Effective Methods and Techniques in Text Mining

Akshata Raut
Department of Computer Engineering, Mumbai University,
Shree L.R. Tiwari College of Engineering and
Technology, Mira Road, India
raut.akshata26@gmail.com

Prof. Vinayak Shinde
Department of Computer Engineering, Mumbai University,
Shree L.R. Tiwari College of Engineering and Technology,
Mira Road, India
vdshinde@gmail.com

ISSN: 2321-8169

Abstract- Text Mining is an indispensable step comes under knowledge discovery process. Text mining extracts undiscloseddata from unstructured to semi-structured data. It is the discovery by automatically extracting information from various written resources. There are fewapproaches used in text mining for information retrieval and are explained with their merits and demerits. In this paper review of some novel researches relevant to mining association is discussed.

Keywords: Text mining, data mining, Knowledge Discovery from Database.

I. INTRODUCTION

To extract useful information and association from massive text data some text data mining approaches are available. Data mining isusedwhere analyzing data to find rules and patterns depict the characteristic of the data. The term "Data Mining" also known as Knowledge Discovery in Databases (KDD) is formally defined as: "the relevant extraction of fixed, previously undiscovered, and useful information from massive chunk of data" [1]. The 'mined' information is represented as a well-formed structure of the dataset, where the structureperhaps used on new data for prediction or classification. Roughly, data mining works on structured data, while text works on special characteristics and is unstructured. An unstructured data is wholly different from databases, where mining techniques are usually applied to manage structured data. Text mining works unstructured or semi-structured data sets.

Data mining techniques have progressively been studied, especially in the real-world databases. The goal of a data mining approach might be e.g. to allow a corporation either to improve marketing, sales, and customer support operations or to identify a fraudulent customer through better understanding of its customers. Data mining techniques are utilized in many fields such as marketing, manufacturing, process control, fraud detection, bioinformatics, information retrieval, adaptive hypermedia, electronic commerce and network management [2].

Text mining technique initially collect document from numerous resources. Text mining application retrieves a document, pre-processes it through checking format and character sets. Then the document insist ogo through a next stage i.e. text analysis phase. Text analysis is semantic analysis to collect high quality information from text. There are many text analysis techniques available depending on objective of organization combinations of techniques could be used. Frequently text analysis techniques are repeated until information is extracted. The processed information can be placed in system called management information, yielding sufficient amount of knowledge for the user of that system.

Text mining approachis as shown in fig.1.

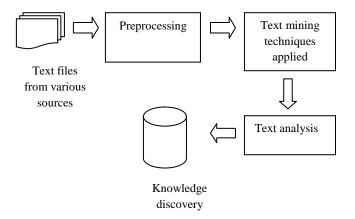


Fig 1: Text mining approach

II. RELATED WORK

Text mining is the technique which helps user tofinds useful information from massive chunksof digital text documents on the Web or databases. Therefore it is crucial to find a satisfying text mining technique that should retrieve the information which meets users' needs within a less amount of time. Traditional Information Retrieval (IR) has the same purposeto automatically retrieve suitable documents as many as possible [6].

Considering the algorithmic level, taking a common Data Mining (DM) task, namely Frequent Set Counting (FSC), as a case study, performed in depth analysis of performance problems in FSC algorithms. This conveys to create a new algorithm which will solve the FSC issue, called Direct Count and Intersect (DCI) [1].

There is a vast research on addressing automatic text categorization. For instance, in the initial work of Lewis [2], he used Bayesian independent classifiers to do categorization. He generally studies what are the effects of selection and clustering in categorization.

General data mining methods are applicable to text analysis tasks [4]. The framework follows the general knowledge discovery (KDD) process, thus containing steps from preprocessing to the utilization of the results.

The categorization is extracted fromaggregation of a model called learning model known as instance-based learning modeland an progressive information retrieval technique known as retrieval feedback. It is determined the efficiency of categorization using two legitimate information collections from the MEDLINE database. Next was an investigation for utilization of automatic categorization to retrieval of text. These experiments clearly indicated that automatic categorization enhance the retrieval execution compared with no categorization [3].

III. METHODS USED IN TEXT MINING

There are so various techniques developed for solving problems of text mining those are nothing but relevant information retrieval according to user's requirement. Basedon information retrieval techniques there are some methods explained below.

1) Term based method:

Term in document is word having well-formed meaning. In term based method document is scrutinized on the basis of term and has benefits of productive computational performance as well as well understood theories for term weighting. These techniques are developed over few decades from the information retrieval and machine learning association. This methodhas disadvantages such as polysemy and synonymy. Polysemy means a word have multiple meanings and synonymy is multiple words having the same meaning. The allowable meaning of many discovered terms is ambiguous for answering what users want. Information retrieval approach provides many term-based methods to solve raised challenge.

2) Phrase Based Method:

Phrase gives more semantics like information and is uncertain. In this, document is estimated on phrase basis as phrases are less doubtful and more selective than individual terms. Some reasons which deter the performance:

- 1) Due tosecondaryanalytical properties to terms
- 2) Less occurrence
- 3) Massiveduplicate and noisy phrases

3) Concept Based Method:

In this method, terms are estimated on sentence and document level. Text Mining techniques are often based on analytical analysis of word or phrase. The term analytical analysiscaptures the importance of word without any document. Two terms might have same frequency in same document, but one term might contribute more appropriate meaning. A novel concept based mining is introduced to acquire the semantics of texts. This model contains three components. The first component evaluates semantic arrangement of sentences. The second component

evaluates a conceptual ontological graph (COG) which describes semantic structures and the final component extracts top concepts based on the first two components to build feature vectors by using the standard vector space model. This model has ability to separateunnecessary terms and meaningful terms which describe a meaningful sentence. It is sometime depends upon natural language processing methods. A special aspect selection is enforced on the query concepts to strengthen the representation and remove noise and ambiguity.

ISSN: 2321-8169

4) Pattern Taxonomy Method:

In pattern taxonomy, documents are evaluated on pattern basis. Patterns are constructed in taxonomy by applying is-a relation. From many years, pattern mining is been reviewedin data mining. Patterns can be detected by using data mining techniques like association rule; frequent item set mining, sequential and closed pattern mining. Use ofdetected knowledge in the field of text mining is very crucial and inefficient, because some useful long patterns with high selectivity lack in support. It is not always said that all short patterns are useful hence known as misconstructions of patterns and it leads to the ineffective performance. An efficient pattern discovery procedure has been recommended to overcome low-frequency and misconstruction problems for text mining. The pattern relatedmethod uses two mechanism pattern deploying and pattern evolving. This technique refines the discovered patterns. The pattern based model performs better than any other pure data mining-based methods.

IV. TECHNIQUES USED IN TEXT MINING

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

A. Information extraction:

Information extraction is the first step for computer to evaluate unstructured text by identifying key phrases and accordance within text. For this, pattern matching is used to look for predefined sequences in text. Information extraction includes tokenization, identification of named entities, sentence segmentation, and part-of-speech assignment. Initially phrases and sentences are parsed and semantically interpreted then required pieces of information are stored in database. General information extraction process is as shown in fig.2. This technology can be very useful when dealing with large amount of text. For many applications most challenging is electronic information which is in the form of free natural language documents rather than structured databases such as

relational databases. Information extraction is a solution for this problem of reconstructing a collection of textual documents into a more structured database. For more mining of knowledge database which is built by an information extraction method can be then submitted to the KDD module.

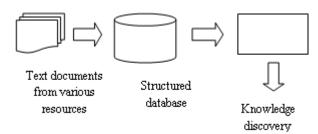


Fig 2: Information extraction process

In the accepted an draw itemsets, the transaction based algorithm discards individual items, fromthe individual item set depending on the system workload, i.e., if the system can allow 10 items from an itemset, whichisof12items,the transaction based would remove the last 2items from the itemset. These removed items from the item sets are then back to the buffer for next processing cycle.

B. Categorization:

Categorization assigns one or more category to independent text document. Categorization is a supervised learning method because it is based on input output examples to segregate new documents. Predefined classes are being assigned to the text documents based on their content. This process consists of pre-processing, indexing, dimensionally reduction, and classification. The objective to train classifier using known examples and then unknown examples are categorized automatically. Analytical classification such as Naïve Bayesian classifier, Nearest Neighbor classifier, Decision Tree, and Support Vector Machines are useful to categorize text.

C. Clustering:

Clustering method is useful to find groups of documents with similar content. As a result of clustering a partition called clusters are generated and each cluster holds a number of documents. The contents of the documents in single cluster are much similar and the contents of different clusters are dissimilar then the quality of clustering is considered better. Clustering technique is usefulto group similar documentswhich it differs from categorization because in clustering documents are clustered on the fly instead of use of predefined topics. K-means is often used clustering algorithm in data mining; in text mining field also it obtains good results. A basic clustering algorithm maintains a track of topics for every document and calculates the weightage of how well the document fits into each cluster. The management information systems uses clustering technology as organizational database contain thousands of documents.

D. Visualization:

Visualization can enhance and clarify the discovery of relevant information. For discriminating individual documents or chunks of documents text flags are used to show the category of document and to show density colors are used. Visual text mining collects huge textual sources in a visual hierarchy. The user can use the document by zooming and scaling. Information visualization is useful to government to classify terrorist networks or to find information about crimes. Following fig.3 shows steps involved in visualization process.

ISSN: 2321-8169

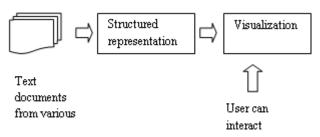


Fig 3: Visualization process

E. Summarization:

Main objective of text summarization is to reduce the length and details of a document while retaining most important points and general meaning. Text summarization is helpful for resolving whether or not a lengthy document fulfills the user's needs and whether it is worth reading for further information hence summary can be replaced by the set of documents. When user reads the first paragraph, text summarization software processes and summarizes the large text document in minimal time as compared to user. Even though computers are able to identify people, places, and time it is difficult to teach software to analyze semantics and to interpret meaning of text document. Humans first read entire text section to summarize it and then try to develop a full understanding. Then finally they makehighlights to show main points. Steps in summarization process are as follows:

- 1) Structured representation of the original text is a pre-processing step.
- 2) Algorithm is applied to translate summary structure from text structure in next processing step.
- 3) In the development step the final summary is retrieved from the summary structure.

V. CONCLUSION

This paper has presented overview of techniques and methods in text mining. Different fundamental methods have been emphasized for conducting text mining. Two terms can have same frequency from statistical analysis this problem can be solved by combined two methods in a single framework. This approach helps to mine efficient pattern and avoid unnecessary time wastage.

VI. REFERENCES

- [1] Paolo Palmerini, "On performance of data mining: from algorithms to management systems for data exploration", Technical Report, Universit'a Ca' Foscari di Venezia, 2004.
- [2] D.D. Lewis, "Feature Selection and Feature Extraction for Text Categorization," Proc. Speech and Natural Language Workshop, pp. 212-217, Arden House, 1992.
- [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.Tavel, P. 2007 Modeling and Simulation Design. AK Peters Ltd.
- [4] 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,s 1998..
- [5] 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. Brown, L. D., Hua, H., and Gao, C. 2003. A widget framework for augmented interaction in SCAPE.
- [6] D. A. Grossman and O. Frieder. Information retrieval algorithms and heuristics. Kluwer Academic publishers, 1998

ISSN: 2321-8169