

Techniques of Text Mining: A Survey

Paramjit Kaur¹, Neena Madan²

¹Student Department of Computer Science, ²Lecturer Department of Computer Science
Guru Nanak Dev University, Regional Campus
Ladhewali Road, Jalandhar, Punjab, INDIA

Abstract: Discovering useful patterns or knowledge from textual databases is difficult task because of data available in large sets and unstructured form. This paper gives an overview of text mining. As there is need of process for text mining, so this paper focuses on text mining process and different text mining techniques. The main techniques that are discussed in this paper are natural language processing, information extraction, clustering, summarization, visualization, concept linkage and categorization, with the help of these techniques textual data can be easily mined.

Keywords: Text mining, Natural language processing, Natural language generation, Clustering, Data mining, Information extraction, Categorization.

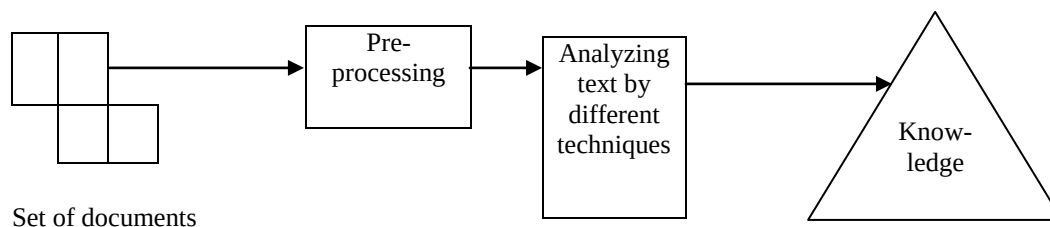
I. Introduction

Text mining is an important area in computer science which is an extension to data mining field. With the help of text mining useful information is extracted by analyzing the text, basically text mining deals with unstructured or semi structured text like text documents, emails, HTML files etc. As text mining is an extension to data mining but there is a difference between both of them. The patterns extracted with the help of data mining is from structured databases and with the help of text mining useful information is extracted from unstructured text. It is difficult to deal with such text because text used for text mining is formless and indefinite. So, text mining can also be described as looking for some useful patterns in text. Information is usually extracted from several written resources or large text documents because text is the most natural form for information storage. It helps to discover the unknown information that is some new information. Text data mining and knowledge extraction from textual databases are also used for the reference of term text mining. Extracting unknown and useful information is not a simple task. There is need of some methods which can help in discovering information. The most basic techniques for text mining are Information Extraction and Natural Language Processing.

II. Process for text mining

As extracting knowledge or patterns from any unstructured text is difficult task. So, there is a need of some kind of process which helps in this task. The process for text mining involves some steps which are as follows:

Figure 1 Text mining process



A. Collection of documents

Documents for processing the data are collected. Documents can be structured or unstructured. But in text mining unstructured documents are mainly used for mining.

B. Pre-processing of text

If documents are pre-processed before applying techniques then its mining is also easy. Firstly stop words are removed and then next is stemming in which suffixes are removed to generate word stems. Then with the help of filtering those words are removed in which there is no relevant information.

C. Apply text mining techniques

For analyzing the text different text mining techniques like clustering, classification, summarization etc are applied. Sometimes techniques are repeatedly applied until desired information is extracted.

D. Analyze the text- discovery of knowledge

Results are analyzed for the discovery of useful information and knowledge. Knowledge is discovered in the form of summary, words, clusters, objects or relationships etc [7] [9].

III. Techniques of text mining

A. Natural language processing

NLP is an important area related to human and computer interaction. Actually it concerns with two parts: Natural Language Generation and Natural Language Understanding. NLP helps computer in deriving meanings from natural language that is called NLG. Natural Language Understanding helps in representation of meaning [1]. It may consist some of the following components: Tokenization, in which a sentence is divided into tokens, where a token represents a word or a symbol. Morphological analysis is that in which a word is represented by its possible separate forms. Syntactic analysis, in which a syntactic structure that is actual structure of words is assigned to sentence by, broke them down all the way. Semantic analysis, in which syntactic structure is translated into semantic representation, that is in their precise meanings. It is actually a specified language that has expressions for real world objects, concepts and relationships [1][2].

B. Information extraction

Information extraction means extracting some useful information like entities or relationships from unstructured text or text sources. Extracting information means some particular data or information from natural language text. The important job of information extraction is to find entities from text, once this has been found then to identify the relationship between these entities. Information extraction uses unstructured or semi structured text as its input like text within natural language, tables or lists after that identified facts or events are stored into structured form. Then for discovering some new and useful knowledge different mining techniques can be applied. Information extraction is useful for large sets of text. An approach is used for IE that is called as pattern matching which is used for match the patterns against text to find the useful information [2].

C. Summarization

Summarization of text gives the summary of whole document or may be group of documents. As information is available in great amount so, it helps in reducing the large documents into small summary by focussing on the main things. It points out the main information in compressed form and large documents are replaced by small summaries. Summarization is helpful to meet the needs of users by reducing the lengthy documents which is helpful in conveying the important information. Extractive and Abstractive are the two methods used for text summarization. In extractive method, the portion of existing data is used for producing the summary like words, sentences or phrases. In abstractive method, natural language generation techniques are used for producing the summary.

D. Categorization

It is a traditional technique for information retrieval. In this method unstructured text documents are categorized on the basis of pre-defined sets of themes. Documents are assigned to one or more categories. In categorization actual information is not processed, only words are counted from the document which is considered as 'bag of words' and from that counted words main themes are identified and relationships are also determined by finding the synonyms. Documents are ranked on the basis of most matching pre-defined content. Also with the help of topic tracking most similar documents are prioritized by content weighting.

E. Visualization

Large documents can be easily mined with this method. It helps to anticipate different data sets. This method helps to visualize colors, maps or graphs etc. With the help of visual representation, extracting useful patterns from structured documents is easy. 2-D representations of document set are used with the help of vector space model. Users can easily determine the data from documents by moving them to small or large scale or by zooming it. Visualization can be easily performed with simple steps: acquire the original data, analyzing and extracting useful data from it and then creating maps or sub-maps for displaying data for the easy interaction of users. Visualization can also be done with the help of Self Organizing Maps, an unsupervised technique used by reducing the dimensions of documents.

F. Concept Linkage

As searching among large documents for information is difficult and time consuming but with the help of concept linkage information can be easily browsed. Because of the availability of large amount of data, one may not come across to important information, so this method helps to find the commonly shared concepts or information within the multiple documents. Potential link between the multiple documents can be easily detected. Concept linkage is important in biomedical field as it helps to find new treatments based on the existing treatments by linking diseases [2].

G. Clustering

It is unsupervised technique which is used for grouping similar data. It is unsupervised because there are no predefined sets. Groups or clusters are made on the basis of some similarity between the documents. Documents

in one cluster are similar to each other and documents in different clusters are different from one another. If there will be high similarity within a cluster and high difference between clusters, then cluster quality will be also high and better. The quality is measured by the ability of cluster to discover hidden patterns. For document clustering mainly two techniques of clustering are used that are agglomerative and K-means clustering. The choice of technique depends upon the desired output. K-means method is used for large data sets. In this method there are k clusters and one centroid for each cluster is defined. Each document is assigned to its closest centroid. Until the stability of centroid it is recalculated [12][13].

V. Issues in text mining

- In textual databases text used for mining is unstructured and formless which is difficult to process or mine.
- Complexity and Ambiguity is one of the main problems in text mining. Because natural language is complex to mine. Ambiguity means there are different and many meanings of one word that is one sentence can be interpreted in various ways which gives different meanings.
- The information to be processed is not as it is available. There is need of process for transformation of information to some form for processing [1][7].

VI. Conclusion

Text mining is used for mining the unstructured and indefinite data. In this paper we have discussed an overview of text mining and its process. Techniques of text mining have been also discussed. The main focus of the discussion was on techniques used for mining the textual databases. As text is available in unstructured manner so it is difficult to process it. There are also some issues in text mining due to complexity and ambiguity of natural language which are also discussed in this paper.

References

- [1] Shaidah Jusoh and Hejab M. Alfawerah, "Techniques, Applications and Challenging Issue in Text Mining", International Journal of Computer Science, Vol. 9, Issue 6, No 6, 2012.
- [2] Vishal Gupta and Gurpreet S. Lehal, "A Survey of Text Mining Techniques and Applications", Journal of Web Technologies in Web Technologies, Vol. 1, No 1, 2009.
- [3] Zakaira Elberrichi, Abdelattif Rahmoun and Mohamed Amine Bentaalah, "Using WordNet for Text Categorization", The International Arab Journal of Information Technology, Vol. 5, No 1, January 2008.
- [4] R. Sagayam, S. Srinivasan, S. Roshni, "A Survey of Text Mining: Retrieval, Extraction and Indexing Techniques", International Journal of Computational Engineering Research, Vol. 2, Issue 5, September 2012.
- [5] Rashmi Agrawal and Mridula Batra, "International Journal of Soft Computing and Engineering", ISSN: 2231-2307, Vol. 2, Issue 6, January 2013.
- [6] S. Niharika, V. Sneha Latha and D. R. Lavanya, "A Survey of Text Categorization", International Journal of Computer Trends and Technology, Vol. 3, Issue 1, 2012.
- [7] M. Sukanya and S. Biruntha, "Techniques on Text Mining", IEEE International Conference on Advanced Communication Control and Computing Technologies, 2012.
- [8] Rafeeq Al-Hashemi, "Text Summarization Extraction System (TSSES) Using Extracted Keywords", International Arab Journal of e- Technology, Vol. 1, No 4, June 2010.
- [9] Divya Nasa, "Text Mining Techniques- A Survey", International journal of Advanced Research in Computer Science and Software Engineering, Vol 2, Issue 4, April 2012.
- [10] Menaka S and Radha N, "Text Classification Using Keyword Extraction Technique", International journal of Advanced Research in Computer Science and Software Engineering, Vol 3. Issue 12, December 2013.
- [11] Ms. Sarika Y. Pabalkar, "Web Text Mining For News By Classification", International journal of Advanced Research in Computer and Communication Engineering, ISSN: 2278-1021, Vol. 1, Issue 6, August 2012.
- [12] Varsha C. Pandey and Dr. A. S. Khandelwal, "A Survey of Different Text Mining Techniques", IBMRD's Journal of Management and Research, ISSN: 2277-7830, Vol. 3, Issue 1, March 2014.
- [13] Michael Steinbach, George Karypis and Vipin Kumar, "A Comparison of Document Clustering Techniques".