Text Mining Techniques, Applications and Challenging issues

Sadia Patka
Asst. Prof., Department of Computer Science and Engineering
Anjuman College of Engineering and Technology
Nagpur, India
Sadiya.patka13@gmail.com

Nazish Khan
Asst. Prof., Department of Computer Science and Engineering
Anjuman College of Engineering and Technology
Nagpur, India
nazish07@rediffmail.com

Tasneem Hasan
Asst. Prof., Department of Computer Science and Engineering
ITM College of Engineering
Nagpur, India
tasneemh@itmnagpur.ac.in

Abstract—Text Mining has become an important research area since in today's world, the amount of stored information has been enormousl y increasing day by day. This information is generally not in structured form and so cannot be used for further processing to extract useful information. It is the requirement of fast moving and competitive world to extract and maintain meaningful information from s uch large amount of data. Text Mining is an important step of Knowledge Discovery in Text which is used to extract previously unknown and potentially useful information from unstructured textual resources. The applications of text mining have enriched the various interesting and recent fields of human life including bioinformatics, business intelligence, human resource management, security applications, competitive intelligence etc. This paper describes different text mining techniques such as information extraction, summarization, categorization, clustering, Natural language processing and etc. The objective of this paper is to describe the detail of steps involved in the overall process of text mining, its techniques and number of current and future applications. The challenging issues in text mining are also addressed in this paper which will be helpful in future research work.

Keywords- *text mining; knowledge discovery; information extraction; summarization; categorization; clustering; natural language processing*

*****

I. Introduction

Text Mining [1] is the discovery by computer of new, previously unknown information by automatically extracting potentially useful and meaningful information from unstructured textual resources. A key element is the linking together of the extracted information to form new hypotheses or new facts which are to be explored further by more conventional means of experimentation.

Text mining (TM), also known as Knowledge-Discovery in Text (KDT), Intelligent Text Analysis or Text Data Mining, refers generally to the process of extracting interesting and non-trivial information and knowledge from unstructured text. Text mining is an interdisciplinary field which draws on data mining, information retrieval, web mining, computational linguistics and natural language processing and statistics as shown in figure 1. Text mining is believed to have a high commercial potential value as most of the information (over 80%) is stored as text. Knowledge may be discovered from many sources of information yet, unstructured texts remain the largest readily available source of knowledge.

A. *Text Mining Vs. Data Mining*

Text Mining is a variation of Data Mining. In text mining patterns are extracted from natural language text however in data mining patters [2] are extracted from databases.

B. *Text Mining Vs. Web Mining*

In Text Mining, the input is in the form of unstructured text, but in Web Mining [1] web sources are structured.

The rest of the paper is organized as follows. Section II describes the process of text mining. Section III includes various Text mining techniques. Applications of text mining are discussed in Section IV. Section V includes challenging issues in text mining. Finally, Section VI presents conclusion and future work.

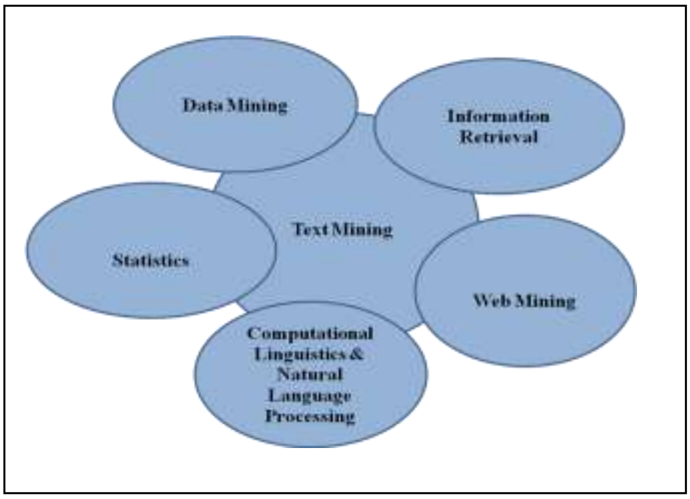

Figure1. Text Mining as Interdisciplinary Field

II. Text Mining Process

The [3] Steps involved in the overall process of text mining are as follows and depicted in fig. 2.

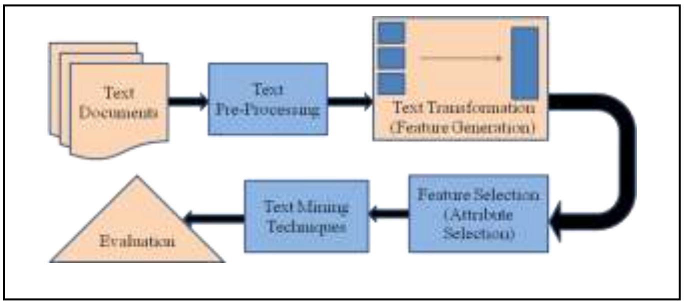
Text Documents
Text Pre-Processing
Text Transformation (Feature Generation)
Feature Selection (Attribute Selection)
Text Mining Techniques
Evaluation

Figure 2. General Text Mining Process Flow

A. *Text Preprocessing*

The text pre-processing step is further divided into three steps:

1) Tokenization: Text document includes collection of sentences. This step divide whole statement into words by removing spaces, commas etc.

2) Stop word Removal: This step involves removing of HTML, XML tags from web pages. Then process of removal of Stop words like "a", "of" etc. is performed. Finally word stemming is applied.

3) Stemming: These techniques are used to find out the root or stem of a word. In stemming words are converted to their stems. For e.g. Flying or Flew converted to Fly.

B. *Text Transformation / Feature Generation*

Text transformation means convert text document into bag of words or Vector space document model notation, which can be used for further effective analysis task.

C. *Feature Selection/Attribute Selection*

This phase removes features that are considered irrelevant for mining purpose. This procedure gives advantage of minimum search space, reduced dataset size, and less computations.

D. *Text mining methods*

As in Data mining, there is also different text mining techniques such as Information extraction, Summarization, Clustering, Classification, Concept Linkage, Information Visualization, Question Answering, Association Rule Mining and Natural Language Processing (NLP)/ Computational Linguistics are applied.

E. *Interpretation or Evaluation*

This phase performs Evaluation and Interpretation of results in terms of calculating Precision and Recall, Accuracy and measure etc.

III. TEXT MINING TECHNIQUES

This section describes various text mining techniques such as Information extraction, Summarization, Topic Tracking, Classification, Clustering, Concept Linkage, Information Visualization, Question Answering, Association Rule Mining and Natural Language Processing (NLP)/ Computational Linguistics.

A. *Information Extraction*

Information extraction (IE) technique [4] identifies key phrases and relationship within a text. For that it uses pattern matching method. In Pattern matching predefined sequences of text are matched with user text. This technique is very useful in analyzing large amount of text in text dataset. The information which is extracted by IE cannot be represented directly into a structured form therefore, post processing is required.

B. *Summarization*

This technique condenses the source text into a shorter version preserving its information. Large documents cannot be summarizes by human manually. In big research organization, researchers do not have time to read all documents so they summarize document and highlight summary with main points. Text Summarization [5] methods are classified into extractive summarization and abstractive summarization. An extractive summarization method selects important sentences; paragraphs etc. from the original document and concatenate them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in natural language. It uses linguistic methods to determine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. One of the strategies most widely used by text summarization tools, a sentence extraction extracts important sentences from an article by statistically weighting the sentences. Further heuristics such as position Information are also used for summarization.

C. *Topic Tracking*

Topic tracking [5] facilitate user by maintaining the topic searched or viewed by the user previously. System predicts about the user's next time other search documents related to previous topic very effectively. Topic detection addresses the problem of detecting new and upcoming topics in time ordered documents. The methods are frequently used to detect and monitor news tickers or news broadcasts.

D. *Classification (Categorization)*

Classification technique [6] classifies text documents into predefined class label (categories). Classification has been used in many applications like mobile messages classification [3], online customer feedback classification, business reports classification and etc. Classification and topic tracking can be integrated to classify the documents by topic and thus making the process faster.

E. *Clustering*

Clustering [6, 7, 8] is a technique used to group similar documents, but it differs from categorization in which documents are clustered on the fly instead of through the use of predefined topics. The advantage of clustering is that documents can emerge in multiple subtopics, thereby ensuring that a useful document will not be absent from search results. A basic clustering algorithm produces a vector of topics for each document and determines the weights of how well the document fits into each cluster. Clustering technique can be useful in the organization of management information systems that contain thousands of documents.

F. *Concept Linkage*

Concept linkage [5] finds related documents who share common concepts between them. The primary aim of concept linkage is to provide browsing for information rather than searching for it as in information retrieval. For example in biomedical, concept link used to link diseases and treatment. In future, Text mining can be applied as a concept linkage to discover new treatments by associating treatments that had been used in related fields.

G. *Information Visualization*

To increase the use or acquisition of knowledge, we need interactive visual representation of abstract data. Visual text mining, or information visualization [5], puts large textual sources in a visual hierarchy or map and provides browsing capabilities and simple searching. DocMiner is a tool that shows mapping of large amount of text, allowing the user to visually analyze the content. The user can visualize the document map by scaling, zooming and creating sub-maps. Information visualization is required when a user wants to narrow down a broad range of documents and explore related topics. Information visualization can be used by the government to identify terrorist networks or to find information about crimes that may have been previously thought unconnected. It could provide them a map of possible relationships between suspicious activities so that they can investigate connections that they would not have come up with on their own.

H. *Question Answering*

Many websites that are equipped with question answering technology, allow end users to “ask” question to the computer and get exact or related answer [9]. Question and Answering technique utilizes multiple text mining methods for the same. First step is the passage retrieval (PR) method. It allows passages with the highest probability of containing the answer to be retrieved, instead of simply recovering the passages sharing a subset of words with the question. Second step is Answer Extraction-aims to establish the best answer for a given question. It is based on a supervised machine-learning approach. It consists of two main modules - one for attribute extraction and the other one for answer selection.

I. *Association Rule Mining*

Association rule mining (ARM) is a technique used to discover relationships among a large set of variables in a data set [10]. It has been applied to a variety of industry settings and disciplines but has not been widely used in the social sciences, counseling, education, and other disciplines. Database containing two or more variables and their respective value, ARM determines variable value by calculating variable’s frequency. ARM used in decision making process. It discover customer purchasing pattern and find relation or associations between different products. Therefore marketing concept is clear for organization to decide product selling approach.

J. *Natural Language Processing (NLP)/ Computational Linguistics*

The Goal of NLP is to design and build a computer system that will analyze, understand and generate NLP [11]. Application includes machine translation of one human language text to another human-language text, used in fiction, robotic systems etc. Thus it is useful for enabling the use of human language for providing a summary after understanding any text document, for commands and queries understanding and analysis purpose.

IV. TEXT MINING APPLICATIONSS

Text mining has a very high commercial value since it is an emerging technology for analyzing large collection of unstructured documents for the purpose of extracting interesting and non-trivial pattern or knowledge.

There are many domain specific application of Text mining, some of the applications we had explained here:

A. *Customer Profile Analysis*

Companies use text mining [12] to draw out the occurrences and instances of key terms in large blocks of text such as articles, Web pages, complaint forums. The software converts the unstructured data formats into topic structures and semantic networks which are important data drilling tools. By the semantic network, one can learn the general tone of the complaints and reasons of complaints. It also finds common words used in complaints and their relationships to other words in the text via semantic weight.

B. *Security applications*

Many text mining software packages are marketed for security applications, like monitoring and analysis of online plain text sources such as Internet news, blogs, etc. for the purposes of national security. It is also useful in the study of text encryption/decryption.

C. *Biomedical Application*

Text Mining is used in biomedical for identification and classification of technical terms in the domain of molecular biology corresponding to concepts.

D. *Company Resource Planning*

Mining company’s reports and correspondences for activities, so its resource status and problems reported can be handled properly and future action planned can be design.

E. *Open-ended survey responses*

Analyzing a certain set of words or terms that are commonly used by respondents to describe the pros and cons of service or product, suggesting common misconceptions or confusion regarding the items in the study. Industries take the advantage of this for marketing, as per the response of customers.

F. *Competitive Intelligence (CI)*

CI facilitate the companies to organize and modify the company strategies [1] according to present market demands and the opportunities based on the information collected by the company about themselves, the market and their competitors. CI enable them manage enormous amount of data for analyzing to make plan. The goal of Competitive Intelligence is to select only relevant information by automatically reading this data. The material which has been collected is classified into categories to develop a database, and the database is analyzed to get answers to specific and crucial information for company strategies.

G. *Customer Relationship Management (CRM)*

In this domain [1] the most widespread applications are related to the management of the contents of client’s messages. This type of analysis aims at automatic rerouting of specific requests to the appropriate service or providing immediate answers to the questions that are most frequently asked. Services research has emerged as a green field area for application of advances in computer science and information Technology.

H. Technology Watch

Text mining techniques are used extensively for identification of the relevant Science and Technology literatures and extraction of the required Information from these literatures efficiently. The technological monitoring [1] analyses the characteristics of existing technologies, as well as identify emerging technologies. It is characterized by two elements – one is the capacity to identify in a non-ordinary way what already exists and that is consolidated and the other one is the capacity to identify what is already available by identifying through its potential, application fields and relationships with the existing technology.

I. Organize Repositories of document-related meta-information

To create structured metadata, Automatic text categorization methods are used. It is used for searching and retrieving relevant documents based on query.

J. Human Resource Management

Text mining techniques are used for the applications aiming at analyzing staff's opinions, to monitor the level of employee satisfaction as well as reading and storing CVs for the selection of new personnel. Often utilize to monitor the state of health of a company by means of the systematic analysis of informal documents.

V. CHALLENGING ISSUES IN TEXT MINING

The major challenging issues in text mining arise from the complexity of a natural language itself. The natural language is not free from the problem of ambiguity. Ambiguity refers to the capability of being understood in two or more possible ways or senses. Ambiguity gives flexibility and usability to natural language; therefore it cannot be entirely eliminated from the natural language. Various meanings can be obtained, as one word may have multiple meanings one phrase or sentence can be interpreted in various ways. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature for a specific domain.

On the other hand, most of the information extraction systems that involve semantic analysis exploit the simplest part of the whole spectrum of domain and task knowledge, that is to say, named entities. Information Extraction [13] does a more limited task than full text understanding. However, the growing need for information extraction application to domains such as functional genomics requires more text understanding. Named entity recognition (NER) describes an identification of entities in free text. For example, in biomedical domain, entities would be gene, protein names and drugs. NER often forms the starting point in a text mining system, which means that when the correct entities are recognized, the search for patterns and relations between entities can begin. The major problem in NER is ambiguous protein names; one protein name may refer to multiple gene products.

The work of [14] have demonstrated an effort to resolve ambiguous terms using sense-tagged corpora and unified medical language system (UMLS) with the motivation that the UMLS has been used in natural language processing applications such as information retrieval and information extraction systems. In their work, machine-learning techniques have been applied to sense-tagged corpora, in which senses (or concepts) of ambiguous terms have been most manually annotated. Sense disambiguation classifiers are then derived to determine senses (or concepts) of those ambiguous terms automatically. However, they conclude that manual annotation of a corpus is an expensive task.

Consequently [15] have extended the previous work by mining in biological named entity tagging (BNET) that identifies names mentioned in text and normalizes them with entries in biological databases. They concluded that names for genes/proteins are highly ambiguous and there are usually multiple names for the same gene or protein. Recognizing and classifying named entities in texts require knowledge on the domain entities. To tag text entities list entities are used with the relevant semantic information; however exact character strings are often not reliable enough for precise entity identification. Research work in [13] demonstrated on using possibility theory and context knowledge in resolving an ambiguous entity. The obtained results show that the approach was successful; however, the context of the texts should be defined by a user.

VI. CONCLUSION AND FUTURE WORK

Text mining, also known as 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. General Process of Text mining is described in this paper. Various text mining techniques and current successful application domains along with future work are discussed in this paper. Efforts are still required in developing systems that interpret natural language queries and automatically performs the appropriate mining operations. The major challenging issues in text mining arise from the complexity of a natural language and hence ambiguity is still the major "world problem" in text mining applications. Text mining used in security purpose like bug or roomer messages classifies on mobile station and removed. Therefore in context mobile messages classification is also required more future work in the area of text mining.

REFERENCES

[1] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications," Journal of Emerging Technologies In Web Intelligence, Vol. 1, No. 1, August 2009, pp. 60-76.

[2] Sadia Patka, M.S. Khatib and Kamlesh Kelwade, "Recent Trends and Rapid Development of Applications in Data Mining," International Conference on Advances in Engineering & Technology 2014, IOSR Journal of Computer Science, e-ISSN: 2278-0661, p-ISSN: 2278-8727, 2014, pp. 73-78.

[3] Falguni N. Patel, Neha R. Soni, "Text Mining: A Brief Survey," International Journal of Advanced Computer Research, Volume-2 Number-4 Issue-6 December-2012, pp. 243–248.

[4] Raymond J. Mooney and Un Yong Nahm, "Text Mining with Information Extraction," Multilingualism and Electronic Language Management, Proceedings of 4th International MIDP Colloquium, September 2003, Bloemfontein, South Africa, Daelemans, W., du Plessis, T., Snyman, C. and Teck, L. (Eds.), Van Schaik Pub., South Africa, 2005, pp. 141-160.

[5] Mr. Rahul Patel and Mr. Gaurav Sharma, "A survey on text mining techniques," International Journal of Engineering And Computer Science, Volume 3 Issue 5 May, 2014 pp. 5621-5625.

[6] Kapil Wankhade, Sadia Patka and Ravindra Thool, "An Overview of Intrusion Detection Based on Data Mining Techniques," 2013 IEEE, 2013 International Conference on Communication Systems and Network Technologies, pp. 626-629.

[7] Kapil Wankhade, Sadia Patka and Ravindra Thool, "An Efficient Approach for Intrusion Detection Using Data Mining Methods," 2013 IEEE, 2013 International Conference on Advances in Computing, Communications and Informatics (ICACCI), pp. 1615-1618.

[8] Sadia Patka, "Intrusion Detection model Based on Data Mining Technique," International Conference on Advances in Engineering & Technology 2014, IOSR Journal of Computer Science, e-ISSN: 2278-0661, p-ISSN: 2278-8727, 2014, pp. 34-39.

[9] Antonio Juarez-Gonzalez, Alberto Tellez-Valero and Claudia Delicia-Carral, "Using Machine Learning and Text Mining in Question Answering," Language Technologies Group, Computer Science Department, National Institute of Astrophysics, Optics and Electronics (INAOE), Mexico.2006.

[10] Dion H. Goh and Rebecca P. Ang, "An introduction to association rule mining: An application in counselling and help seeking behaviour of adolescents," Journal of Behaviour Research Methods39 (2), Singapore, 2007, pp. 259-266.

[11] Varsha C. Pande1 and Dr. A.S. Khandelwal, "A Survey Of Different Text Mining Techniques," IBMRD's Journal of Management and Research, Volume-3, Issue-1, March 2014, pp. 125-133.

[12] Shantanu Godbole and Shourya Roy, "Text to Intelligence: Building and Deploying a Text Mining Solution in the Services Industry for Customer Satisfaction Analysis", IEEE, 2008, pp. 441-448.

[13] H. M. Alfawareh and S. Jusoh, "Resolving ambiguous entity through context knowledge and fuzzy approach", International Journal on Computer Science and Engineering (IJCSE), vol. 3, no. 1, 2011, pp.410-422.

[14] H. Liu, S. B. Johnson, and C. Friedman, "Automatic resolution of ambiguous terms based on machine learning and conceptual relations in the UMLS",Journal of the American Medical Informatics Associations (JAMIA) 2002,vol.9,pp.621–636.

[15] H. Liu, Z. Hu, M. Torii, C. Wu, and C. Friedman, "Quantitative assessment of dictionary-based protein named entity tagging," Journal of the American Medical Informatics Associations (JAMIA) 2006, vol.13,pp.497–507.