Survey on Text Clustering Techniques

K.V. Kanimozhi[1] and M. Venkatesan[2]

[1]Girijananda Choudhury Institute of Management and Technology
[2]School of Computing Science and Engineering VIT University, Vellore.
E-mail: [1]kani_kayal@rediffmail.com, [2]mvenkatesan@vit.ac.in

Abstract—*Recently Text mining plays an important role in the areas such as competitive Intelligence, life sciences, social media, sentiment analysis, trend spotting so on. Since 80% of organizations data is unstructured content which includes web pages, legal documents, blog articles, surveys therefore organizations must analyze not just transactional information but also textual content to gain insights and improve performance. So Text clustering is used for organization of collection of text documents into clusters based on similarity. In this paper an elaborate survey is done on various techniques of text clustering and provides the overview of future trends in text clustering.*

1. INTRODUCTION

Since recent developments in technology, science, user habits, organizations, business, etc gave rise to production and storage of massive amounts of data, not surprisingly; the intelligent analysis of big data has become more important for both academics and business. A gigantic amount of data is stored as unstructured text. Examples include newspaper articles, books, email messages, research papers, web pages, and XML Documents.

A knowledge repository typically contains vast amounts of formal knowledge elements which generally are available as documents. With increasing globalization and Internet technology advancements many organizations works on documents.

With the advancement of technologies in World Wide Web, huge amounts of rich and dynamic information's are available. With web search engine a user can quickly look through and locate the documents. Usually search engines returns numerous documents, a lot of which are relevant to the topic and some may contain irrelevant documents with poor quality. Clustering plays an important role in organizing such massive amount of documents returned by search engines into meaningful clusters. A cluster is a collection of data objects where documents within a cluster have high intra similarity and low inter-similarity to other clusters. Document clustering is closely related to data clustering.

2. BACKGROUND WORK

Clustering of text documents is used to group documents into relevant topics. The major complexity in document clustering is its high dimension. It requires efficient algorithms which can solve this high dimensional clustering. A document clustering is a major topic in information retrieval area. Example includes search engines. The basic steps used in processing the document are

2.1 Bags of words model

For many pattern discovery tasks, each document is treated simply as a bag of words, the set of all words appear in that document. Sometimes this word augmented with frequency of words in the document. These ways of modeling documents are typically quite effective for most tasks.

2.2 Vector space model

In this model, the dimension of the vector space consists of all the words there in all the documents in the collection. This is a very high dimensional space. Each document is then represented as a vector in this space contains the frequency of all the words in that document.

For example, if the dimensions are ('and', 'the', 'computer', 'science') then a document may be represented as (4, 3, 2, 1) the numbers indicate the frequency of the corresponding words in that document. The advantage of the vector space model is that it is likely to use mathematical techniques developed in linear algebra to solve problems. The major steps involved are preprocessing the text documents.

2.2.1 Preprocessing the Text Document:

The set of all the words in all documents is a large number to deal with. Most text processing and web mining uses some pre-processing in order to reduce the number of dimensions.

a. Filtering
The process of removing special characters and punctuation that are not thought to hold any discriminative power under the vector model.

b. Tokenization

Tokenization Splits sentences into individual tokens, typically words. More sophisticated methods, drawn from the field of NLP, parse the grammatical arrangement of the text to pick significant terms.

c. Stop words

The process of removal of words that do not have much value to the process of pattern discovery. Words such as 'and' and 'the' are frequent and their presence or absence should not really influence the result of mining. Such words are called stop words. List of stop words are widely obtainable and are used to remove their presence in documents before processing them further.

d. Stemming

The task is to find the stems of words. Documents may have a word in many different forms. For example, 'learn' may be present as 'learn', learning', 'learnable', 'learned' etc. In all forms the concept of learning is present is significant than the form. To handle this situation, stemming is done where all the different forms are transformed to its root from by removing suffixes and or prefixes. Standard stemmers which handle the task well are widely available.

e. Pruning

Pruning removes words that come into view with very low frequency throughout the corpus. The underlying assumption is that these words, even if they had any discriminating power, would form too small clusters to be useful

2.2.2 Termfrequency-inversedocument frequency (TF-IDF)

TF-IDF is another widely used technique to encode the text document to further reduce the number of dimensions. It has the added advantage that it provides the relative importance of words in each document and in the entire document collection.

It also generalizes upon the stop word removal technique.

The basic idea is that the relative importance of a term in the document is directly proportional to the frequency of that term in that document. However if that term is also frequent in several other documents, then its relative importance in the original document is reduced. Thus its relative importance in the original document is reduced. Thus its relative importance is inversely proportional to the number of documents that contain it.

The precise formulae used to compute the TF-IDF values differ slightly according to implementation. Care is taken to avoid improper division by zero and to normalize the values to a logarithmic space.

These TF-IDF values are computed for every term. Only terms that have a TD-IDF value more than some threshold is considered worthy of inclusion into the vector space model. Other terms are neglected. This greatly reduces the number of dimensions that one must deal with.

However there are also some tasks that are unique to this kind of data. These include text and web search, mining themes and hot topics, document understanding.

Whenever search is performed, the user typically gives a set of keyword as a query and the required output is the set of documents that contain all those keywords. Usually the number of documents that match a user's query is huge. We need a way to organize the results. This way the user can choose to see only the category of results of interest. To solve this problem, we need an effective way to organize documents such that only the relevant documents for a query need to be retrieved and processed. The approach to be solved first is too many results problem and then efficiency problem. The relative frequency or TF-IDF score of the keywords in the document is also an indication of its relevance to the query.

2.1 Frequent patterns:

For many kind of analysis, the frequent patterns discovered from a document database can be used instead of the original database. So frequent Item Set Mining has been an essential part of data analysis and data mining which extracts information from database based on frequently occurring events according to the user given minimum frequency threshold.

3. CLUSTERING TECHNIQUES -LITERATURE REVIEW

3.1 Traditional Text Clustering Techniques:

Generally there are two categories of clustering methods: partioning method and agglomerative hierarchal methods are applied in text clustering.

D.R.Cutting et,al.[1] uses the k-means algorithm uses the idea of Centroid can represent a cluster, after selecting a k initial Centroid each document is assigned to a cluster based on distance measure, then k centroids are recalculated. This step is repeated until an optimal set of k clusters are obtained based on heuristic function.

A.K.Jain et.al.[2] proposed Agglomerative hierarchal clustering algorithms initially treat each document as a cluster, use different kinds of distance functions to compute the similarity between all pair of clusters, and then merge the closest pair.

Congan Lu et al.[3] proposed three different methods of using the neighbors and link in the k-means and bisecting k-means algorithms for document clustering. Comparing with the local information given by the cosine function, the link function provides the global view in evaluating the closeness between two documents by using the neighbor documents.

3.2. Clustering methods using frequent item sets.

H.Edith et al [4] proposed CMS method .The basic idea of CMS is to use maximal frequent sequences (MFS) of words as features in vector space model for document representation and then k-means is employed to group documents into clusters.

Beil et.al [5] proposed a method called frequent term -Based Clustering (FTC) for document clustering. The motivation of FTC is to produce a document clusters with overlaps as few as possible.

B.Fung et.al [6] proposed Frequent Item set-based Hierarchal clustering (FIHC) to produce document clusters by finding global frequent items, initial clustering, tree clustering ad pruning.

Wen Zhang et.al [7] proposed a new method called Maximum Capturing for document clustering using frequent patterns which includes three similarity measures proved better than other methods like CFWS, CMS, FTC, FIHC.

Sandy Moens et al [8] proposed the recent work is done on two frequent item set mining algorithms for map-reduce where Dist-Éclat focuses on speed by using a simple load balancing scheme based on k-FIs. BigFIM focuses on mining very large databases by utilizing hybrid approach.

3.3. Clustering methods using Optimization techniques.

Wei Song et.al [9] proposed a variable string length genetic algorithm for automatically evolving the proper number of clusters as well as providing near optimal data set clustering.

Eisa Hasanzadeh [10], [11] proposed work PSO+LSI are faster than PSO+Kmeans algorithms using the vector space model for all numbers of dimensions.

Stuti Karol et al [12] introduced hybrid PSO based algorithm. The two partitioning clustering algorithms Fuzzy C-Means (FCM) and K- Means each hybridized with Particle Swarm Optimization (KPSO and FCPSO). The performance of hybrid algorithms provided better document clusters against traditional partitioning clustering techniques (K-Means and Fuzzy C Means) without hybridization.

Nihal M. AbdelHamid et.al [13] proposed algorithm that has been tested on a data set containing 818 documents and the results have revealed that the algorithm achieved its robustness. This model was compared with the Genetic Algorithm and K-means and it was concluded that Bees algorithm outperforms GA by 15% and the K-means by 50%. And also the results revealed that the Bees Algorithm takes more time than the Genetic Algorithm by 20% and the K-means by 55%.

Kayvan Azaryuonet.al [14] proposed shows that the proposed algorithm presents a better average performance than the standard ants clustering algorithm, and the K-means algorithm.

4. EVALUATION METHODS:

Evaluation methods like F-measure is used to estimate performances of the clustering results are normalized.

The formula for F-measure is depicted as Equations,

$$P(i, j) = \frac{n_{ij}}{n_j}$$

$$R(i, j) = \frac{n_{ij}}{n_i}$$

$$F(i, j) = \frac{2*P(i,j)*R(i,j)}{P(i,j)+R(i,j)}$$

$$F = \sum_i \frac{n_i}{n_j} \max F(i, j)$$

Here n_i is the number of documents of class i, n_j is the number of documents of cluster j, and n_{ij} is the number of documents of class i in cluster j, n is the total number of documents in the collection. P(i,j) is the precision of cluster j in class i, R(i,j) is the recall of class i in cluster j, F(I,J) is the measure of cluster j in class i.In general, the larger the F-measure is , the better the clustering result.

5. FUTURE TRENDS

During the literature review certain research gap has been noticed which are the scalability problem where the existing model does not work on large or big data sets, Efficient, parallel and incremental clustering techniques is required, Due to the high dimension of text documents effective techniques are required to reduce the size. And once the clusters are made the cluster topics using good topic modeling should be provided.

And also automatic clustering should be implemented by finding the number of clusters to the clustering algorithm by itself has yet to be done.

One peculiar issue with respect to document clustering is that typically a document may belong to more than one cluster. This is an important aspect to note while designing document clustering algorithms. More research works have to be carried out based on semantic to make the quality of text document clustering.

6. CONCLUSION

Since clustering is often the foremost data mining task applied on a given collection of data and is used to explore the possibility of any underlying patterns existing in the data. This paper has presented a survey on the research work done on text document clustering based traditional partitioned and hierarchal algorithm, using frequent patterns, and also the optimization techniques.

More research works have to be carried out to make the quality of text document clustering.

REFERENCES

[1] D.R.Cutting,D.R.Karger,J.O.Pederson,J.W.Tukey,Scatter/gather: a cluster-based approach to browsing a large document collections, in: Proceedings of Annual ACM SIGIR Conference on Research and Development in information Retrieval, 1992, pp.318-329.

[2] A.K.Jain,R.C.Dubes, Algorithms for Clustering Data, Prentice Hall, Englewood Cliffs, 1998.

[3] Congan Luo, Yanjun Li, Soon M.Chung. Text document clustering based on neighbors. Doi:10.1016/j.datak.2009.06.007.pp.1271-1288.

[4] H.Edith ,A.G.Rene.J.A.Carrasco-Ochoa, J.F.Martinez-Trinidad, Document clustering based on maximal frequent sequences, in: Proceedings of FinTAL 2006,ln.vol.4139,2006,00.257-267.

[5] F.Beil, M.Ester, X.W.Frequent term based text clustering, in: Proceedings of the 8th ACM SIGKDD International Conference on knowledge Discovery and Data Mining, 2002, pp. 436-442.

[6] B.Fung, K.Wang, M.Ester, Hierarchal document clustering using frequent item sets, in: Proceedings of the 3rd SIAM International Conference on Data Mining, 2003.

[7] Wen Zhang, Taketoshi Yoshida, Xijin Tang, Qing Wang. Text Clustering using frequent item sets. Doi:10.1010/j.knosys.2010.01.011. pp.379-388.

[8] Sandy Moens, Emin Aksehirli and Bart Goethals. Frequent Item set Mining for Big data. 2014.

[9] Wei Song & Soon Cheol Park,(2009), ―Genetic Algorithm for text clustering based on latent semantic indexing, Computers and Mathematics with applications‖, 57, 1901-1907.

[10] Eisa Hasanzadeh & Hamid Hasanpour, (2010),―PSO Algorithm for Text Clustering Based on Latent Semantic Indexing‖, The Fourth Iran Data Mining Conference , Sharif University of Technology, Tehran, Iran.

[11] Eisa Hasanzadeh, Morteza Poyan rad and Hamid Alinejad Rokny,(2012),Text clustering on latent semantic indexing with particle swarm optimization (PSO) algorithm‖, International Journal of the Physical Sciences Vol. 7(1), pp. 116 – 120.

[12] Stuti Karol , Veenu Mangat, (2012), ― Evaluation of a Text Document Clustering Approach based on Particle Swarm Optimization‖, CSI Journal of Computing , Vol. 1 , No.3.

[13] Nihal M. AbdelHamid, M. B. Abdel Halim, M. Waleed Fakhr, (2013), ―BEES ALGORITHM-BASED DOCUMENT CLUSTERING‖, ICIT 2013 the 6th International Conference on Information Technology.

[14] Kayvan Azaryuon , Babak Fakhar,(2013), ―A Novel Document Clustering Algorithm Based on Ant Colony Optimization Algorithm‖, Journal of mathematics and computer Science Vol.7 , pp. 171-180.