planning requires both internal as well as external recorded data, the nature of data is in form of highly unstructured text and hence for information retrieval and knowledge innovations, organizations are using text mining for salvation of textual information. The decision maker review both the historical and current information by applying analytical models over it for effective and efficient decision making to explore new opportunities and trends in business world and keep abreast with the demands and sentiments of the market. The support of text mining in business knowledge management and intelligence is discussed in the current paper.

Keywords: Mining, Text Mining, Business Intelligence, Business Planning.

1. INTRODUCTION

The data in the repositories are increasing with the business growth of the organizations. The information collected is majorly in tacit forms and could be unstructured in nature. To develop effective business policies, this information has to be mined so as to discover hidden and unknown patterns in it. The traditional mining methods are turning obsolete and incompatible with the growth of repositories and the amount of information which they hold and hence new methodologies such as Text Mining (TM) have evolved in the current time. TM is an important procedure for knowledge discovery and build up to help the business organizations for furnishing and extracting text from sea of unstructured information for effective business conduct [1]. The enterprises are suffering form information overload as the accessibility to large amount of data is within reach due to effective automation and improved standards of information technology [2]. With the growth of Artificial Intelligence (AI) and its allied techniques like NLP, the Text Mining has given new dimensions to Business Intelligence [3], which is considered to be an analytical procedure comprising of techniques to structure data and tools for better data analysis [4] through computation linguistics, statistical packages, and learning systems.

Also, with the availability of data management tools, managing huge data repositories do not pose problems to the enterprises. With improved ETL (Extract, Transform, Load) techniques been made integral parts of data repositories, prompt data collection and its analysis could be efficiently accomplished [4]. Hence, the BI technologies have become more proficient and provide the enterprises with appropriate historical and current data for strategy formulation and execution. The BI applications and tools could bring in efficacy to the organization by giving effective analytical approach, reporting models, mining inclusive of data, text, & web, and predictive & prescriptive analysis, and hence help the enterprises to keep at par with the market requirements.

2. TEXT MINING IN BUSINESS

Mining is the process of finding out unknown or hidden pattern from the information which is previously undiscovered. It's a zest behind all business policy making. If the approach is to discover and extract text from various versatile textual materials for knowledge management, then it could be termed as Text Mining [5] which provides the facility of managing information system for research and analysis [6] and help the enterprises in better business management.

The most indispensable business information is countered in the unstructured text documents that may include web pages. A tool that can perform retrieval and analysis on unstructured text documents, analogous to the way these operations are performed on structured database, is required. Since text data is processed through text mining, it is evident that text mining along with ETL and OLAP could enlarge effectual platform as a basis for Business Intelligence that could incorporate information Retrieval, Computational Linguistics, and Pattern Recognition [8]. Enterprises are bound to pay emphasis on incorporating Business Intelligence theories and technologies to manage and organize their business and decision making through the use of artificial intelligence, statistics, and probability theory. Text mining may extract textual information from multiple document repositories containing information in raw textual format. TM could be an add-on to the concept of Data Mining (DM) which enables the business organizations to buoy up business planning through Business Intelligence.

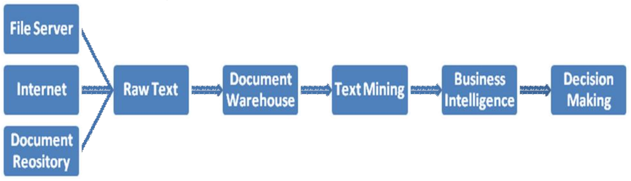

Figure 1 Text Mining in Business Intelligence

The primary goal of the text mining is to identify the useful information without duplication from various documents with synonymous understanding. Text Mining is an empirical tool that has a capacity of identifying new information that is not apparent from a document collection.

Text Mining Process includes the following Steps:

- Text pre-processing: It involves the syntactic/semantic analysis of text.
- Text Transformation, which includes the attribute generation. Two main approaches of document representation "bag of words, vector space".
- Text Representation, involves selecting a subset of features to represent a document, further reduction of dimensionality, irrelevant feature reduction, e.g. - sampling, statistics etc.
- Data Mining: It includes application of Classical data mining techniques, e.g. - Classification, Clustering etc. This is purely application dependent stage.
- Interpretation/Evaluation- Analysing results.

Text Mining can be defined as a sub field of data mining and is used to discover patterns or information from textual data. It also inherently requires techniques from other fields of information retrieval, data mining and computational linguistics. Text Mining techniques are also aimed at finding the Business Intelligence solution to help companies to remain competitive in the market.

3. Extraction Transformation Loading (ETL)

The data inflow for business organization could be from multiple sources of applications like CRM, ERP, DB, mainframes, textual files, spreadsheets, which are dynamic and versatile in nature. The formats of these versatile sources of data tend to be dissimilar. This dissimilarity is a major hurdle for managing the data coming from these sources. The first and foremost requirement is to create a Data Warehouse having unified format based on which the data from these varied sources be integrated into a single entity. To formulate such data warehouse, Extraction, Transformation, and Loading (ETL) process is adopted to extract unstructured data from varied sources, creating unified format to transforming the data to fit business needs, and finally uploading the data into unified data warehouse [9].

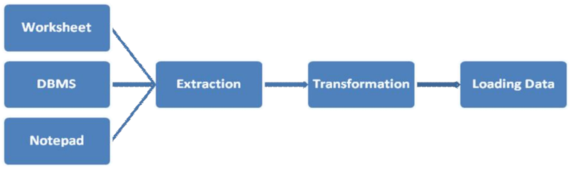

Figure 2 ETL Framework

4. Document Warehouse

Document warehouse is a unified repository with uniform structure to store document types from varied sources to extract and store the significant qualities of documents for Business Intelligence. The sources of documents could be internal as well as external to the enterprise compared to Data Warehouse which endows historical and current operations and the data is internal to the enterprise [10]. The document warehouse is entailed to extract the raw textual information to fulfill business needs as per the query in context of an event. It helps in text mining with proficient repositories for making valuable decisions and support business intelligence [11]. In addition to what, when, who, where, and how aspects, document warehouse is capable of answering 'why' which is a lag in case of data warehouse.

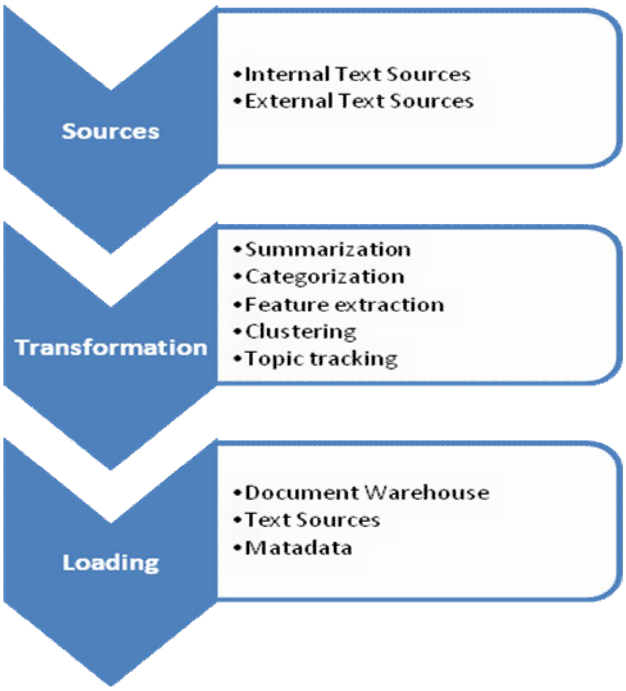

Figure 3 Document Warehouse Construction

The data from varied sources that has to be unified in document warehouse undergo summarization, categorization, feature extraction, clustering and topic tracking so that it could be stored appropriately in document warehouse.

The first and foremost action for this is that data be extracted from these sources through ETL process. These sources may have different formats ranging from RDBMS, Virtual Storage Access Method (VSAM) or Index Sequential Access Method (ISAM) and hence it is important that the extraction tool or application must be compatible with these formats and understand basic construct of these formats in order to efficiently pull out the data from these sources. A good ETL tool should be able to communicate with sources of different formats used throughout the organization.

In the second phase of transformation, the extracted data from varied sources having different format, must be converted into a unified single format for amalgamation of information collected from various sources. The three operations of selection, translation, and summarization, are applied over the data to unify it.

In the last phase of loading, the transformed data is stored in the document warehouse where it becomes non-updatable and could be used by the enterprise for formulating business strategies and develop business intelligence. Once stored, the data in document warehouse, the data could be used through various reporting and analytical tools for knowledge management and development.

The building of document warehouse is largely dependent of ETL tools which are now the integral part of enterprise application software and are quite dynamic in nature to cover up more that extraction, transformation, and loading of data.

5. TEXT MINING IN VARIOUS STAGES OF BUSINESS INTELLIGENCE

Businesses zeal for achieving competitive advantage and stay ahead in the competition by improving their decision making methodologies and effective application of knowledge in strategy planning & implementation at various stages of Business Intelligence like Data Collection, Data Analysis, Situation Awareness, Risk Analysis & Assessment, and Decision Support.

At each stage of Business Intelligence, text mining and related document repository could be used effectively and efficiently to enhance the knowledge management [10]. In the very first stage of data collection, an amalgamation of text mining and document warehouse, which stores external and internal data important for planning, will result in gaining the strategic advantage.

In the second step of Data Analysis, TM uses complex Natural Language Processing (NLP) techniques and incorporates computational linguistics, statistics, and machine learning. TM involves linguistically and semantically analysis of plain text, in realising hidden traits in the text, like frequency of use of specific words, entity extractions, and documents summarizations, and help in synthesis of useful knowledge from hoarded documents.

In the third stage of situation awareness, the prime objective is to establish a trend from historical data, establish a relationship among the data and the model that suits the dataset, and identify the relevant information in the context of event where decision making is required.

In the fourth step, Text Mining tools reinforce Business Intelligence which leads to effective Strategic Information System (SIS) to gain competitive edge and effective risk management. It is helpful in selecting the most credible action to overcome a problem at the time of decision making. It can help in the analysis of application of various policies, inferring the risks associated with each along with the positives, and advantages & drawbacks of choosing one over another. This helps the enterprise select the most appropriate strategy to gain business advantage.

In the fifth step of Business Intelligence, the quality text extracted through Text Mining has huge significance since the BI reports are based on the quality of input. Accurate the quality of text, better is the BI decision. In case, the text & data supplied to the BI applications in incorrect or inappropriate, the quality of output will suffer and will deteriorate the business operations of an enterprise.

6. CONCLUSION

TM is related field to DM, but differs in its techniques and methodologies used. It deals with textual data rather than records. TM is defined as detection of hidden patterns, traits, or unknown information from textual data that makes up huge quantity of data repository. TM involves linguistically and semantically analysis of plain text, in realizing hidden traits in the text, like frequency of use of specific words, entity extractions, and documents summarizations. It is one of the prominent mining techniques as huge amount of knowledge is stored in form of text and is largely in unstructured form. It uses complex Natural Language Processing (NLP) techniques and incorporates computational linguistics, and statistics to extract huge amount of textual information from sources within and outside the business systems and add value to each stage of Business Intelligence, thus giving the enterprises a business advantage.

References

[1] Rashid A., “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.

[2] Ranjan J., “Business Intelligence: Concepts, Components, Techniques and Benefits”, Journal of Theoretical and Applied Information Volume 9, Issue 1, November 2009, pp. 60 - 70.

[3] Sumathy K.L., Chidambaram M., “Text Mining: Concepts, Applications, Tools and Issues – An Overview”, International Journal of Computer Applications, Volume 80, Issue 4, October 2013, pp.29-32.

[4] Kumari N., “Business Intelligence in A Nutshell”, International Journal of Innovative Research in Computer and Communication Engineering Volume 1, Issue 4, June 2013, pp. 969-975.

[5] Gupta V., Lehal G., “A Survey of text mining techniques and applications”, Journal Of Emerging Technologies In Web Intelligence, Volume 1, Issue 1, August 2009, pp. 60-76.

[6] Nasa D., “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.

[7] Obeidat M., North S., Rattanak V., “Business Intelligence Domain and Beyond”, Universal Journal of Industrial and Business Management, Volume 2, Issue 6, August 2014, pp. 127-134.

[8] Dumais S.T., “Latent Semantic Analysis”, Annual Review of Information Science and Technology, Volume 38, Issue 4, 2004, pp. 189-230.

[9] Chhillar R.S., “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] Gao L., Chang E., and Han S., “Powerful Tool to Expand Business Intelligence: Text Mining”, Proceedings Of World Academy Of Science, Engineering And Technology, Volume 8, October 2005.

[11] Gupta P., Narang B., “Role of Text Mining in Business Intelligence”, Gian Jyoti E-Journal, Volume 1, Issue 2, March 2012.

[12] Khan R. A., Dr. Quadri S.M., “Business Intelligence: An Integrated Approach”, Business Intelligence Journal, Volume 5, Issue 1, January 2012, pp.64-70.

AUTHOR

Jitendra Singh Tomar is Mathematics Graduate and received his MCA degree in 2001. He is a Microsoft Certified Systems Engineer since 2002 and is working as an IS Consultant and an Academician. He has worked upon various software and network security projects in the industry. As a trainer and academician, he has conducted various MDPs and training programs for the professionals and has been an active associate for various academic activities including curriculum design. He is currently working with Amity University, UP, India, since 2006 and held imperative positions over a period of time.