Volume : 2 | Issue : 2 | february 2013 ISSN - 2250-1991

Research Paper Computer Science

Introduction of Text mining and an Analysis of Text mining Techniques

* Shah Neha K

* Research Scholar (Singhania University, Pacheri Bari, Jhunjhunu)

ABSTRACT

In Computer Science Text Mining has become an important research area. The process of getting valuable information from unstructured text is known as Text mining. Text mining [1] is similar to data mining, except that data mining tools are designed to handle structured data from databases, but text mining can work with unstructured or semi-structured data sets such as emails, full-text documents and HTML files etc. In this paper I discuss Text mining a very useful technique and analysis of different text mining techniques.

Keywords: Text mining, Information extraction, Topic tracking, Categorization, etc.

I. Introduction

A very large number of information stored in different places in unstructured form. This unstructured text can not be easily used by computer for further processing. So there is a need for some technique that is useful to extract some valuable information from unstructured text. Text Mining is the process [1] of getting new, previously unknown information by extracting information from different written resources using different techniques.

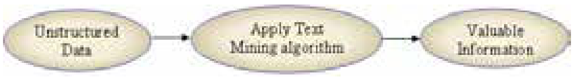

Figure 1: Process of Text Mining

Steps of Text Mining:

1. Collection of Data: Data can be collected from different written resources. These are unstructured data.
2. Structuring the input text: Collected unstructured input text converted into some structured form using any text mining algorithm.
3. Subsequent insertion into a database: Valuable information that are structured is then stored into a database.

II. Technologies and their role in text mining

Some of the technologies [1] that have been developed and can be used in the text mining process are information extraction, topic tracking, categorization, clustering, summarization, concept linkage, information visualization, and question answering. Some of them are as follows.

Information Extraction

Information Extraction (IE) is used to analyze unstructured text using computer. This technology is useful for large volume of text. A pattern matching process is used by software that is used for IE. IE identifies key terms and relationship within text. For this, predefined sequences in a text are looked by this process. The software infers the relationships between all the identified people, places, and time to provide the user with meaningful information.

In traditional data mining [2] assumption is that the information to be "mined" is already in the form of a relational database. But for many applications electronic information is available in the natural language only, it is not in the structured databases. At that time IE is useful that transform textual documents into a more structured database.

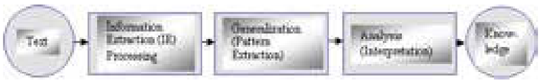
Text
Information Extraction (IE) Processing
Generalization (Pattern Extraction)
Analysis (Interpretation)
Know-ledge

Figure 2: Process of IE

Topic Tracking

Topic tracking system provides documents of user's interest to the users. This system works by keeping user profiles, based on the documents predicts other documents of interest to the user. Yahoo offers [3] a free topic tracking tool (www.alerts.yahoo.com) that allows users to choose keywords and notifies them when news relating to those topics becomes available. Topic tracking technology has limitations also. For example, if a user sets up an alert for a particular topic, he/she receive several news stories on that word but very few of them are actually on that topic. Some text mining tools available that provides the user's interested document based on her/his reading history and clicking information.

Topic tracking can be applied in many areas for example business industry, medical field and educational field. In business industry it can be used to alert company for their competitor, competitive products or any changes in the market. It is also be used in the medical field by doctors and other people looking for new treatments for illnesses and want to use newly advance technologies. It can be used by people in the field of educational field for sharing knowledge in the area of research.

Categorization

In categorization main topic of a document is identified by placing the document into a pre-defined set of topics. Categorization does not try to process the actual information but only counts words that appear and, from the counts, identifies the main topics that the document covers. When categorizing a document, a computer program will often consider the document as a collection of words. Documents having the most content on a particular topic are arranged in order and rank is given to it as per content. Categorization can be used with topic tracking to further specify the relevance of a document to a person seeking information on a topic. The documents returned from topic tracking could be ranked by content weights so that individuals could give priority to the most relevant documents first.

Categorization can be applied in many areas where communication is required between particular field and individuals.

For example [1] in business many business industry provide customer support or have to answer questions on a variety of topics from their customers. If they can use categorization schemes to classify the documents by topic, then customers or individuals will be able to access the information they seek much more readily.

Clustering

Clustering method can be used in order to find groups of documents with similar content but it is different from categorization. Text categorization [4] is a kind of "supervised" learning where the categories are known beforehand and determined in advance for each training document. In contrast, document clustering is "unsupervised" learning in which there is no predefined category or "class," but groups of documents that belong together are sought. For example, document clustering assists in retrieval by creating links between similar documents, which in turn allows related documents to be retrieved once one of the documents has been deemed relevant to a query. Another benefit of clustering is that documents can appear in multiple subtopics, thus ensuring that a useful document will not be omitted from search results.

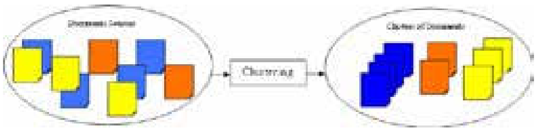

Figure 3: Process of Clustering

Clustering technology can be useful in the organization where thousands of documents are used.

Summarization

Microsoft Word's AutoSummarize function is an example of text summarization. When the document is lengthy text summarization is helpful to user for find out whether the required information is in the document or not. Text summarization software takes large text and prepares a summary for what time is used by user to read a paragraph. The main purpose of summarization is to reduce the length of a document with keeping its main points and overall meaning as it is. When humans summarize text they read whole text, understand it very well and then prepare a summary and highlights the main points. But computers do not yet have the language capabilities of humans it is difficult task for computers.

Sentence extraction tool is most widely used by text summarization. It extracts important sentences from document by statistically weighting the sentences. Summarization can be also done using position of information. For example, extraction tool may extract the sentences which are written after 'that is' or 'in conclusion' etc. that are used to indicate main topics of document. Summarization tool also search for heading or other marking that show main points of a document. Many text summarization tools allow the user to choose the percentage of the total text they want extracted as a summary. Summarization can work with topic tracking tools or categorization tools in order to summarize the documents that are retrieved on a particular topic.

Interesting topic of individuals can be quickly search using summarization tool from hundreds of documents.

Conclusion

At last I conclude that, Text mining is the process of extracting interesting and required information from large amount of unstructured text. Text mining is also known as Text Data Mining or Knowledge-Discovery in Text (KDT) [1], refers generally to the process of extracting knowledge from unstructured text. As most of the information is stored as text, text mining is required in many areas like medical, industry, education etc.

Some of Text Mining techniques are as follows:

- Information Extraction.
- Topic Tracking
- Categorization
- Clustering
- Summarization

Information Extraction technique identify facts and relations in text. This technique gives relationships between all the identified people, places, and time to provide the user with meaningful information.

Topic tracking system provides documents of user's interest to the users. The related news can be available to user using topic tracking.

In categorization main topic of a document is identified by placing the document into a pre-defined set of topics. Categorization counts words that appear, and from the counts, identifies the main topics that the document covers.

A cluster is a group of related document. Clustering method can be used in order to find groups of documents with similar content.

The main purpose of summarization is to reduce the length of a document with keeping its main points and overall meaning as it is.

REFERENCES

[1] Vishal Gupta and Gurpreet S. Lehal (2009) "A survey of text mining techniques and applications" in "Journal of emerging technologies in web intelligence, vol. 1, no. 1". | [2] Un Yong Nahm and Raymond J. Monney (2002) "Text Mining with Information Extraction" in AAAI-2002 Spring Symposium on Mining Answers from Texts and Knowledge Bases. | [3] Weiguo Fan, Linda Wallace, Stephanie Rich, and Zhongju Zhang, (2005), "Tapping into the Power of Text Mining", Journal of ACM, Blacksburg. | [4] N. Menaga and B. Hemapriya (2013) "An Efficient Concept-Based Mining Model for Enhancing Text Clustering" in International Journal of Computer Trends and Technology- volume4Issue1