ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(44), DOI: 10.17485/ijst/2016/v9i44/105075, November 2016* ISSN (Online) : 0974-5645

Data Mining Techniques for Text Mining

Gagandeep Kaur* and Hardeep Singh

Department of Computer Science and Engineering, Lovely Professional University, Phagwara – 144411, Punjab, India; gagandeepsaini7172@gmail.com, Hardeep.16869@lpu.co.in

Abstract

Semantic text mining is an abstraction of acknowledge based on the meaning. Semantic terms are explained, phrases or words. The searching terms concern their weight is computed corresponding to their synonyms, and the term which has maximum weight is at the top. The determined technique will make use of neural technique for clustering the document present to their meaning. If various words which have similar meaning are present in document then it will cluster it in similar cluster. Increment the cluster quality neural network approach, with semantic based analyzer is used popularized.

Keywords: Data Mining, Pattern Taxonomy Model, Text Clustering, Text Mining

1. Introduction

In figure 1Data mining can be expressed as an abstraction of knowledge from the large data set. This knowledge can be used for the various fields. Data mining can be used in huge way like market analysis, games, fraud detection. Data mining is fact finding for information. The association rule is based on discoverer of frequent item set. Data mining has number of techniques and numbers of algorithms are available. Data cleaning, Data selection, Data integration, Data transformation, Data mining, Pattern evaluation, Knowledge presentation these are following steps KDD.

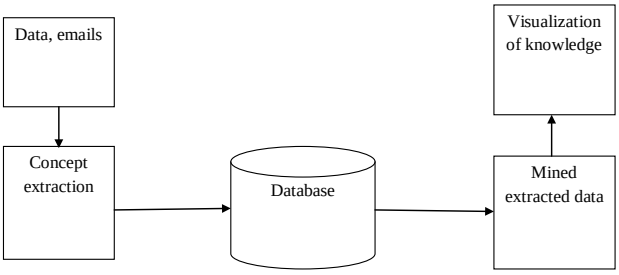

Figure 1. Concept of Data Mining.

1.1 Text Mining

Text mining is used for high quality knowledge from text documents and displays the invisible meaning. Information retrieval, abstraction, clustering, categorization are fields in text mining. Text mining used to abstract some useful knowledge from the pages which are gather to gather.

There are two types of techniques:

- Natural processing language
- Extraction of information

1.2 Text Clustering

The technique of grouping large same documents in the form of groups is called clustering. K-mean algorithm is used in text clustering (centroid-based). The advantage of clustering is that documents are classified according to their topic and subtopic, which gives the best results when searching is done. It means, the search easy. Text clustering contains four main parts: (a) text reprocessing (b) word relativity computation (c) word clustering (d) text classification.

There are four methods of text clustering:

- Partitioning method.

* *Author for correspondence*

- Hierarchical method.
- Density-based method.
- Grid- based method.

2. Literature Review

In[1] in this paper discussed the number of applications used in data mining. Data mining is used for various numbers of fields. In this paper target the various approaches for searching the new areas. Data mining used in: universities and schools, hospitals, games and domain specific. In conclusion, domain specific produce they are more correct and useful information.

In[2] in this paper based on new model that is concept-based which analysis both sentence and documents. The document analysis was based by previous model. This model has two parts: firstly, based on analysis of term and secondly, is based on measure of similarity. F measure and entropy can be measure in cluster quality. Minimum value of entropy is beneficial for the cluster quality. More the F-measure improve the quality of results.

In[3]text mining have developed into a great search field. The implicitly and explicitly approach to abstract in knowledge discovery in text. Number of applications is used in Feature extraction, Text base navigation, Clustering, Summarization, Topic tracking, Search and retrieval. In this papers in which discussed the text mining applications: text mining is used in telecommunication, health care center, research center and banks.

In[4] in this paper in which discussed the technique and challenging issues in text mining are used. The two main techniques for text mining: (a) Natural Processing language and (b) Extraction of Information. The text mining used in hospitals, government organization and they are used in business is taking the right decision. There are many favorable problems faced by text mining on the one hand natural language complexity. On the other hand, words can have many meanings but these meanings can be explained in different ways, this give arise certainty.

In[5] In this paper designed the number of mining approach for new pattern. We are used the number of patterns to make the techniques. The pattern deploying and pattern evolving are two mainly process used by this. Text mining will focus implementation for bioinformatics and it is include applying the discovered patterns for different time's series analysis.

In[6] in this paper in which discussed the number of data mining application. However we will effectively use and update discover patterns are still an open research problem. Closed sequential pattern, frequent and closed pattern are describe in this paper. Polysemy and synonymy are suffering problems in term based approach. The pattern evolving and pattern deploying are used in proposed technique.

In[7] in this paper meaning idioms based is cluster. Processing idioms, POS tagging, the documents of pre-processing, the terms of the semantic based, the terms of calculation semantic weight, semantic grammar, similarities of document, applying clustering algorithm these are steps used under in model consisting. The clustering is used to find the beneficial possible result by hierarchical algorithm. The clustering is used by chameleon algorithm. F measure and entropy is used in measuring efficiency.

In[8] in this paper we target on large data collection is discovering the patterns by deploying efficient algorithm. Extracting and non- trivial is used to refer in text mining. In which implement the number of approaches: (a) Pattern Taxonomy model, (b) Inner pattern, (c) Stemmer. Better mean exploration will be investigated by us in the long pattern. In the conclusion patterns of repetitions have been focused by us.

In[9] in this paper there are many mining application used for pattern developing. Polysemy and synonymy are main problem facing in term based approach. Recently research gives us pattern evolving and pattern deploying is useful patterns.

In[10] collection of objects is considered by us for information retrieval. One or more properties are characterizing each object associated. In literate to used measure vector similarities is wide laid by dice and jaccard. A large mass of data is basing the most of the information retrieval. Such data are defaulted by the manipulation. Classification is grouping of similar items into common classes. The classification methods are mainly used for two purposes. (a) To classify the set of index keywords. (b) To classify the documents into subject classes. The clustering process improves the search process.

In[11] text mining is used for the research areas. In this papers in which discussed the information extraction, pattern taxonomy model. Effective discovery approach has been designed to overcome the problem; the problem is low frequency and misinterpretation. This paper is

solving to cover all challenging in data mining. Pattern method is used to identify the pattern. The successful techniques are discussed in this research paper. We will make strong application for the solving problems.

3. Explain Base Paper

In[12] text mining is used for finding the effective knowledge. They are used to searching the correct information. In this paper, in which discuss the pattern taxonomy model. There are two stages in pattern taxonomy model. Firstly how to abstract the pattern in text page and secondly how to improve the effectiveness. Closed and frequent pattern are used in pattern taxonomy model. They are used to abstract and update the discovered pattern. Pattern approach is used to searching new pattern. There are main issues under this pattern based approach:

- Low frequency
- Misinterpretation

When decrease the minimum support value then occur the low frequency. Pattern mining is used for misinterpretation. Text preprocessing module is discussed in the proposed scheme. (a) Stop word removal and (b) word stemming, these two techniques are used in text preprocessing. The effective pattern discovery is defined in this paper. These techniques:

(a) Pattern evolving.

(b) Pattern deploying.

are used to improve the effectiveness. These are used for searching the interesting knowledge. In proposed approach, the pattern taxonomy model is used to abstract the pattern.

We can search the problem in this paper. The pattern discovery for text mining is selected base paper. We will be working in calculate the weight in the shortest time.

4. Problem Formulation

Clustering is the technique in which similar and dissimilar data are clustered. The clustering is generally done to analyze data in more efficient manner. In this work, algorithm for text clustering has been analyzed. In the weight based algorithm for text clustering, weights are assigned to each word in the document according to their weight age.

The main problem exists in weight based algorithm is of sorting and time. In figure 2 the proposed algorithm will not sort the member of the clusters and it takes long time to calculate weights of the words. In work, improvement will be proposed in weight based algorithm to reduce clustering time and to sort members of the cluster.

5. Flow Chart

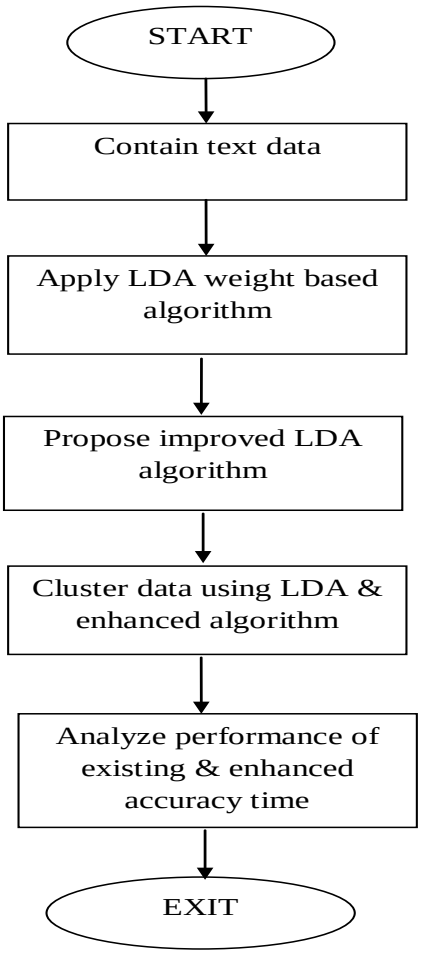

Figure 2. Proposed Technique Work Flow.

6. Expected Outcomes

The expected outcomes will improve the cluster quality after applying neural network technique with semantic based analyser; it will reduce the processing time and reduce the algorithm escape time. Taking together terms and their corresponding synonyms weight gives the better performance because it can measure the vital concept of sentence and the document. To evaluate the results, we need to improve the cluster quality using F-measure and entropy.

7. Conclusion

The proposed work will link text mining with natural language processing and minimum the gap between them. The new proposed hypothesis of semantic model will improves the cluster quality. The proposed technique will relates terms and their corresponding words to calculate their weights. The term which has maximum weight is brought to the top, with this better clustering result is obtained. Using neural network with semantic algorithms will improve the efficiency. There are number of ways to extend the work. First link it with the web documents. This technique is work only with the terms and their synonyms; it can be further improved by working with the concept of hypernyms or working on both synonyms and hypernyms. To improve the cluster quality we can used hybrid clustering. F-measure and entropy is used to improve the cluster quality.

8. Acknowledgment

I am using this opportunity to express my gratitude to everyone who supported me in the research work. First I offer my sincerest gratitude to my supervisor, Hardeep Singh, who has supported me throughout my Dissertation. Without him, this Dissertation would not have been completed or written. I am thankful for his aspiring guidance, invaluably constructive criticism and friendly advice during the research work. I am sincerely grateful to him for sharing their truthful and illuminating views on a number of issues related to the research. Finally, I thank my parents for supporting me throughout all my studies at university.

9. References

1. Padhy N, Mishra P, Panigrahi R. The Survey of Data Mining Applications and Feature Scope, International Journal of Computer Science Engineering and Information Technology, 2012; 2(3), pp.44-58.
2. Shehata S, Karray F. Enhancing Text Clustering using Concept-based Mining Model, Proceedings of the 6thInternational Conference on Data Mining, ICDM, 2006.
3. Gupta V, Lehal G S. A Survey of Text Mining Techniques and Applications, Journal of Emerging Technologies in Web Intelligence, 2009; 1(1), pp.60-76.
4. Jusoh S, Alfawareh H M. Techniques, Applications and Challenging Issue in Text Mining, International Journal of Computer science Issues, 2012; 9(2), pp.431-36.
5. Mythili K, Yosodha K. Research scholar. Pattern Taxonomy Model with New Pattern Discovery Model for Text Mining, International Journal of Science and Applied Information Technology, 2012;1(3),pp.88-92.
6. Zhong N, Li Y, Wu S T. Effective Pattern Discovery for Text Mining, IEEE Transactions on Knowledge and Data Engineering, 2012;24(1),pp.30-44.
7. Drakshayani B, Prasad E V. Semantic Based Model for Text Document Clustering with Idioms, International Journal of Data Engineering (IJDE), 2013;4(1), pp.1-13.
8. Charjan D S, Pund M A. Pattern Discovery for Text Mining using Pattern Taxonomy, International Journal of Data Engineering, 2013; 4(10),pp.1-6.
9. Radhakrishnan A. Efficient Updating of Discovered Patterns for Text Mining: A Survey, International Journal of Computer Science and Network Security, 2013; 13(10), pp.4550-55.
10. Salton G, McGill M J. Introduction to Modern Information Retrieval, ACM Digital Library: USA ,1986.
11. Bhushan V I, Ujwalapatil. A Comparative Study on Different Types of Effective in Text Mining: A Survey, International Journal of Computer Engineering & Technology, 2013; 4(2), pp.535-42.
12. Aswini V, Lavanya S K. Pattern Discovery for Text Mining, International Conference on Computation of Power, Energy, Information and Communication, 2014.