ISSN XXXX XXXX © 2017 IJESC

Research Article Volume 7 Issue No.7

Short Text Clustering using Web Content Mining: A Survey

Huma Khan[1], Mayur Rathi[2]
M.Tech Scholar[1], Assistant Professor[2]
Department of CSE
LNCTS (RIT) Indore, India

Abstract:
Social network chatting, microblogs and different types of mobile based texts communication are mainly based on short text. Short text is used according to user convenience. Data scientists are working on accuracy of analytical process which deals with text conversations. This work also describes one approach for enhancing text clustering accuracy. Web based knowledge is used for better understanding of textual contents.

Keywords: Text Mining, Text Clustering and Web Content Mining.

I. INTRODUCTION

Data mining is the technique of finding fruitful patterns from the large data and web mining is an important branch of data mining for dealing with huge amount of data from the web. All data also have text contents, so for dealing with those contents data mining has another branch known as text mining. A number of researchers are trying to achieve better performance of text clustering techniques. Web mining is another field of data mining which deals with web data, if anyone wants to apply dynamic cluster analysis at the time of word inserted then he should use web content mining technique. This is useful for any kind of data and capable for analyzing huge amount of web data. Today electronic messages, chatting, micro blogging are increasing rapidly. If any researcher wants to analyze chat data or wants to title any document or any other operation on these type of textual contents. Now every user is doing short message chatting, nobody wants to write full English words. Everyone has his own shortcuts also, so if any researcher wants to perform any analytical approach of mining then it is very crucial task. This work proposes new techniques for enhancing the performance of text clustering by using web contents. Present study shows the text analysis of social networking chat. The proposed clustering technique enhances the performance of text clustering. Clustering of short texts is a difficult task, this work intends to deal with it. Text clustering is one of the major area of research for data mining researchers. Clustering is the process of grouping similar words in without having previous knowledge. This work intends to achieve the solution for short text clustering task by using web contents mining techniques. In the field of data mining, when we deal with huge online contents and extract useful information from the web, this is known as web contents mining.

II. PROBLEM DEFINITION

Now every user is accessing information from the high speed networks continuously. This network offers services and in formations all-time therefore that becomes a part of life of new generation. An imaginary social world is created around every people because of networking websites like twitter, fb and others. Data scientist continuously wants to analyze user's data for getting knowledge about their interests. For throwing advertisements on their personal computers according to their interests. But this process of extracting knowledge from their huge web data is very difficult which needs web mining techniques. According to chat or microblog analysis we can extract knowledge about user's behaviour, but this is very difficult to analyze chat communication because it contains short words those are not proper English words. For identifying the interesting patterns from the social networking web application, information about short text is required. The existing algorithms are not able to provide effective solution, therefore, some additional data mining technique that provides ease in clustering text data is required.

- Clustering of short and unknown words are very difficult.
- Web based knowledge is required to learn new words or short cuts.
- Selection of correct word from web page contents is typical to know.
- Text clustering needs proper cluster analysis technique to check accuracy of cluster.

III. LITERATURE SURVEY

Text mining, sentence mining and document mining are the different fields of analysis on text modules. All text modules have different aspects of analysis and extract knowledge with highest accuracy. Many researchers are doing researches in these all directions. Text content mining and text clustering are the favorite area for many data analyst. Microblogs are text collections with short words; text clustering technique for microblog analysis is also an important field [1]. Many researchers are working on document-summarization, abstracting and title identification. Different machine learning techniques are used for doing human-machine interaction and automatic answering machines, these all systems are based on text analysis, text clustering techniques [2]. Text mining different ranges that having a number of applications on data processing, data retrieval and machine learning. At present most of the analysis are progressing with multiple language support, means able to gain information across languages and capable to group similar data from different kind of language sources according to their original semantics [3]. them properly for further use. In this proposed algorithm, short texts are analyzed using web content mining technique and filter

these gathered data. This algorithm collects all information about that word at real time and store them for future use. Clustering has knowledge about that word like what is the actual form of that word and which is best suitable cluster for it.

The proposed system may work on following steps:
Text document: A document which includes a lot of short text is used as an input, this is the first step of this algorithm. This may be a chat document or microblog which is used for clustering process.
Stop word removal: Stop words are those words which give some pause in language for example full stop, comma etc. This section removes the punctuations and eliminate the repeated words which are not useful. Stop phrases is generally used in every sentences and do not play any essential role in cluster knowledge. Area of microblog feature selection is another field of research, Lio and chen has done work on it [4]. Microblog topic detection using topic modeling method is proposed by huang, yang in 2014 [5]. Text clustering is the hot area of research, in this way gao and zhang propose iteration method for text clustering on microblog posts [6]. A different thing in short text research is duplicate blog and spam detection in small chat or blogs [7]. Rong and song has done work on blogs of government affairs on social media, this paper proposes mining technique for social network analysis [8].

IV. METHODOLOGY

Present clustering algorithms need some addition to resolve this short form of text documents to analyze COMPARISON

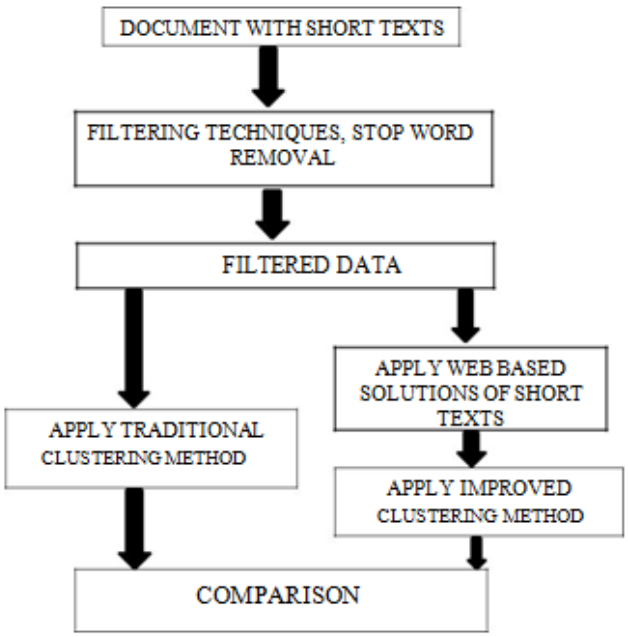

Figure.1. Proposed model for improved clustering
Present clustering: Present clustering are the techniques which are already used in existence but those need some improvements for enhancing their performance. Mostly used text clustering techniques are hierarchical clustering, K-means etc. Proposed clustering: For dealing with short texts of document, web content mining feature is attached to clustering technique. This proposed technique enhances the accuracy of text clustering. Comparison: Last step of proposed method is to compute the performance of method by using some distinct parameters and compare them with different traditional methods.

V. CONCLUSION

The proposed work deals with the text clustering with improved accuracy. For this task web contents are used. This work observes that general text clustering analyzes only text structure or sentence structure, that does not deals with proper definitions or meaning. But using web based search for each text we can identify perfect cluster for that word. This work shows a method for extracting knowledge from web contents, because web contents have all types of data like advertisements. It fetched only related information to words. Finally after observations on different expects this concludes that proposed technique is better than other present methods.

VI. REFERENCES

[1].Jiliang TANG, Xufei WANG, Huiji GAO, Xia.HU, Huan LIU, "Enriching short text representation in microblog for clustering" IJCSI -2012.

[2].H. P. Luhn, "A Business Intelligence System",. Non-topical Issue, IBM Research Journals, Volume 2, Number 4, pp. 314 2013.

[3]. B. V. Rama Krishna, B. Sushma, "Novel Approach to Museums Development & Emergence of Text Mining", International Journal of Computer Technology and Electronics Engineering (IJCTEE), ISSN 2249-6343, Volume 2, Issue 2, 2012.

[4].Zintao Lio and Wei Chen "Short Text Feature Selection for Micro-blog Mining" IEEE, 2010.

[5].Siqi Huang, Yitao Yang, HuakangLi, GuoziSun,."Topic Detection from Microblog Based on Text Clustering and Topic Model Analysis" IEEE 2014.

[6].Kai Gao, Bao-quan Zhang "Modelling on Clustering Algorithm Based on Iteration Feature Selection for Micro-blog Posts" ICMIC (2014).

[7].Qunyan Zhang, Haixin Ma, Weining Qian, Aoying Zhou "Duplicate Detection for Identifying. Social Spam in Microblogs" IEEE 2013.