Text Mining Methods and Techniques

Sonali Vijay Gaikwad
Department of Computer Engineering
Padmashree Dr. D. Y. Patil College of Engineering Pimpri Pune India

Archana Chaugule
Department of Computer Engineering
Padmashree Dr. D. Y. Patil College of Engineering Pimpri Pune India

Pramod Patil
Department of Computer Engineering
Padmashree Dr. D. Y. Patil College of Engineering Pimpri Pune India

ABSTRACT

In recent years growth of digital data is increasing, knowledge discovery and data mining have attracted great attention with coming up need for turning such data into useful information and knowledge. The use of the information and knowledge extracted from a large amount of data benefits many applications like market analysis and business management. In many applications database stores information in text form so text mining is the one of the most resent area for research. To extract user required information is the challenging issue. Text Mining is an important step of knowledge discovery process. Text mining extracts hidden information from not-structured to semi-structured data. Text mining is the discovery by automatically extracting information from different written resources and also by computer for extracting new, previously unknown information. This survey paper tries to cover the text mining techniques and methods that solve these challenges. In this survey paper we discuss such successful techniques and methods to give effectiveness over information retrieval in text mining. The types of situations where each technology may be useful in order to help users are also discussed.

Keywords
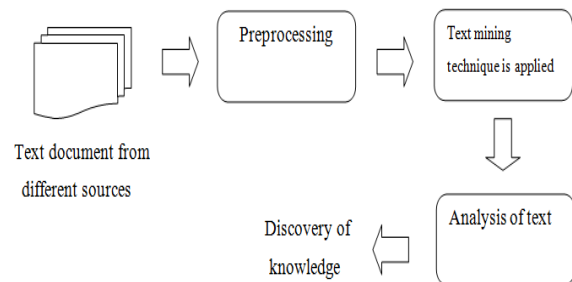Text mining process, methods, technologies

1. INTRODUCTION

Data mining technology helps to extract useful information from various databases. Data warehouses turned out to be doing well for numerical information, but unsuccessful when it came to textual information. The 21st century has taken us beyond the limited amount of information on the web. This is good in one way that more information would provide greater awareness, and better knowledge. Text data mining refers to the process of extracting interesting and non-trivial patterns or knowledge from text documents. As text mining is extraction of useful information from text data it is also known as text data mining or knowledge discovery from textual databases. It is challenging issue to find accurate knowledge in text documents to help users to find what they want.

Nowadays most of the information in business, industry, government and other institutions is stored in text form into database and this text database contains semi structured data. A document may contain some largely unstructured text components like abstract additionally few structured fields as title, name of authors, date of publication, category, and so on. Text mining is a variation on a field called data mining that tries to find interesting patterns from large databases. The great deal of studies done on the modeling and implementation of semi structured data in recent database research. On the basis of these researches information retrieval techniques such as text indexing methods have been developed to handle unstructured documents. In traditional search the user is typically look for already known terms and

has been written by someone else. The problem is in result as it is not relevant to users need. This is the goal of text mining to discover unknown information which is not known and yet not written down.

Text mining process starts with a document collection from various resources. Text mining tool would retrieve a particular document and pre-process it by checking format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to derive high quality information from text. Many text analysis techniques are available; depending on goal of organization combinations of techniques could be used. Sometimes text analysis techniques are repeated until information is extracted. The resulting information can be placed in a management information system, yielding an abundant amount of knowledge for the user of that system.

Text mining process is as shown in following fig.1

Fig. 1 Text mining process

Rest of this paper presents challenging issues, merits and demerits, methods, and techniques of text mining.

2. CHALLENGING ISSUES
Complexity of natural language is main challenging issue in text mining. The natural language is not free from the ambiguity problem. One word may have multiple meanings and multiple words can have same meaning. The capability of being understood in two or more possible ways means ambiguity. This ambiguity leads to noise in extracted information. Ambiguity cannot be entirely eliminated from the natural language as it gives flexibility and usability. There are various ways to interpret one phrase or sentence thus various meanings can be obtained. Although a number of researches have been conducted in resolving the ambiguity problem, the work is still immature and the proposed approach has been dedicated for a specific domain. It is challenge to answer what user wants as semantic meanings of many discovered words are uncertain.

Merits of Text mining:

i) The names of different entities and relationship between them can easily be found from the corpus of documents set using the technique such as information extraction.

ii) The challenging problem of managing great amount of unstructured information for extracting patterns e is solved by text mining.

Demerits of Text mining:

i) The information which is initially needed is no where written.

ii) To mine the text for information or knowledge no programs can be made in order to analyse the unstructured text directly.

3. METHODS AND MODELS USED IN TEXT MINING

Traditionally there are so many techniques developed to solve the problem of text mining that is nothing but the relevant information retrieval according to user's requirement. According to the information retrieval basically there are four methods used

1) Term Based Method (TBM).

2) Phrase Based Method (PBM).

3) Concept Based Method (CBM).

4) Pattern Taxonomy Method (PTM).

3.1 Term Based Method

Term in document is word having semantic meaning. In term based method document is analyzed on the basis of term and has advantages of efficient computational performance as well as mature theories for term weighting. These techniques are emerged over the last couple of decades from the information retrieval and machine learning communities. Term based methods suffer from the problems of polysemy and synonymy[1]. Polysemy means a word has multiple meanings and synonymy is multiple words having the same meaning. The semantic meaning of many discovered terms is uncertain for answering what users want. Information retrieval provided many term-based methods to solve this challenge.

3.2 Phrase Based Method

Phrase carries more semantics like information and is less ambiguous. In phrase based method document is analyzed on phrase basis as phrases are less ambiguous and more discriminative than individual terms[2]. The likely reasons for the daunting performance include:

1) Phrases have inferior statistical properties to terms,

2) They have low frequency of occurrence, and

3) Large numbers of redundant and noisy phrases are present among them.

3.3 Concept Based Method

In concept based terms are analyzed on sentence and document level. Text Mining techniques are mostly based on statistical analysis of word or phrase. The statistical analysis of the term frequency captures the importance of word without document. Two terms can have same frequency in same document, but the meaning is that one term contributes more appropriately than the meaning contributed by the other term[7]. The terms that capture the semantics of the text should be given more importance so, a new concept-based mining is introduced.

This model included three components. The first component analyzes the semantic structure of sentences. The second component constructs a conceptual ontological graph (COG) to describe the semantic structures and the last component extract top concepts based on the first two components to build feature vectors using the standard vector space model. Concept-based model can effectively discriminate between non important terms and meaningful terms which describe a sentence meaning [8]. The concept-based model usually relies upon natural language processing techniques. Feature selection is applied to the query concepts to optimize the representation and remove noise and ambiguity.

3.4 Pattern Taxonomy Method

In pattern taxonomy method documents are analyzed on pattern basis. Patterns can be structured into taxonomy by using is-a relation. Pattern mining has been extensively studied in data mining communities for many years. Patterns can be discovered by data mining techniques like association rule mining, frequent item set mining, sequential pattern mining and closed pattern mining[5]. Use of discovered knowledge (patterns) in the field of text mining is difficult and ineffective, because some useful long patterns with high specificity lack in support (i.e., the low-frequency problem). Not all frequent short patterns are useful hence known as misinterpretations of patterns and it leads to the ineffective performance.

In research work, an effective pattern discovery technique has been proposed to overcome the low-frequency and misinterpretation problems for text mining. The pattern based technique uses two processes pattern deploying and pattern evolving[6]. This technique refines the discovered patterns in text documents. The experimental results show that pattern based model performs better than not only other pure data mining-based methods and the concept-based model, but also term-based models.

4. TECHNIQUES USED IN TEXT MINING

To teach computers how to analyze, understand and generate text, technologies are produced by natural language processing. The technologies like information extraction, summarization, categorization, clustering and information visualization, are used in the text mining process. In the following sections we will discuss each of these technologies and the role that they play in text mining. The types of situations where each technology may be useful in order to help users are also discussed.

4.1 Information Extraction

Information extraction is initial step for computer to analyze unstructured text by identifying key phrases and relationships within text. To do this task process of pattern matching is used to look for predefined sequences in text. Information extraction task includes tokenization, identification of named entities, sentence segmentation, and part-of-speech assignment. Firstly phrases and sentences are parsed and semantically interpreted then required pieces of information entered into the database. General information extraction process is as shown in fig.2. The most accurate information extraction systems involve handcrafted language processing modules, substantial progress has been made in applying data mining techniques to a number of these steps. This technology can be very useful when dealing with large volumes of text. For many applications challenging is electronic information is
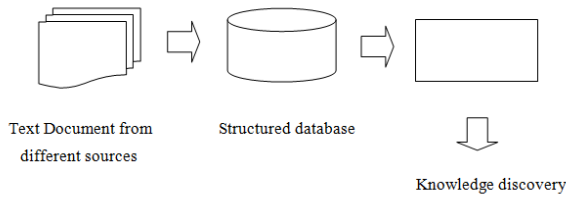
Fig.2 Information Extraction

in the form of free natural language documents rather than structured databases like relational databases. Information extraction solves this problem of transforming a corpus of textual documents into a more structured database. For further mining of knowledge database constructed by an information extraction module can be provided to the KDD module.

4.2 Categorization

Categorization automatically assigns one or more category to free text document. Categorization is supervised learning method because it is based on input output examples to classify new documents. Predefined classes are assigned to the text documents based on their content. The typical text categorization process consists of pre-processing, indexing, dimensionally reduction, and classification[3][4]. The goal of categorization is to train classifier on the basis of known examples and then unknown examples are categorized automatically. Statistical classification techniques like Naïve Bayesian classifier, Nearest Neighbour classifier, Decision Tree, and Support Vector Machines can be used to categorize text.

4.3 Clustering

Clustering method can be used in order to find groups of documents with similar content. The outcome of clustering is typically a partition called clusters *P* and each cluster consists of a number of documents *d*. The contents of the documents within one cluster are more similar and between the clusters more dissimilar then the quality of clustering is considered better. Even though clustering technique used to group similar documents it differs from categorization because in clustering documents are clustered on the fly instead of use of predefined topics. As documents can appear in multiple subtopics clustering ensures that a useful document will not be omitted from search results[7].

In data mining K-means is frequently used clustering algorithm, in text mining field also it obtains good results. A basic clustering algorithm creates a vector of topics for each document and measures the weights of how well the document fits into each cluster. The organization of management information systems makes use of clustering technology as organizational database contain thousands of documents.

4.4 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. Information visualization is applicable to government to identify terrorist networks or to find information about crimes. Following fig.3 shows steps involved in visualization process.
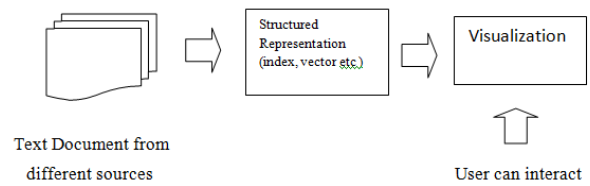
Fig.3 Visualization

The goal of information visualization divided into three steps: (1) Data preparation step includes deciding and obtaining original data of visualization and form original data space.

(2) The process of analyzing and extracting visualization data needed from original data and to form visualization data space is known as Data analysis and extraction.

(3) Visualization mapping step employ certain mapping algorithm to map visualization data space to visualization target.
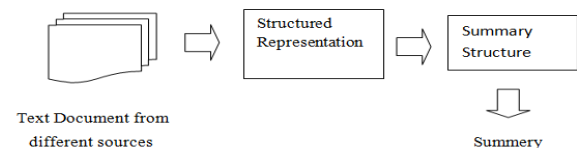
4.5 Summarization

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.

Summarization process include following steps:

(1)Pre-processing obtain a structured representation of the original text.

(2) To transform summary structure from text structure algorithm is applied in next processing step.

(3) In the invention step the final summary is obtained from the summary structure.

Fig.4 Summarization

5. CONCLUSION

This paper has presented overview techniques, methods and challenging issue in text mining. The focus has been given on fundamental methods for conducting text mining. The paper also addressed the most challenging issue in developing text mining systems. Four methods of text mining term based, phrase based, concept based and pattern taxonomy model discussed. Term based approach suffer from polysemy and synonymy while phrase based approach performs better as phrase carries more semantics like information and is less ambiguous. Two terms can have same frequency from

statistical analysis this problem can be solved by concept based approach by finding term contributing more meaning. In pattern based approach pattern taxonomy is formed to solve low frequency problem and misinterpretation problem. Then in next half Text Mining is discussed with its various techniques and usages. To extract structured information from the unstructured text Information Extraction is used. In information extraction data mining techniques can be applied for getting useful patterns or knowledge from the documents. To produce the relevant information from the corpus is known as summarization. Classification is a supervised technique because before it can be used to classify the newly arrived document it has all the input output patterns which are used to train the model. Clustering is unsupervised learning technique because no pre-defined input-output patterns are there. According to summery of documents text is clustered, this process is known as clustering. To provide improved understandable information for mining the documents Graphical Visualization is used.

6. REFERENCES

[1] G. Salton and C. Buckley, "Term-Weighting Approaches in Automatic Text Retrieval," Information Processing and Management:An Int'l J., vol. 24, no. 5, pp. 513-523, 1988.

[2] H. Ahonen, O. Heinonen, M. Klemettinen, and A.I. Verkamo, "Applying Data Mining Techniques for Descriptive Phrase Extraction in Digital Document Collections," Proc. IEEE Int'l Forum on Research and Technology Advances in Digital Libraries (ADL '98), pp. 2-11, 1998.

[3] W. Lam, M.E. Ruiz, and P. Srinivasan, "Automatic Text Categorization and Its Application to Text Retrieval," IEEE Trans. Knowledge and Data Eng., vol. 11, no. 6, pp. 865-879, Nov./Dec. 1999.

[4] H. Lodhi, C. Saunders, J. Shawe-Taylor, N. Cristianini, and C. Watkins, "Text Classification Using String Kernels," J. Machine Learning Research, vol. 2, pp. 419-444, 2002.

[5] S.-T. Wu, Y. Li, Y. Xu, B. Pham, and P. Chen, "Automatic Pattern- Taxonomy Extraction for Web Mining," Proc. IEEE/WIC/ACM Int'l Conf. Web Intelligence (WI '04), pp. 242-248, 2004.

[6] S.-T. Wu, Y. Li, and Y. Xu, "Deploying Approaches for Pattern Refinement in Text Mining," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1157-1161, 2006.

[7] S. Shehata, F. Karray, and M. Kamel, "Enhancing Text Clustering Using Concept-Based Mining Model," Proc. IEEE Sixth Int'l Conf. Data Mining (ICDM '06), pp. 1043-1048, 2006.

[8] S. Shehata, F. Karray, and M. Kamel, "A Concept-Based Model for Enhancing Text Categorization," Proc. 13th Int'l Conf. Knowledge Discovery and Data Mining (KDD '07), pp. 629-637, 2007