International Journal of Latest Trends in Engineering and Technology
Special Issue SACAIM 2016, pp. 572-576
e-ISSN:2278-621X

A COMPARATIVE STUDY OF DIFFERENT TEXT MINING METHODS

Nihal Abdulla[1] and Abdul Rahman Hameed[2]

Abstract-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 detection by mechanically mining info from dissimilar written capitals and also by computer for mining new, formerly unknown information. This survey paper attempts to cover the text mining techniques and approaches that solve these tests. In this survey paper we deliberate such fruitful techniques and methods to give efficiency over info recovery in text mining. The types of circumstances where each skill may be valuable in order to assistance users are also deliberated. The text analytics describes a set of linguistic, statistical and machine learning techniques. Keywords – Text mining, methods, Technologies.

Keywords –*Text mining, methods, Technologies.*

I. INTRODUCTION

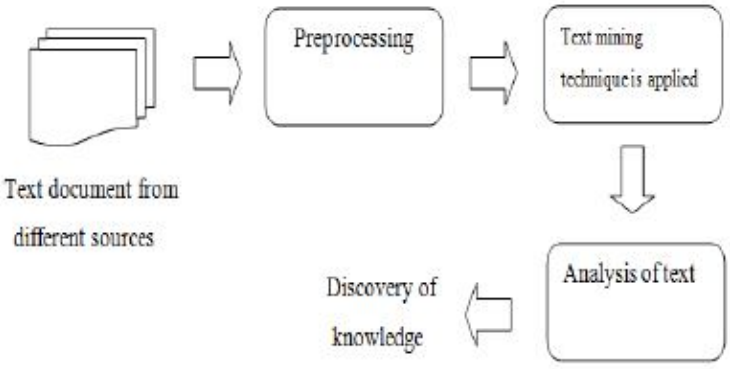


Fig.1 Text mining process

Text data mining mentions to the procedure of mining stimulating and non-trivial patterns or information from text documents. As text mining is removal of valuable info from text data it is also recognized as text data mining or knowledge discovery from textual databases. It is stimulating issue to find precise knowledge in text documents to help users to find what they want

[1] *Aloysius institute of management and Information Technology St.AloysiusCollege(AUTONOMOUS) Beeri, Kotekar, Mangalore, Karnataka, India*
[2] *Aloysius institute of management and Information Technology St.AloysiusCollege(AUTONOMOUS) Beeri, Kotekar, Mangalore, Karnataka, India*

Text mining is a difference on a field called data mining that tries to find stimulating patterns from big databases. The countless deal of studies done on the demonstrating and application of semi structured data in new database research. On the base of these researches info recovery methods such as text indexing methods have been established to switch unstructured documents. In old search the user is naturally appearance for now known terms and has been inscribed by somebody else. The problem remains in outcome as it is not applicable to user's requirement. This is the aim of text mining to learn indefinite info which is not identified and yet not inscribed down. Text mining method starts with a document group which differs in resources. Text mining tool will recover a particular document and pre-process it by testing format and character sets. Then document would go through a text analysis phase. Text analysis is semantic analysis to develop high quality info from text. Many text analysis methods are accessible; dependent on aim of organization arrangements of methods could be used. Sometimes text analysis techniques are repeated until information is extracted. The resulting info can be located in a management info system, compliant a plentiful amount of knowledge for user of that system.

II.PROBLEM STATEMENT

Compared algorithms works for simple data. So for huge complex data, concept based method is used.

III.LITERATURE SURVEY

Usually there are so many skills developed to solve the problem of text mining that is nothing but the related info recovery according to user's necessity. According to the info recovery mostly there are three methods used:

a) Term Based Method (TBM).
b) Phrase Based Method (PBM).
c) Concept Based Method (CBM).

A. Term Based Method

Term in document is word having semantic meaning. In term based method text is analyzed on the base of term and has benefits of effective computational presentation as well as developed concepts for term increment. These methods are developed over the last couple of years from the info recovery and machine learning groups. Term based methods suffer from the problems of polysemy and synonymy [1]. Polysemy means a word has various senses and synonymy is several words consuming the similar meaning. The semantic significance of many learned terms is undefined for responding what user's want. Information retrieval provided many term-based methods to solve this challenge.

B. Phrase Based Method

Phrase brings more semantics like info and is less indefinite. In phrase based method document is analyzed on phrase basis as phrases are less uncertain and further discriminative than different terms [2]. The expected reasons for the intimidating performance include:

a) Phrases have lower statistical properties to terms,
b) They have small frequency of occurrence
c) Large numbers of terminated and piercing expressions are present among them.

C. Concept Based Method

In concept based method the terms are evaluated on thebasis of sentence and document level. Text Mining methods are mainly based on overall analysis of term. The overall analysis of the term frequency acquires the significance of the word without document. In this method two

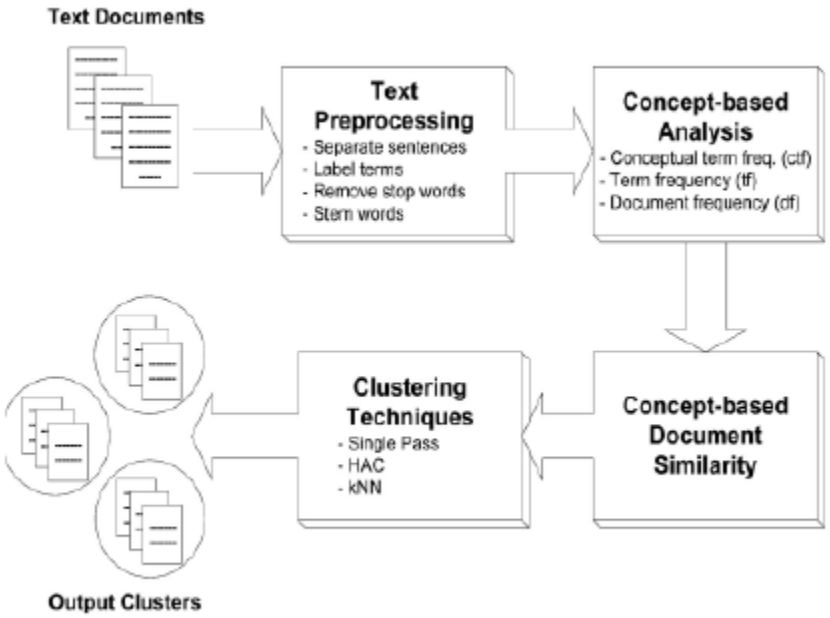
terms can have same frequency in same document. But here the significance is that one term gives more accurately than the significance given by the other term [7]. The terms that acquires the semantics of the text must be given more significance. So a new concept-based text mining is introduced in this method.This model includes:

a) Analyzation of the semantic structure of sentences.

b) Construction of a conceptual ontological graph (COG) to define the semantic structures.

IV.ANALYSIS OF CONCEPT BASED METHOD

The proposed model can efficiently find significant matching concepts between documents, according to the semantics of their sentences. The similarity between documents is calculated based on a new concept-based similarity measure. A raw text document is the input to the proposed model. Each document has well defined sentence boundaries. Each sentence in the document is labeled automatically based on the Prop Bank notations. The sentence that has many labeled verb argument structures includes many verbs associated with their arguments. The labeled verb argument structures, the output of the role labeling task, are captured and analyzed by the concept-based model on the sentence and document levels. In this model, both the verb and the argument are considered as terms. This means that this term can have more than one semantic role in the same sentence.



Concept-Based Analysis Algorithm

The concept-based analysis algorithm consists of the following steps:

1. *ddoci*is a new Document
2. L is an empty List
3. *sdoci*is a new sentence in *ddoci*
4. Build concepts list *Cdoci*from *sdoci*
5. for each concept *ci* ∈ *Ci*do
6. compute*ctfi*of *ci* in *ddoci*

7. compute*tfi*of *ci* in *ddoci*
8. compute*dfi*of *ci* in *ddoci*
9. *dk*is seen document, where
k={0,1,..., doc-1}
10. *sk*is a sentence in *dk*
11. Build concepts list *Ck*from *sk*
12. for each concept *cj*Є *Ck*do
13. if (*ci* == *cj*) then
14. update*dfi*of *ci*
15. compute*ctf*weight = avg(c*tf*i, c*tf*j)
16. add new concept matches to L
17. end if
18. end for
19. end for
20. output the matched concepts list L

The proposed concept-based analysis algorithm describes the process of calculating the *ctf*, *tf*, and *df*of the matched concepts in the documents. The procedure begins with processing a new raw document (at line 1) which has well defined sentence boundaries. The lengths of the matched concepts and their verb argument structures are stored for the concept-based similarity calculations. Each concept (in the for loop, at line 5) in the verb argument structures, representing the semantic structures of the sentence, is processed in a sequential manner. Each concept in the current document is matched with the other concepts in the previously processed documents. A concept list L is maintained to find the match to previous documents. The concept list L holds the entry for each of the previous documents that shares a concept with the current document. After processing, L contains all the matching concepts between the current document and any previous document that shares at least one concept with the new document. Finally, L is output as the list of documents with the matching concepts and the necessary information about them. The concept-based analysis algorithm is capable of matching each concept in a new document *d* with all the previously processed documents in *O(m)* time, where *m* is the number of concepts in *d*.

V.CONCLUSION

The paper presented here deals with different methods and different challenging issues in text mining.We mainly concentrated on different methods for conducting the text mining process.The text miningmethods which we discussed in this paper are text mining term based, phrase based and concept based model. Term based method onlyconcentrates on semantics. So it lacks from polysemy and synonymy. Compared to term based model, phrase based model performs well as it carries more semantics like the information and is less ambiguous. From text mining analysis we had found that two terms can have same frequency. This issue can only be solved using the concept based method. This is solved by finding the term which contributes more meaning. Henceforth Concept Based model is the best method used for text mining.

REFERENCE

[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.