[Volume 3 issue 6 June 2015]
Page No.1025-1031 ISSN :2320-7167

INTERNATIONAL JOURNAL OF MATHEMATICS AND COMPUTER RESEARCH

Short Text Clustering; Challenges & Solutions: A Literature Review

Dr. Tamanna Siddiqui,[1]Parvej Aalam[2]

'1Associate Professor, Department of Computer Science, Aligarh Muslim University, Aligarh, U.P. India,
E-mail-drtamannasiddiqui.cs@gmail.com
2Research Scholar, Department of Computer Science, Aligarh Muslim University, Aligarh, U.P. India,
E-mail-parvezalig2006@gmail.com

ABSTRACT

Clustering has been an interesting research area due to its usability. It is an unsupervised learning technique which is used in many fields such as machine learning, data mining, pattern recognition, image analysis and bioinformatics etc. Also, Clustering helps in finding targeted patterns and structures directly from very large data sets with little or none of the background knowledge. Researchers are working on it by using various clustering methods/algorithms like K-means, C-means, Fuzzy C-means etc. to make it useful and have obtained meaningful results. In Short Text Clustering, clustering is performed using short text data like tweets, facebook messages, various news feeds etc. Short text as its name suggests is a text that contains only a few words; for instance, the length of a short text in Twitter is less than 140 words; Search engine queries are mostly short texts. However, it may prove very helpful in extracting meaningful information if this large unorganized data may be grouped on the basis of some similarity. But, the major problem in clustering short text is its *sparse feature vector*. 'Feature Vector' is the key element in clustering technique. So, the major solution proposed by researchers is to expand short text data as a long text using various concepts. In this paper, we have discussed various challenges in short text clustering, concepts used to overcome these challenges and also we have discussed the other possible solutions which may further improve clustering results.

KEYWORDS: Short Text, Feature Vector, Clustering, Sparse

INTRODUCTION

Clustering is a popular technique used in various fields for probing and analyzing data. It is classified as a subfield of data mining and as data mining itself, clustering analysis has its own multidisciplinary nature. It has been widely studied by experts in a number of research communities. More formally, the clustering problem describes the problem where the goal is to partition a given set of n entities (also known as patterns or points etc.)into several groups based on how similar/dissimilar they are, such that entities within the same group are similar

to each other and entities that belong to two different groups are dissimilar to some extent. Clustering techniques are used to solve a wide variety of research problems. It also plays an important role in solving many engineering and practical problems. It is used in medicine field for finding cures and symptoms of diseases. In the field of archaeology researchers find the taxonomies of stone tools, funeral objects by applying cluster analytic techniques. In the field of Image segmentation, different clustering algorithms are used to obtain segment labels for each pixel of an image. Clustering is also used in information retrieval, market analysis.

Short Text Clustering is the clustering done on short text data. With tremendous growth of web2.0, more and more short texts are generated in all kinds of websites. Facebook, Twitter and Microblog are few to be named amongst them. These are very short text containing a few words only. It may help to extract meaningful results if arranged on the basis of some similarity. Readers have to spend a lot of time to find the useful information which they expect in this large scale data set. Moreover, different web sites often publish the same information related to a topic which will lead to redundancy of browsing results, thus increasing the size of short text data set. Therefore, the need to cluster these short texts is felt so that they may help the people in finding more accurate and unique browsing results related to a particular topic, no matter whether the results are fetched from same website or different websites. Since several years, researchers are working on it and have obtained meaningful results.

RELATED WORK

To solve the issue of sparsity of feature vector, the short text in short text clustering is expanded by the use of external knowledge sources [2][3][4][6][7][12]. Hotho et al. [4] believe that expanding the WordNet synsets to the documents is able to achieve better results than using "bag of word" in documents. Recently, Wikipedia is explored as a knowledge database for all types of documents clustering [2][6][7]. The results of web search engine can be also used as the knowledge expansion of short texts [10][11]. Sahami and Heilman [10] addressed the problem of short text clustering by introducing a new method for measuring the similarity between short texts. Banerjee et al. [2] created a Lucene index of the Wikipedia articles and then used the query strings to retrieve the top matching Wikipedia articles from the Lucene index as the expansion of the short texts. In 2006, a Scholar named Gabrilovich[1] has demonstrated the value of using Wikipedia as an additional source of features for text categorization. A popular approach within the literature has been the application of lexical resources such as WordNet to aid in the comparison of textual data[4]. WordNet provides a manually annotated lexical database of the English language. Taking advantage of the semantic relationships expressed between terms in WordNet, several methods have been proposed for compensating issues of semantic ambiguity when comparing text (Hotho et al. (2003), Jing et al. (2006), Li et al. (2008)). One drawback to these methods is that the

creation and maintenance of such resources can be very expensive, and obtaining a suitable resource may be difficult for some domains. Ni et al. presented Term Cut method [3] that clusters short texts by finding core terms in the corpus. Janruang and Guha [9] proposed the semantic suffix tree clustering method that constructs the semantic suffix tree through an on-depth and on-breadth pass by using semantic similarity and string matching. Quan et al. [8] mined the relation of the non-common terms between the short texts based on a series of third-party topics. After that, the feature vectors of two short texts are modified according to the discovered relation, and then the cosine metric is employed to calculate the similarity of modified vectors. Xiaohui et.al. [12] given the idea of Short Text Clustering with Expanding Keywords through Concept Graph. They showed that concept graph method used to expand short texts is superior to the methods, which do not expand keywords with Wikipedia in terms of precision, recall and F-score.

We reviewed large number of research papers to find the challenges in short text clustering. As a result, following findings are obtained.

CHALLENGES IN CLUSTERING SHORT TEXTS

3.1 Sparse Feature Vector

In document or large text clustering methods, each document is represented by a feature vector. This vector is a 2-d vector of numerical values corresponding to terms in the document and other metadata so that a number of machine learning algorithms can be run on them. The numerical values in a feature vector are the weights of terms in the documents to be clustered. Terms in a document/documents can be weighed using different schemes. The most common is tf-idf i.e. term frequency-inverse document frequency. In feature vector, each term is assigned a weight equal to its tf-idf value. The tf-idf is a product of term frequency represented as f(t,d) which in simplest way is calculated by counting the number of times a term ‘t’ occur in a particular document ‘d’ and inverse document factor represented as idf(t,D) which is calculated by dividing the total number of documents ‘D’ by no. of documents containing that term.

Mathematically,

$$tfidf(t,d,D) = tf(t,d) * idf(t,D)$$

In short texts, the number of words being very less, the feature vector generated from short text are generally sparse in nature. This sparsity of feature vector is a major problem in clustering short text data and resolving this problem is a challenging task. The other problems related to clustering short texts are:

3.2 Polysemy

The existence of more than one meaning for a single word. For example, the word table may refer to a type of furniture or a row-column table or to a 2,3,4,5……table. So, in which category, the word table should be placed? In a document, it may not be so challenging

because we can identify easily from the context that to which 'Table', the word 'table' is referred. But in short text it is merely a challenging task; the reason being simple that there are only few words and so context of a particular word cannot be understood.

3.3 Synonymy

Two or more words having the same meaning. For example; the words Beautiful, Attractive, Pretty, Lovely, Stunning have same meaning. So it is also a challenging task to decide in which cluster such words would be placed especially in case when such words are found in short texts.

CONCEPTS USED

In this section, we discuss some of the prominent methods used by researchers in expanding the short texts for the purpose of clustering. These methods have improved the Short Text Clusteing accuracy to a greater extent and also have some shortcomings which are discussed in further section of the paper. These methods include:

Wiki_Method [2]

They downloaded English Wikipedia dump of 1 day containing 1,174,107 articles.A Lucene(http://lucene.apache.org/) index of these articles was then created.With the available short text of news feed, a query string was used to retrieve the top matching articles from the Lucene index.The titles of the retrieved Wikipedia articles served as additional features of the feed item for clustering.Then different clustering algorithms were run using this representation and traditional simple bag of words representation.Experiment was done with the snapshot of Google news homepage on a day and gathered 1557 articles on 26 different topics.Each article here consisted of a title and one line of description.Then using lucene index feature vectors related to these topics were created. Freely available clustering package SenseClusters was used.

SenseClusters uses CLUTO as clustering engine and it provides 6 different clustering algorithms; rb, rbr, direct, agglo, graph and bagglo.All these algorithms were run by representing the news articles as simple bag representation and by using wikipedialucene index. The former way here is known as Baseline and the latter as Wiki_Method.

Table 1: below shows the accuracy (in percentage) achieved by the baseline and Wiki_Method with the different clustering algorithms of Cluto:

(Table 1)

Cluto Algorithm	rb	Rbr	direct	agglo	graph	bagglo
Clustering Accuracy (in %) by baseline method	63.20	79.38	67.05	22.03	81.62	23.57
Clustering Accuracy (in %) by wiki_method	85.42	63.65	82.66	83.88	89.56	43.67

The above table of results shows that except the case of 'rbr' clustering algorithm, the Wiki_Methodachived better accuracy than the Baseline. Thus Wikipedia proved to be an strong source in improving the accuracy of short text clustering.

4.2 Concept Graph Method [14]

It is a graph in which a concept acts as a vertex and the relationship between two concepts which have common terms as the edge.A concept is nothing but a group of keywords having high possibility of co-occurrence. The combination of several keywords in a concept provides a rich semantic meaning.For e.g: The concept "bird flu" has more concrete meaning than the keyword "bird" or "flu".By this method they aimed to improve accuracy in the results of short text clustering by expanding the short texts with the internal knowledge of the data set.

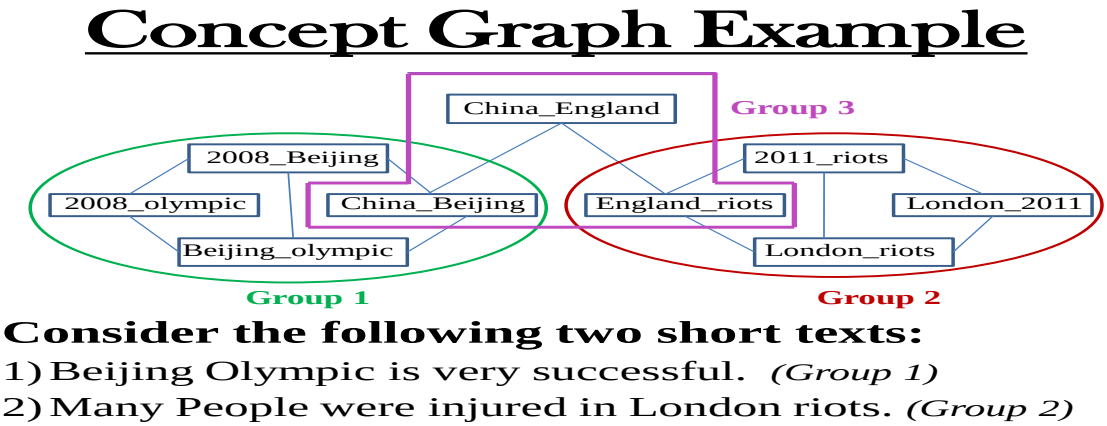
By this method clustering was done in three steps as follows:

I. Extracting concepts from the data set.

II. Constructing the concept graph and obtaining the concept groups by partitioning the concept graph.

III. Expanding the short texts with the concept groups and then clustering the expanding short texts replacing the original short texts. Figure 1 below demonstrates the use of 'Concept Graph' for clustering Short text Data.

(Figure 1)



MAJOR FINDINGS

As the major problem in clustering short text is its sparse feature vector, the only solution as also proposed by various researchers is to enrich its feature vector. Obviously, the feature vector would be enriched when the short text would be expanded by taking external help such as Wikipedia, WordNet, Concept Graph as done till date. By expanding short text using external sources, the problem of Polysemy and Synonymy is also resolved to a greater extent and thus better and better results are achieved in the field of Short text Clustering. After

reviewing the work done on short text clustering, we have reached to the following conclusions:

Testing Wiki_Method with several algorithms

The **Wiki_Method**[2] was tested only with Cluto algorithms. Cluto is a clustering engine containing several algorithms for clustering. Wiki_Method was only tested with six algorithms may be tested with other clustering algorithms as K-means, DBSCAN, Fuzzy C-means algorithm, EM algorithm etc. and then accuracy be compared.

5.1 Using additional Wikipedia Concepts

Additional Wikipedia Concepts should be used as an additional feature for enriching the short text. Some of the additional Wikipedia concepts are:

.**1Category Network** :

Wikipedia has different categories of different topics which are then divided into sub-categories and the sub-categories are further divided and divided to form more sub-categories. This forms the category Network of Wikipedia. These different categories can help to expand our short texts and hence to enrich 'feature vector' of short text data.

5.1.2 Page link structure:

This is another Wikipedia concept that can help in short text clustering. In a Wikipedia page, there are many words written in blue color. Each of these has a link to a different page of Wikipedia. The page opened by a particular link will further contain words having link to more Wikipedia pages. This forms the page link structure of Wikipedia.

5.1.3 Anchor text:

The words/text written in blue color in a page forms the anchor text of that Wikipedia page. The feature vector of short text data may also be enriched using anchor text.

5.1.4 Full text:

Wikipedia method used by [2] used only the titles of articles for clustering short text data. Full text of an article may also prove helpful in this.

CONCLUSION AND FUTURE WORK

Short Text Clustering is the major obsession in both academia and industry. Processing short texts is becoming a trend in information retrieval. Since the short text has sparse feature vector that can be extended by using external information, it is more challenging than document clustering. So, the main objective of our research will be to propose a framework that can resolve the issues related to Short Text Clustering. More accuracy in clustering results may be achieved by running existing document clustering algorithms for clustering short texts. Wikipedia is a large electronic knowledge storehouse on the web with millions of articles contributed collaboratively by volunteers. Further, additional Wikipedia concepts such as Word-Net, Anchor text, Page Link Structure may also serve our purpose.

REFERENCES

1. E. Gabrilovich. Feature Generation for Textual Information Retrieval Using World Knowledge. PhD Thesis, Department of Computer Science, Technion – Israel Institute of Technology, Haifa, Israel, 2006.
2. Banerjee, Somnath, Krishnan Ramanathan, and Ajay Gupta. "Clustering short texts using wikipedia." Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval. ACM, 2007.
3. X. Ni, X. Quan, Z. Lu, and W. Liu and B. Hua, Short text clustering by finding core terms, Knowledge and information systems 2011;27(3): 345-365.
4. Hotho, S. Staab and G. Stumme, WordNet improves text document clustering, In Proceedings of the 26th Annual International ACM SIGIR Conference Semantic Web Workshop 2003; 541-544.
5. S. Tian, and X. Zhai, L. Yu, and H. Guo, Uyghur Text Clustering Based on Semantic Word Set, Journal of Computational Information Systems 2013; 9(2): 781-790.
6. G. Spanakis, G. Siolas, and A. Stafylopatis, Exploiting Wikipedia knowledge for conceptual hierarchical clustering of documents, The Computer Journal 2012;55(2): 299-312.
7. X. Hu, X. Zhang, C. Lu, E.K. Park, and X. Zhou, Exploiting Wikipedia as external knowledge for document clustering, In Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining 2009; 389-396.
8. X. Quan, G. Liu, Z. Lu, X. Ni and W. Liu, Short text similarity based on probabilistic topics, Knowledge and information systems 2010; 25(3): 473-491,
9. J. Janruang, and S. Guha, Semantic Suffix Tree Clustering, In Proceedings of the International Conference on Data Engineering and Internet Technology 2011; 35-40.
10. M. Sahami and T.D. Heilman, A web-based kernel function for measuring the similarity of short text snippets, In Proceedings of the 15th international conference 2006 on World WideWeb;377-386.
11. D. Bollegala, Y. Matsuo, and M. Ishizuka, A Web Search Engine-based Approach to Measure Semantic Similarity between Words, IEEE Transactions on Knowledge and Data Engineering 2011; 23(7): 977-990.
12. HUANG, Xiaohui, et al. "Short Text Clustering with Expanding Keywords through Concept Graph." Journal of Computational Information Systems 9.21 (2013): 8649-8657.