ISSN No. 0976-5697

Volume 8, No. 1, Jan-Feb 2017

International Journal of Advanced Research in Computer Science

REVIEW ARTICLE

Available Online at www.ijarcs.info

Study on Text Clustering For Topic Identification

Sindhu Antony
Department of Computer Science
Christ University
Bangalore, India

Rupali Wagh
Department of Computer Science
Christ University
Bangalore, India

Abstract: Due to advent of web technologies, amount data available has grown enormously. Information retrieval from this data thus has become most important operation. Huge portion of data are available as text and poses new challenges in information retrieval and search operations. Large texts can be grouped into clusters which process the text easily. Grouping of text based on similarity of contents is called as text clustering. Topic identification refers to recognizing topic/ideas conveyed in the text. It is necessary to extract the key idea of a text if no categorization of text exists. Topics can be identified by clustering text and then extracting keywords from clusters. There are many algorithms for text clustering. Hierarchical clustering and K-means clustering are two important clustering techniques. This survey paper discusses about various clustering algorithms and application of clustering for topic identification. Also focuses on challenges and issues of text clustering and topic identification.

Keywords: Topic identification, Text clustering, K-means clustering, Hierarchical clustering, Differential cluster labeling, Cluster initial labeling

I. INTRODUCTION

Text mining is one of the most important and efficient technique to extract data from a large text document. It is additionally called knowledge discovery in text (KDT) or intelligent text analysis. Hidden knowledge can be extracted from both structured and semi structured data. The fundamental idea of a text or document can be recognized by its topic. Topic gives the core theme of the document. The user or the reader can understand the text from its topic easily without reading the entire document. The topic of a passage is the center of the passage and tells the users what the passage will be about. Text clustering is one of the main techniques which can be used for automatic identification of document topic from large document corpus. Clustering is most widely used mining technique in almost every domain. Because of its "unsupervised", nature, clustering can aptly be used as a first step to get more insights into the data and hence is used as a preprocessing step in many data mining solutions. Text clustering or document clustering can be defined as grouping of text documents based on the similarities in the contents of text present in the documents. In any information processing system, where the size of text corpus is generally very huge, clustering of related or similar objects has long been considered as a very useful tool. It helps users while navigating through a large document collection by organizing document collection. Document collection can contain documents belonging to different groups. Identification of groups and topic that they represent using cluster labeling is one of very important application of document clustering. There are two major approaches to automatic cluster labeling-

- Differential cluster labeling
- Cluster initial labeling

Text clustering, both hierarchical and flat is thus used for document grouping and improving the results of an information retrieval systems

II. LITERATURE SURVEY

The goal of topic identification is to find out labels or categories from a given large set of documents. Text or documents contains information stored without any structure. This unstructured nature of data poses challenges in extraction of relevant pattern. The unstructured text is then transformed into intermediate form. Such intermediate form is then used for application of mining tools and techniques. Vector space model (VSM) is an algebraic representation of text documents. In VSM every text document is represented as vectors and terms of document are used for indexing. It can be used for information filtering, information retrieval and text analysis. The paper titled [13] "Vector space model: An information retrieval system" explains more about information retrieval techniques and its variations. This paper proposes new variations in information retrieval techniques. Retrieval of information from web is an essential process in today's world. Web is a collection of large amount of data. The paper titled[14] "Analysis of vector space model in information retrieval" presents different approaches of vector space model to compute the similarity score of hits from search engine. It also explains the issues and problems of vector space model in information retrieval and comprehensive comparison of term count model. The paper titled [15] "From frequency to meaning: Vector space models of semantics" describes the use of vector space models for semantic processing of text. This paper presents variety of applications of vector space model.

Text clustering is a prime application of cluster analysis to textual documents. It is a classic area for machine learning and pattern recognition. It is an extensively studied research domain. Due to the advance in web technology, there is a great increase in the quantity of information available on internet. Also the quantity of data set aside in computer files and databases are rising at an amazing rate. Document clustering has been investigated for use in a number of different areas of text mining and information retrieval. The paper titled [1] "A

Survey Paper On Different Techniques Of Document Clustering discusses various document clustering techniques along with their pros and cons. Clustering is currently one of the most essential technique for dealing with massive amount of information on the web and personal system. The principle of clustering depends on the concept of dividing a set of objects into a specified number of clusters on the basis of characteristics found in the actual data. This paper discusses various algorithms and comparisons. Algorithms are compared with on various parameters such as time complexity, space complexity and models like static and dynamic. The paper titled [2] "Survey on clustering techniques in document clustering" explains about clustering and different types of cluster models. This paper also explains the different cluster models. The various cluster models are connectivity models, distribution models, centroid models, density models, subspace models, group models and graph based models. And this paper is describing about different clustering algorithms too. The paper titled [3] "A survey of text clustering algorithms" is a research survey on text clustering algorithms. The problems of text clustering and key challenges of clustering are important presentations of this paper. Key methods used for text clustering and its relative advantages are also been explained in this paper. The recent advances in the area in the context of social network and linked data are also explained. This paper describes some methods for text clustering such as feature selection method and transformation method. Feature selection method is helpful for the quality text clustering. It is easy to apply in the problem of text categorization. Transformation methods are used to transform the text to a new feature. Latent Semantic Indexing is one of the transformation methods.

K-means clustering algorithm is an efficient algorithm for text clustering. It is explained by different researchers. The paper titled [4] "Survey on clustering algorithms and K-Means" describes the comparison between all types of clustering algorithms and in-depth discussion of K-Means algorithm. K-Means clustering is a partition based algorithm which can be used for small as well as large dataset. It is efficient and relatively scalable algorithm. One of the main disadvantages of this clustering is, it is more sensitive to noise. The paper titled [5] "Survey on various enhanced K-Means algorithms" explains the enhanced K-Means algorithm which is simple, scalable and efficient. But this algorithm has some limitations like random selection of centroid, number of cluster K initialized and influenced by outliers. This paper describes the survey of how to improve this algorithm from all its deficiencies. The paper titled [6] "A survey on clustering principles with K-Means clustering algorithm using different methods in detail" elaborates that clustering is an essential task in data mining process. According to the given dataset the clusters can be made and each cluster will have similar object. Based on the similarity there will have K number of clusters. According to this survey K-Means clustering takes more time for its execution. In order to avoid that ranking method and query redirection is proposed. The paper titled [7**]** "Global K-means (GKM) clustering algorithm: A survey" presents about the variants of K-Means clustering algorithm and a critical analysis of it. This paper also proposed a new concept of Faster Global K-Means algorithm of Streamed Datasets (FGKM-SD) which will improve the efficiency, less running time and less storage space. The paper titled [8] "A survey on K-Means clustering and web-Text mining" explains K-Means clustering and web-Text mining. Text mining refers to extracting needed information from text and web mining refers to getting unknown information from the web. This paper highlights the problems of searching of research papers on web and concerns about time efficiency. To overcome this limitation, weighted page rank method is coupled with K-Means clustering algorithm to improve the execution time. This research is mainly related to assign the ranks of research papers on the basis of popularity of papers.

The second most common clustering algorithm is hierarchical clustering algorithm. As the name indicates the structure is in a hierarchical way. By aggregating or dividing, documents can be clustered into hierarchical structure, which is suitable for browsing. It is a connectivity based algorithm. When compared to K-Means clustering it has less efficiency. The paper titled [9] "A survey on Efficient hierarchical Algorithm used in Clustering" explains basics of Hierarchical clustering to build the hierarchy of clusters. It is used with unstructured set of data. This paper concentrates more on inter and intra clusters differences and also explains the advantages and disadvantages of hierarchical clustering. Agglomerative and divisive are two types hierarchical clustering techniques. The paper titled [10] "Survey on Hierarchical document clustering techniques (Fihc and F^2ihc)" describes about the Incremental hierarchical clustering, Frequent item set Based hierarchical clustering (Fihc) and Fuzzy Frequent Item set based hierarchical clustering algorithm(F^2ihc). In this paper, frequent items of a cluster are more accurate and efficient. The paper titled [11] "Survey paper on Clustering techniques" presents different hierarchical clustering algorithms especially. Hierarchical clustering builds a hierarchy of clusters. There are two types of hierarchical clustering. 1. Agglomerative: Proceed by series of fusions of the 'objects' into groups. 2. Divisive: Separate 'n' objects successively into finger groupings. It also explains the advantages and disadvantages of hierarchical clustering. The paper titled [12] **"**A survey of hierarchical clustering algorithm" discusses about clustering algorithms for cost based optimization. It also explains the clustering algorithms and its attributes and the differences between them. Main clustering algorithms explained in this paper are sequential algorithms, Hierarchical clustering algorithm, Agglomerative algorithm and divisive algorithm. Clustering algorithms for cost based optimization is calculated by using cost function. Various algorithms are compared with its attributes such as space complexity, time complexity and models like static and dynamic.

III. TEXT CLUSTERING AND TOPIC IDENTIFICATION

Clustering is the most common form of unsupervised learning. It requires no human expert to group the data which means that the grouping is performed on the basis of inherent similarities and differences among the documents. Since the documents belonging to one cluster are "similar" to each other. Text clustering can be applied aptly in information retrieval. Following are few applications of text clustering in information retrieval –

1. Search result clustering – Instead of displaying a plain list of documents that matched against the query, retrieved list is further clustered to make the understanding of the result better.

2. Scatter Gather – This approach clusters the document set after which the user can manually select documents of his choice which can further be clustered. The process can be repeated till relevant groups are formed.
3. Collection clustering – The collection of documents are clustered to improve the recall of the result of a query.
4. Language Modeling – Based on the appropriate text representation model entire document is clustered to improve the precision/recall of the result
5. Cluster based retrieval – The document collection is cluster for improving the efficiency of information retrieval systems.

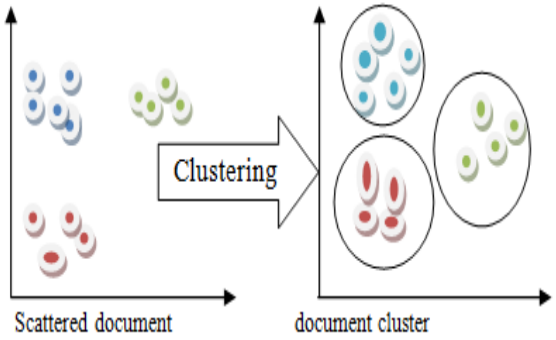

Fig.1: Clustering

Topic identification is one of the most common application of text clustering which is used extensively in information retrieval. The paper titled [20] "Topic identification Frame work and Application" explains about the necessity of topic identification in documents. This paper also presents the topic identification problems along with its framework and desired properties and also some classification scheme of identification algorithms.

The paper titled[19] " Identifying document topics using the Wikipedia category network" narrates that a simple algorithm that exploits only the headings and categories of Wikipedia articles can categorize documents by Wikipedia categories. The paper titled [18] "Topic detection by clustering key words" tells that by clustering keywords the topic can be identified. The keywords are extracted and clustered based on similarity measures by using K-Means bisecting algorithms. The paper titled [17] "Topcat: data Mining for Topic identification in a text Corpus" Explains that it is a technique for identifying topics in articles. NLP is used to identify Key entities in individual articles. This paper describes how to identify topics in text and the association rules among entities.

By clustering Text documents similar documents will be in one group. The paper titled [16] "Document and Topic discovery Based on semantic similarity in Scientific Literature" presents a modified semantic model. Here related terms are extracted as concepts and identified by its topics by bisecting K-Means clustering algorithm and topic detection method.

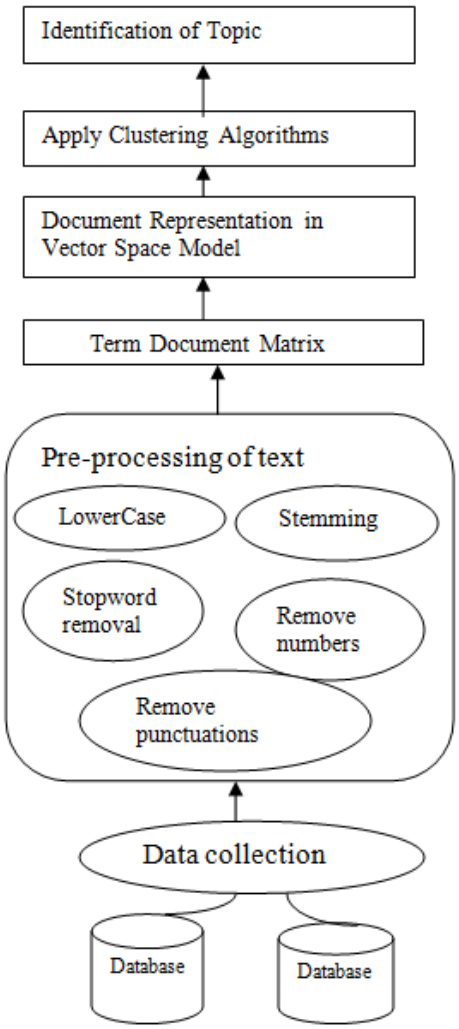

Fig.2: Generic Framework for Topic Identification Using Document Clustering

IV. CHALLENGES AND ISSUES OF CLUSTERING FOR TOPIC IDENTIFICATION

- Large datasets are one of the major issues for text clustering and topic identification. It is difficult to identify the topic from a large dataset. But when the data is clustered then it is easy to find out the topics and key idea of the document.
- The data may be noisy. There may be many outliers in the datasets. These outliers can be removed by accomplishing the preprocessing steps of the text clustering. After preprocessing the dataset will be error free and clustering for topic identification can be done easily.
- Word Ambiguity exists inherently in any text. Sometimes same word may have different meaning, which will make confusion for the users. For example Apple (the company) or Apple (the fruit). Word sense disambiguation strategy can be included to avoid this.
- Context sensitivity is a challenge for text clustering. The text will be sensitive to any context or circumferences. The text will have various concepts.

Each concepts may be high or small relationship between them

V .CONCLUSION

Text clustering and information extraction plays an important role in today's web world. It is very difficult to extract relevant information from the high volume of data. Text clustering technique can be used as a first step for information extraction. Clustering of text helps to increase searching efficiency and reduces searching time. Topic identification from group of documents obtained through clustering further helps in extracting themes in document collection. Generic topic identification model is discussed in this paper. Two most commonly used algorithms for clustering, K-Means algorithm and hierarchical algorithm are emphasized in the paper.

VI. REFERENCES

[1]. MamtaMahilane, Mr. K. L. Sinha, "A Survey Paper On Different Techniques Of Document Clustering", International Journal Of Current Engineering And Scientific Research, Volume-2, Issue-1, 2015.

[2]. MamtaMahilane et.al, "A Survey of clustering Techniques in Document Clustering", International Journal Of Current Engineering And Scientific Research, Volume-2, Issue-1, 2015.

[3]. Charu C Agarwal, ChengXiangZhai, "A survey of Text Clustering Algorithms".

[4] .MeghaMandloi **"**A Survey on Clustering Algorithms and K-Means**"** International Journal of Research in Engineering Technology and Management ISSN 2347 – 7539.

[5].Twinkle Garg, Arun Malik, **"**Survey on Various Enhanced K-Means Algorithms" International Journal of Advanced Research in Computer and Communication EngineeringVol. 3, Issue 11, November 2014

[6]. ManpreetKaur, UsvirKaur **, "**A Survey on Clustering Principles with K-means Clustering Algorithm Using Different Methods in Detail"IJCSMC, Vol. 2, Issue. 5, May 2013.

[7]. ArpitaAgrawal, Hitesh Gupta, "Global K-Means (GKM) Clustering Algorithm: A Survey"International Journal of Computer Applications (0975 – 8887) Volume 79 – No.2, October 2013

[8]. AayushiBindal, AnalpPathak, **"**A Survey on K-means Clustering and Web-Text Mining" International Journal of Science and Research (IJSR) ISSN (Online): 2319-7064 Index Copernicus Value (2013): 6.14 | Impact Factor (2014): 5.611 Volume 5 Issue 4, April 2016.

[9]. G.Thilakavathi et.al, " A Survey on Efficient Hierarchical Algorithm Used in Clustering"International Journal of Engineering Research & Technology (IJERT)Vol. 2 Issue 9, September – 2013.

[10].Ms. DevikaDeshmukh, Mr. SandipKambleMrs. PranaliDandekar"Survey on Hierarchical document clusteringtechniques" International Journal of Advanced Research in Computer Science and Software EngineeringVolume 3, Issue 7, July 2013

[11].AmandeepKaur Mann, NavneetKaur"Survey paper on Clustering techniques"International Journal of Science, Engineering and Technology Research (IJSETR) Volume 2, Issue 4, April 2013

[12]. MarjanKuchaki Rafsanjani et.al, "A survey of ierarchical clustering algorithms"The Journal of mathematics and Computer Science Vol .5 No.3 (2012) 229-240.

[13].Vaibhav Kant Singh, Vinay Kumar Singh, "Vector Space Model: An Information Retrieval System"International Journal of Advanced Engineering Research and Studies E-ISSN2249–8974

[14].JitendraNath Singh , Sanjay Kumar Dwivedi, "Analysis of Vector Space Model in Information Retrieval "National Conference on Communication Technologies & its impact on Next Generation Computing CTNGC 2012 Proceedings published by International Journal of Computer Applications® (IJCA)

[15] Peter D Turney et al, "From Frequency to Meaning: Vector Space Models of Semantics " Journal of Artifcial Intelligence Research 37 (2010)

[16].J. Jayabharathy et.al, "Document Clustering and Topic Discovery based on Semantic Similarity in Scientific Literature"

[17]. Chris Clifton et. al, "TopCat: Data Mining for Topic Identification in a Text Corpus" March 15, 2000.

[19].Peter Sch¨onhofen **"** Identifying document topics using the Wikipedia category network**"**Computer and Automation Research Institute, Hungarian Academy of Sciences,Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence.

[20] Benno Stein et.al , " Topic identification framework and applications" Tochtermann, aurer(Eds.):Proceedings of the I-KNOW '04, Graz 4th International Conference on Knowledge Management Journal of Universal Computer Science, pp. 353-360, ISSN 0948-6968