ISSN:0975-9646

A Comprehensive Review of Text Mining

Tarsem Singh
Post Graduate Department of Computer Science and Applications,
GHG Khalsa College, Gurusar Sadhar (Distt. Ludhiana)

Abstract **- The amount of textual information has been immensely increasing day by day. Text Mining is a rising new field that attempts to assemble novel information from natural language text. To extract significant information from text many text mining techniques are available. In this paper several text mining techniques like text clustering, classification, information extraction, etc. are reviewed along with application areas, issues and considerations in text mining.**

Keywords **– Summarization, classification, clustering, information extraction, visualization**

I. INTRODUCTION

In the recent years, we have perceived an enormous growth of the electronic information, mostly in the form of text. For example, Internet is very important and useful information resource and a huge amount of textual information is available on web. So, discovering meaningful information from the massive data is a big challenge. Text Mining is an interdisciplinary field that draws on information retrieval, data mining, machine learning, statistics and computational linguistics. The goal of text mining is to derive high-quality information from text. Text mining is a process that employs a set of algorithms for converting unstructured text into structured data objects and the quantitative methods used to analyze these data objects.

Text Mining concerns looking for patterns in unstructured text. It is sometimes alternately referred to as *Text Data Mining* (TDM) or *Knowledge Discovery in Textual Databases* (KDT). Generally it is the process of extracting interesting and non-trivial patterns or knowledge from unstructured text documents. All the extracted information is linked together to form new facts or new hypotheses to be explored further by more conventional means of experimentation. The problem of Knowledge Discovery from Text (KDT) is to extract explicit and implicit concepts and semantic relations between concepts using Natural Language Processing (NLP) techniques. Its aim is to get insights into large quantities of text data. KDT, while deeply rooted in NLP, draws on methods from statistics, machine learning, reasoning, information extraction, knowledge management, and others for its discovery process

II. TEXT MINING PROCESS MODEL

Starting with a collection of documents, a text mining tool would retrieve a particular document and pre-process it by checking format and character sets. Then it would go through a text analysis phase, sometimes repeating techniques until information is extracted. The resulting information can be placed in a management information.

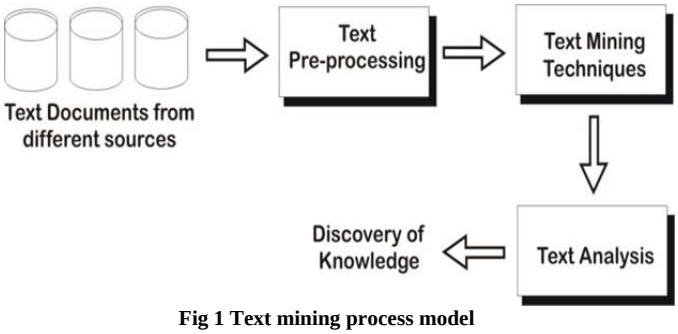

Fig 1 Text mining process model

Text Pre-processing

To apply text mining technique on pre-processed text is easy as compared to natural language text. So text pre-processing is an important task in text mining process model. The goal of document pre-processing is to represent the documents in such a way that their storage in the system and retrieval from the system are very efficient. Text pre-processing includes Tokenization, Stop Words Elimination and Stemming.

Tokenization: Tokenization is a process of chopping up a given stream of text or character sequence into words, phrases, symbols or other meaningful elements called tokens which are grouped together as a semantic unit and used as input for further processing such as parsing or text mining.

Stop Words Elimination: The most common words in text documents are articles, prepositions, and pronouns. These words are treated as stop words. Some of the examples for stop words are: the, in, a, an, with, and, as, at, be, for, from, has, he, in is, etc. The general strategy for determining a "stop list" is to sort the terms by collection frequency and then to make the most frequently used terms, as a stop list, the members of which are discarded during indexing. Stop words are removed from documents because those words are not measured as keywords in text mining applications

Stemming: Stemming is used to identify the stem/root of a word. For example, the words program, programs, programming, programmed etc. all can be stemmed to the word "**program**". This method is used to remove various suffixes, to remove the number of words, to have accurately matching stems, to save time and memory space.

Synonym Expansion: When a word has multiple meaning, synonym expansion tries to find the correct meaning of the word used in a sentence by identifying its synonyms in a given context.

Text Mining Technique
In this stage of Text Mining Process, selected algorithm is applied on text in order to process the text. The various text mining techniques are text summarization, classification, clustering, Information Extraction, Visualization could be used.

Text Analysis
In this stage the text is analysed for discovering the knowledge. Various tools can be used to analyse the text like link discovery tool.

III. TEXT MINING TECHNIQUES

Natural Language Processing (NLP)
NLP is a technology that concerns with natural language generation (NLG) and natural language understanding (NLU). NLG uses some level of underlying linguistic representation of text, to make sure that the generated text is grammatically correct and fluent. Most NLG systems include a syntactic reliazer to ensure that grammatical rules such as subject-verb agreement are obeyed, and text planner to decide how to arrange sentences, paragraph, and other parts coherently. The most well-known NLG application is machine translation system. The system analyses texts from a source language into grammatical or conceptual representations and then generates corresponding texts in the target language. NLU is a system that computes the meaning representation, essentially restricting the discussion to the domain of computational linguistic. NLU consists of at least of one the following components; tokenization, morphological or lexical analysis, syntactic analysis and semantic analysis.

Tokenization: In tokenization, a sentence is segmented into a list of tokens. The token represents a word or a special symbol such an exclamation mark.

Morphological or lexical analysis: It is a process where each word is tagged with its part of speech. The complexity arises in this process when it is possible to tag a word with more than one part of speech.

Syntactic analysis: It is a process of assigning a syntactic structure or a parse tree, to a given natural language sentence. It determines, for instance, how a sentence is broken down into phrases, how the phrases are broken down into sub-phrases, and all the way down to the actual structure of the words used.

Clustering
Text Clustering is a method based on the concept of dividing the similar text into the same cluster. Clustering can be used to classify texts or passages in natural categories that arise from statistical, lexical and semantic analysis rather than the arbitrarily pre-determined categories of traditional manual indexing systems. Once the texts are clustered on the basis of common themes, it may also be useful to correlate their divergent themes. Texts may also be clustered on the basis of length, cost, date etc. or bibliographic data such as author, institution or country of origin.

Text clustering algorithms are divided into a wide variety of different types such as agglomerative clustering algorithms, partitioning algorithms, and standard parametric modeling based methods such as the EM-algorithm. Furthermore, text representations may also be treated as strings (rather than bags of words). These different representations necessitate the design of different classes of clustering algorithms. Different clustering algorithms have different trade-offs in terms of effectiveness and efficiency

Summarization
Text summarization is to reduce the length and detail of a document while retaining most important points and general meaning. Text summarization is helpful for to figure out whether or not a lengthy document meets the user's needs and is worth reading for further information hence summary can replace the set of documents. In the time taken by the user to read the first paragraph text summarization software processes and summarizes the large text document. It is difficult to teach software to analyze semantics and to interpret meaning of text document even though computers are able to identify people, places, and time. Humans first reads entire text section to summarize then try to develop a full understanding, and then finally write a summary, highlighting its main points.

Information Extraction
Information Extraction is an automated method look for facts in the text that may represent combination of terms or patterns. It is the process of using the information to extract meaning from a text. It normally combines part of speech tagging, ontologies and regular expressions to produce a structured, machine readable file that contains essential information. There are three basic steps involved in Information Extraction:

Fact Extraction: At this stage, we are concerned with searching for individual facts contained within the document. Here the domain specific knowledge is crucial, because we can encode pattern recognition for particular facts that we know to expect in the document. Pattern Matching and Lexical Analysis are the common fact extraction techniques. In Pattern Matching technique, the process of using common regular expressions to form the lowest level of extraction, effectively constructing a bottom-up parsing of the text. In Lexical Analysis, we are concerned with breaking the text up into tokens.
Syntactic and Semantic Structure: This stage looking for noun and verb groups can be performed while looking at the local text and provide further clues to the context of word occurrences in a sentence.
Knowledge Representation: This is a trivial phase of the Information Extraction Process. In this phase the actual knowledge derived by text mining process is available for the user.

Naïve Bayes Classifier

It depends upon the probabilistic relationship between different categories. The classifier is based on *Bayes theorem*, which is stated as:

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

where H is a hypothesis to be tested and E is the evidence associated with the hypothesis. From a classification viewpoint, the hypothesis is the dependent variable and represents the predicted class. The evidence is determined by values of the input attributes. P(E | H) is the conditional probability that H is true given evidence E. P(H) is an a *priori probability*, which denotes the probability of the hypothesis before the presentation of any evidence. Naïve Bayesian is simple and efficient to implement as it assumes that all the words of the documents are independent to one another.

The requirements for the Naïve Bayes model are as follows:

- Each model must contain one numeric or text column that uniquely identifies each record. Compound keys are not allowed.
- All columns must be either discrete or discretized columns. It is important to ensure that input attributes are independent of each other.
- At least one predictable column that must contain discrete values. The values of the predictable column can be treated as input and to find relationships among the columns.

Visualization

In text mining visualization methods can improve and simplify the discovery of relevant information. To represent individual documents or groups of documents text flags are used to show document category and to show density colours are used. Visual text mining puts large textual sources in a visual hierarchy. The user can interact with the document by zooming and scaling.

.

IV. Applications Of Text Mining

Business Intelligence: One of the major concerns of any business is to minimize the amount of guess work involved in decision making and thereby reduce risk. Text mining techniques allow business professionals to extract important words/patterns from documents, group related documents together, read only summaries and drill down to the full documents as necessary, thereby saving precious time and energy.

Analysis of the Market Trends: For the growth of an organization, it is compulsory to know various situations like number of competitors, number of employees, product information, sales, purchase etc. of the competing organizations. This analysis requires enormous amount of information in which text mining techniques may help. Classification and Information Extraction techniques can be used to simplify this task.

Document Searching: Text mining is used to automatic search of large number of documents based on keywords. For example, internet search engines provide efficient access to the web pages with certain content.

Analyzing Survey Response: In survey research like marketing, it is not uncommon to include various open-ended questions related to the topic under investigation. The idea is to permit respondents to express their "views" or opinions without constraining them to particular dimensions or a particular response format. This may yield insights into customer's views and opinions that might otherwise not be discovered when relying solely on structured questionnaires designed by experts.

V. Issues And Considerations

*Synonyms and phrases***:** Synonyms, such as "sick" or "ill" or words that are used in particular phrases where they denote unique meaning can be combined for indexing. For example, "Microsoft Windows" might be such a phrase, which is a specific reference to the computer operating system.

Support for different languages: Stemming, synonyms, the letters that are permitted in words, etc. are highly language dependent operations. There, support for different languages is important.

References

[1] Jiawei Han, Micheline Kamber, Jian Pei, *Data Mining : Concepts and Technique,* 2nd edition, Morgan Kaufmann Publishers, 2006.

[2] Richard J. Roiger and Michael W. Geatz, *Data Mining : A Totorial-based Primer,* 1st edition, Pearson Education, Inc., 2003.

[3] N. P. Katariya *et al*, "Text Pre-processing for Text Mining using Side Information", International Journal of Computer Science and Mobile Applications (ijcsma.com), Vol.3 Issue. 1, January-2015, pg. 1-05

[4] S. Jusoh *et al,* "Techniques, Applications and Challenging Issue in Text Mining". International Journal of Computer Science Issues, Vol. 9, Issue 6, No 2, November 2012, pg 431-436.

[5] R. Sagayam, S.Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", International Journal of Computational Engineering Research (ijceronline.com), Vol. 2 Issue. 5, September 2012, pp.1443-1446

[6] Dr. S. Vijayarani et al , "Preprocessing Techniques for Text Mining - An Overview" International Journal of Computer Science & Communication Networks, Vol 5(1), pg. 7-16

[7] S. V. Gaikwad *et al,* "Text Mining Methods and Techniques", International Journal of Computer Applications", Vol 85, No. 17, January 2014, pg 42-45

[8] M. F. Porter, An Algorithm for Suffix Stripping, Program, vol. 14, no. 3, pp. 130-137, 1980.

[9] R. Agrawal *et al*, "A Detailed Study on Text Mining Techniques", International Journal of Soft Computing and Engineering", Vol.-2, Issue-6, January 2013, pg. 118-121.

[10] V. Gupta and G. S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Emerging Technologies in Web Intelligence", Vol-1, No.-1, August 2009, pg 60-76.

[11] D. Nasa, "Text Mining Techniques – A Survey", International Journal of Advance Research in Computer Science and Software Engineering", Vol.-2, Issue-4, April-2012, pg 50-54.