A study on Text Mining Techniques and Applications

Harika Devi Kotha*, Avinash Malladi**
**Assistant professor, IFHE-FST, Hyderabad*
***Assistant professor, IFHE-FST, Hyderabad*

Abstract: Text mining is a branch of artificial intelligence. Text mining allows searching the data to extract valuable information. There are many techniques available to mine the text. Efficient text mining techniques are surveyed and compared in this paper. The methods involved in text mining are classification and clustering. Classification or categorization methods use keywords to classify the documents. Clustering uses similarity based search to find the similar documents. These two methods are providing powerful tools to mine text data. Text mining is an emerging area in computer science and is used to access large quantity of unstructured text efficiently. By using techniques like categorization, entity extraction, sentiment analysis etc, efficient text mining can be implemented. These methods are useful to extract useful information and knowledge from hidden text content.

Keywords: Text mining, entity extraction, sentiment analysis, categorization

1. Introduction

Text mining is an emerging area in computer science and is used to access large quantity of unstructured text efficiently. By using techniques like categorization, entity extraction, sentiment analysis etc, efficient text mining can be implemented. These methods are useful to extract useful information and knowledge from hidden text content. Text mining allows searching the data to extract valuable information. There are many techniques available to mine the text. Efficient text mining techniques are surveyed and compared in this paper. The methods involved in text mining are classification and clustering. Classification or categorization methods use keywords to classify the documents. Clustering uses similarity based search to find the similar documents. These two methods are providing powerful tools to mine text data.

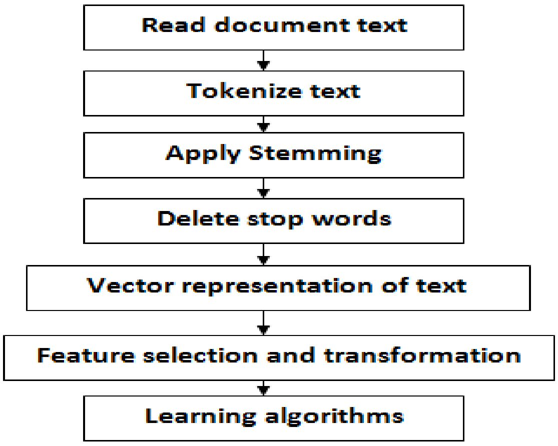
2. Text mining Techniques

By using techniques like categorization, entity extraction, sentiment analysis etc, efficient text mining can be implemented. These methods are useful to extract useful information and knowledge from hidden text content.

2.1. Categorization

This is process of assigning thematic labels for the text documents to categorize them into groups. This technique can be applied in document filtering, document indexing, meta data creation, document organization and document execution in text mining process.

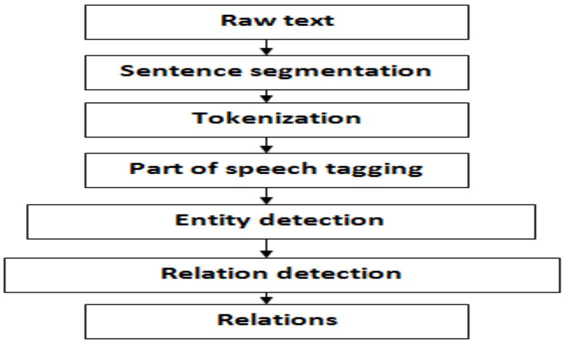
Fig.2. Categorization



2.2. Document indexing

The process of extracting information from the unstructured is called information extraction. The data which is present on the web is semi structured data or unstructured. The data which is in this format is understandable to the user because these are in natural language. Natural language processing can be done by using document indexing. Document indexing takes the following flow.

Fig.3. Document indexing



2.3. Sentiment analysis

This analysis is also known as opinion mining. This is used to systematically identify, extract, quantify, and study affective states and subjective information of text which is in natural language. The unstructured or semi structured text present on web can be analyzed or mined based on this sentiment analysis or opinion mining.

3. Applications of text mining

There are many applications of text mining. Some of the applications can be listed as: Knowledge management, Risk management, Cybercrime prevention, Customer care service, Fraud detection through claims investigation, Contextual Advertising, Business intelligence, Content enrichment and Social media data analysis.

4. Conclusion and future work:

Through techniques such as categorization, entity extraction, sentiment analysis and others, text mining extracts the useful information and knowledge hidden in text content. In the business world, this translates in being able to reveal insights, patterns and trends in even large volumes of unstructured data. In fact, it's this ability to push aside all of the non-relevant material and provide answers that is leading to its rapid adoption, especially in large organizations.

References:

[1]. Wen Zhang, A comparative study of TF_IDF, LSI and multi-words for text classification, Expert Systems with Applications 38 (2011) 2758–2765.

[2]. Shu-Hsien Liao, Data mining techniques and applications – A decade review from 2000 to 2011, Expert Systems with Applications 39 (2012) 11303–11311.

[3]. Durga Bhavani Dasari, Text Categorization and Machine Learning Methods: Current State of the Art, Global Journal of Computer Science and Technology Software & Data Engineering, Volume 12 Issue 11 Version 1.0 Year 2012

[4]. Text Mining: Concepts, Applications, Tools and Issues – An Overview, International Journal of Computer Applications (0975 – 8887) Volume 80 – No.4, October 2013

[5]. Content Analysis of Verbatim Explanations". Ppc.sas.upenn.edu. Retrieved2015-02-23, A Business Intelligence System, H.P. Luhn, IBM journal article, 1958

[6]. "Getting started in text mining", Cohen, K. Bretonnel; Hunter, Lawrence (2008). PLoS

[7]. Computational Biology 4 e20. doi: 10. 1371/journal.pcbi.0040020.

[8]. Overview and Semantic Issues of Text Mining, Anna Stavrianou, Periklis Andritsos, Nicolas Nicoloyannis, SIGMOD Record, September 2007 (Vol. 36, No. 3)

[9]. Andreas Hotho, A Brief Survey of Text Mining, May 13, 2005