International Journal of Advance Research in Science and Engineering
Volume No.07, Special Issue No.04, April 2018
www.ijarse.com
IJARSE
ISSN: 2319-8354

A Brief Survey on Labeling Methods used in Text Document Clustering

Harsha Patil[1], Ramjeevan Singh Thakur[2]

[1]*Research Scholar, Department of Computer Applications,*

[2]*Associate Professor, Department of Computer Applications,*

[1,2]*Maulana Azad National Institute of Technology(MANIT),Bhopal, Madhya Pradesh, India.*

ABSTRACT

As different document clustering methodologies proved that clustering is the efficient method for solve the query of search engine on internet and provide fast and precise output. Document clustering results in different clusters of documents which are internally very cohesive in reference of similarity with the documents belongs to the same cluster. Labeling to clusters are another very important era for understand cluster properly by name and to get that which types of documents it consist, In this paper we provide brief survey on clustering methods which are used by researcher. This source will definitely help researcher for deciding method for cluster labeling.

Keywords: Chi-Squared Selection, Document Clustering, Labeling, Text Mining, Topic.

I INTRODUCTION

Cluster labeling is the method of providing precise, understandable and descriptive labels for the clusters. Traditional clustering algorithms do not have standard method to generate any label for produced clusters. Cluster labeling techniques introspect theinsides of the documents per cluster to find a labeling that encapsulate the topic of each cluster and discriminate the clusters from each other.

Many efficient algorithms are proposed and implemented by many researcher for grouping the text documents into different clusters without annotations alongside the resulted clusters. Few researcher are contributed in this field and have been work for many techniques of cluster labeling like Wikipedia-based cluster labeling [1], hierarchical cluster labeling [2], [3] etc.

II CLUSTER LABELING METHODS

Clustering labeling methods broadly can be divided in to two categories: 1. Differential cluster labeling 2. Cluster-Internal Labeling

2.1 Differential Cluster Labeling

Differential cluster labeling labels a cluster by comparing term distributions across clusters. Frequent terms are used to represent the whole clusterTerms having can be omitted in labeling a cluster. By omitting less frequent terms and using a differential test good labeling can be generated.

2.1.1 Pointwise mutual information

Mutual information measures the degree of dependence of two random variables.

In the case of cluster labeling, the variable X is associated with membership in a cluster, and the variable Y is associated with the presence of a term. Both variables can have values of 0 or 1. Function p(C,T) is used to represent the probability that two events occur simultaneously. where C is the particular cluster ant T represent the term.

2.1.2 Chi-Squared Selection

The Pearson's chi-squared test can be used to calculate the probability that the occurrence of an event matches the initial expectations. It will identify cluster labels that characterize one cluster in contrast to other clusters.

2.2 Cluster-Internal Labeling

Cluster-internal labeling selects labels that only depend on the contents of the cluster of interest. No comparison is made with the other clusters. Cluster-internal labeling can use a variety of methods, such as finding terms that occur frequently in the centroid or finding the document that lies closest to the centroid.

2.2.1 Centroid Labels

This method is used when document and their terms frequencies are represented by Vector Space Model (VSM). In VSM importance of terms are calculated by its weights of the terms .For weight calculation term frequency are used.

We can calculate the centroid by finding the mean of all the document vectors. If an entry in the centroid vector has a high value, then the corresponding term occurs frequently within the cluster. These terms can be used as a label for the cluster.

2.2.2 External knowledge labels

External Knowledge database like Wikipedia, WordNet can be used for Cluster labeling. Richness of this Knowledgebase supports to find more semantically and precise cluster labels

III RELATED WORKS

Clustering is the methodology of categorized documents in different clusters, on the basis of their similarity with each other. Similar documents are placed in one cluster. Two different clusters' documents are dissimilar to the documents of another cluster and similar to the documents of same cluster. These clustering process does not generate cluster labels. So here is the challenges for researcher to generate cluster labels suitable for the clusters and provide understandable meaning which is self-explanatory for the cluster details.Few researcher are contributed in this field and have been work for many techniques of cluster labeling like Wikipedia-based cluster labeling [1], hierarchical cluster labeling [2], [3] etc.

After the clustering many researchers works for cluster labeling. Many researcher used Statistical Based method [9] , NLP based and on based on Semantically similarity by using External Knowledge base[10]. In [4] researcher used frequent terms for finding clusterlabels.In some research work [5] NLP based methods are adopted to find suitable cluster labels.

Anotherwork that talks about the preservation of semantics whileforming topics is by Zhou Chong et al. [6] used the semantic similarity for generate more meaningful clusters , where they consider a window within which they find theitemsets which are candidates for topics. Ling Zhuanget al. [7]used maximal frequent itemsets for document clustering and used them as cluster labels. XueyuGeng and JinlongWang [8] presents analysis of academic documents and provide high quality topic summarization.They used LDA model.But it was manual based on manual topics modeling

IV CONCLUSION

In this paper we present brief study regarding methodologies used for Cluster Labeling. Many Methods which are belongs to Differential Cluster labeling or internal cluster labeling are discussed. This source will definitely help researcher for deciding method for cluster labeling.

REFERENCES

[1] Carmel D, Roitman H and Zwerdling N 2009 Enhancing Cluster Labeling Using Wikipedia in *Proc. of the 32nd Intl. ACM SIGIR Conf. on Research and Development in Information Retrieval* pp. 139–146.

[2] Treeratpituk P and Callan J. 2006 Automatically Labeling Hierarchical Clusters *Proc. of the 2006 Intl. Conf. on Digital Government Research* pp 167–176

[3] Moura M F and Rezende S O 2007 Choosing a Hierarchical Cluster Labelling Method for a Specific Domain Document Collection *New Trends in Artificial Intelligence* pp. 812–823

[4] C. Wartena, and R. Brussee, "Topic Detection by Clustering Keywords," Proceedings of the 19th International Conference on Database and Expert Systems Applications, 2008.

[5] E. Lloret, "Topic Detection and Segmentation in Automatic Text Summarization," http://www.dlsi.ua.es/~elloret/publications/SumTopics.pdf, 2009.

[6] Z. Chong, L. Yansheng, Z. Lei, and H. Rong, "FICW:Frequent Itemset Based Text Clustering with WindowConstraint," *Wuhan Journal of Natural Sciences, Vol:11*, No: 5, pp: 1345-1351, 2006.

[7] L. Zhuang, and H. Dai, "A Maximal Frequent Itemset Approach for Web Document Clustering," Proceedings of the 4th *International Conference on Computer and Information Technology, 2004.*

[8] X. Geng, and J. Wang, "Toward theme developmentanalysis with topic clustering," Proceedings of the Ist *International Conference on Advanced ComputerTheory and Engineering*, 2008.

[9] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. I.Griffiths, "Probabilistic Author-topic Models forInformation Discovery," *Proceedings of SIGKDD,2004.*

[10] Xu T and Oard D W 2011 Wikipedia-based Topic Clustering for Microblogs *Proc. of the American Society for Information Science and Technology* pp. 1–10.