A Survey on Document Clustering Approach with Multi-Viewpoint based on different similarity measures

[1]Anjali Gupta, [2]Rahul Dubey
[1]Student, [2]Professors
[1]Cse department,[2]Cse department
SAGAR Institute of Science & Technology
Gandhi Nagar,Bhopal Madhya Pradesh, India

Abstract: There is a fast-growing large amount of available electronic information such as online newspapers, journals, conference proceedings, Web sites, e-mails, etc. Using all this electronic information controlling, indexing or searching is not possible for human and for search engines also. Thus, automatic document organization is a critical issue. By using document clustering methods, we can understand data distribution or we can pre-process data for other applications. For an example, if a search engine uses clustered documents to search an item or data, it can produce results more effectively and efficiently.

Document clustering is an approach of automatic clustering operation of text documents so that related documents are presented in same cluster, unrelated documents are presented in different clusters. Document Clustering is a technique of an unsupervised learning to group related documents into a cluster, each cluster consist of documents that are related to one another within the same clusters and are unrelated to documents belonging to other cluster. Before applying any clustering methods, a similarity (related)/distance measures must be determined. The similarity(related) measure reflects the degree of closeness of the target objects.

In this paper, we introduce document clustering on Multiview point based similarity measure and two related clustering methods. The existing traditional dissimilarity/similarity based document clustering measure uses only a single viewpoint. This is the origin, while ours usesmany other different viewpoints.

Index Terms: Text mining, document clustering and information extraction

I. INTRODUCTION

Knowledge discovery is also known as data mining is a process of finding interesting unknown knowledge from a vast collection of data and moreover data mining is the process of analysing data from different point of view and summarizing useful information from it. Data mining is the process of technically finding patterns and correlations among variety of fields in database. These data are in any forms such as it can be text, numbers, images, audio, video facts that can be processed by computer. Data clustering which is the unsupervised classification is one of the important techniques of data mining, where similar data objects are groups into clusters so that data in each cluster share some common characteristics or properties according to defined distance measure. Document clustering is the process of arranging a collection of text documents into clusters based on some similarity measures. Document within same clusters are same with each other each other than those documents belong to a different cluster. Many of the algorithms are available to performing document clustering. The mainly used document clustering algorithm is k-mean. It is most widely used document clustering algorithm, it uses sum of square error objective function that uses Euclidean distance for similarity measures. Effectiveness of clustering algorithms depends on the appropriateness of the similarity measures. It seems that success and failure of clustering algorithm depends on the natures of similarity measures.

II. LITERATURE SURVEY

There are number of clustering approaches. These are partitioning (e.g. K-means, kmedoids), hierarchical (e.g. DIANA, AGNES, BIRCH), density-based (e.g. DBSACN, OPTICS), grid-based (e.g. STING, CLIQUE), model based (e.g. EM, COBWEB), frequent pattern-based (e.g. p-Cluster), constraint-based (e.g. COD), and link-based (e.g. SimRank, LinkClus) clustering approaches [1].

The widely used method which is applied to documents is hierarchical clustering method. In 1988, Willett applied agglomerative clustering methods to documents by altering the calculation method of distance between clusters [3]. These algorithms have many problems with clusters that finding stopping point is very difficult and they run too slowly for thousands of documents.

Hierarchical clustering algorithms are applied to documents for number of times by Zhao and Karypis [4,5] and in 2005 they tried to improve agglomerative clustering algorithm by appending constrains [6]. Hierarchical clustering is often depicted as the better-quality clustering approach, but is limited because of its quadratic time complexity.

K-means and its variants, which are partitioning clustering algorithms that create a non-hierarchical clustering consisting of k clusters, are applied to documents [2]. They are more efficient and scalable, and their complexity is linear to the number of documents. A disadvantage of k-means is that estimating the value of k wrongly leads worse accuracy. Furthermore, k-means can have stuck on a local maximum because of randomly chosen initial centroids. K-means as well as its variants have a time complexity whichis linear in the number of documents, but are thought to produce inferior clusters.To solve this problem, Kaufman and Rousseeuw proposed k-medoids 4 algorithm but this algorithm is computationally much more expensive and does not scale good with huge document sets [2].

Sometimes K-means and agglomerative hierarchical approaches are combined to “get the best of both worlds.

III. DOCUMENT CLUSTERING PROCESS

Clustering is an automatic knowledge method meant at combination a set of object into subsets or clusters. Themain objective is to generate clusters that are consistent inside, however significantly dissimilar from each other [8]. In very simplelanguage, substance in the same group should be as analogous as possible, while matter in one cluster should be as different as possible from matter in the other clusters.

Clustering is the most usual form of unsupervised learning which deals with finding a structure in a collection of unlabelledor raw data.Automatic grouping of text document in a cluster which has high similarity in comparison to one another, but is different from document in other clusters is a method of document clustering. It is important to spotlight that getting from a collection of document to a clustering of the collection is not only a single process, but is a process of multiple stage.

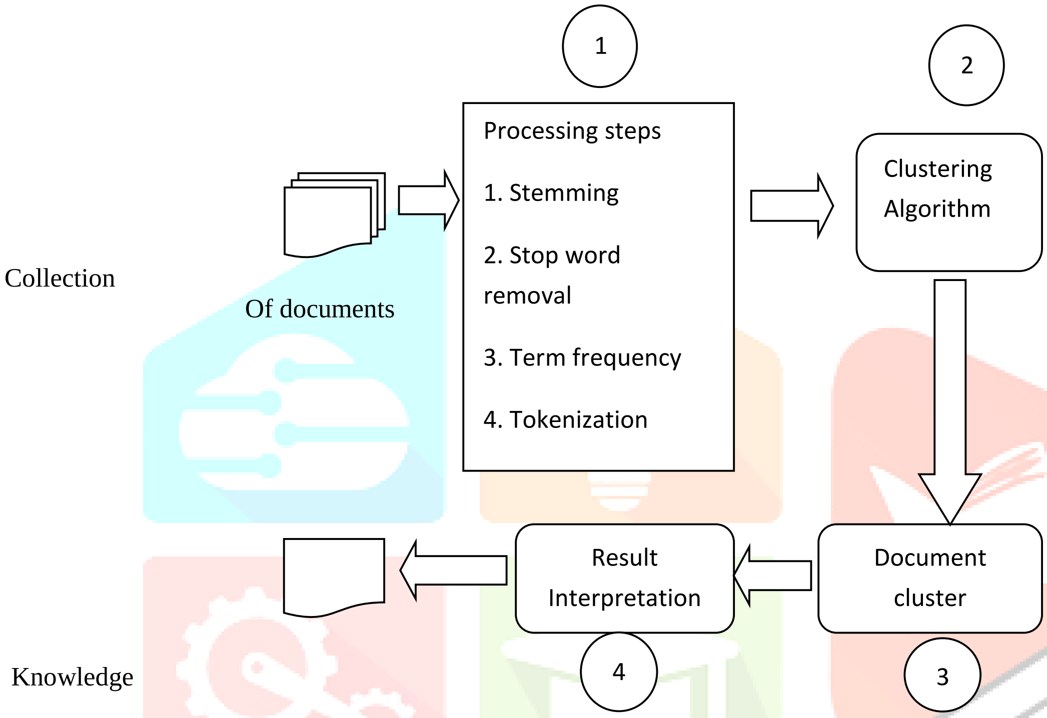

figure1: document clustering process

3.1 COLLECTION OF DATA:

Method used to collect the documents that needs to be clustered is like crawling, indexing and filtering etc.

3.1.1 Pre-processing Steps:

Pre-processing is step where data is pre-processed to represent the data in a form that can be used for clustering.

Stemming:

Stemming is a technique for the reduction of words into their steam or base form many words e.g. same, similar, dissimilar, similarity, and non-similar belong to similar.

Stop word Removal

Prepositions, articles, and pronouns etc are the most familiar words in any text document do not provide meaning of the document. These words are removed. These words are not important for text mining application.

Term Frequency

Theterm frequency is a method of feature selection in document clustering, document frequency that is used to filter out irrelevant or unimportant feature. In other word, words which are too frequent in the collection can be removed.

Tokenization

Splits sentences into separates tokens, the main use of tokenization is to recognise meaningful keyword.

3.2 CHALLENGES IN DOCUMENT CLUSTERING

Document clustering is being studied from many years but still it is far from insignificant and solved problem. The challenges are [8]:

- Selection of required features of the documents.
- Selection of necessary similarity measure.

- Selection of right clustering method.
- Evolution of the quality of the clusters.
- Implementation of the clustering algorithm in an efficient manner so that it make optimal use of available memory and CPU resources.
- Associate meaningful label to each final cluster [9].
- To consider semantic relationship between words like synonyms like synonyms of submit is assert is treated as submit.

3.3 CLUSTERING ALGORITHM

The clustering algorithm is used in the process of documents analysis. These methods convert unstructured document to structured document for further investigation. In this work we used differentkind of clustering algorithm as follows*:*

3.3.1 K-Means

K-means is a kind of clustering method. K-means is algorithm to find the positions of clusters that minimize the distance from data points to the cluster.

Method

I. First find the centroid-Randomly select the centroid at 1st iteration
II. Then assign cluster to each value of attributes.
III. And repeat the all process until error rate stop changing and cluster also stop changing.

K-means which is an essentialflat clustering algorithm which allow minimizing the average squared distance of objects from their cluster center, here a cluster center is defined as the mean or centroid µ of the objects in a cluster C:

In K-means the ideal cluster is a sphere with the centroid as its centre of gravity. Ideally, K-mean considers the clusters should not overlap. K-mean use a term Residual Sum of Squares (RSS) which measures how well the centroids represent the members of their clusters. RSS is squared distance of each vector from its centroid summed over all vectors.

K-means select initial clusters centres K randomly chosen objects, they mainly named as the seeds. To minimize RSS it then moves the cluster centreson in space. This procedure is done iteratively by repeating two steps until a stopping criterion is met.

1. The one is reassigning objects to the cluster with closest centroid.
2. And other is re-computing each centroid based on the current members of its cluster.

Stopping criterion as termination condition is given as following:

- Stable number of iterations I has been completed.
- Steps continue until centroids µ do not change between iterations.
- End up when RSS falls below a pre-established threshold.

3.3.2 K-Medoids

The k-mean algorithm is sensitive to outlier. It is sensitive because an object with so manyvalues may substantially distort the distribution of data. The problem of K-mean can be overcome by using K-medoids to represent the cluster in place of centroidhere;medoid is the center data object in a cluster. K-Medoids selectk data objects are randomly as medoids to represent k-cluster while remaining all data objects are placed in a cluster having medoids nearest to that objects. New medoid is evaluated after processing all data objects, which gives cluster in a better way and the entire process is repeated. All data objects are grouped to the cluster again based on the new medoids.Eachiterationchanges the medoids location step by step. This process is remains same until there is no move.

K-Medoidsresult;itcreates k-cluster which represents a set of n data objects.

3.3.3Expectation Maximization

The Expectation Maximization is also called as Model based clustering algorithm. It is a subcategory of the flat clustering algorithm. The model-based clustering generates dataset and then tries to recover the model from the data. This model is then used to define clusters and the cluster membership of data.

The EM algorithm is a generalized form of K-Means algorithm which considers k-centroide as model then generates the original data. It alternate among expectation step, corresponding to reassignment, and a maximization step, corresponding to re-computation of the parameters of the model.

3.3.4 Hierarchical Clustering

Hierarchical Clustering techniques cluster collections of documents in the form of tree structure or in hierarchy structure. They are typically called dendogram. These types of clusters are constructed using two approaches bottom-up or top-down approaches. The tree's root represents one cluster and node represent all data points, while at the levels of tree there are n clusters. Hierarchical clustering approaches create a hierarchical decomposition of given collection of documents and create hierarchical structure.

A hierarchical clustering algorithm creates a hierarchical decomposition of the given set of data objects. Depending on the decomposition approach, hierarchical algorithms are classified as agglomerative (merging) or divisive (splitting). Hierarchical Clustering algorithms are further classified into two categories: agglomerative and divisive;

A) Agglomerative Clustering:

The agglomerative clustering supports a variety of searching methods as it creates a hierarchy of tree which is used for the searching operations. These algorithms merge document successively based on their similarity into clusters. The agglomerative approach starts with a large number of clusters. This approach merges the two clusters on each step based on some similarity measures. Thus after each step, the cluster's number decreases tremendously. This process is continuously repeated until the desired number of clusters is form or repeated until only one cluster left.

Agglomerative methods used an initial clustering of the term space, where all documents are represented as separate cluster. The closest clusters uses given inter-cluster similarity measure are then they are merged continuously. It continueuntil only 1 cluster or a predefined number of clusters remain.

Steps of simple Agglomerative Clustering Algorithm:

1. Compute the similarity between all pairs of clusters in other words itcalculates a similarity matrix whose ij entry gives the similarity between the I and j clusters.
2. Merge the most similar (closest) object to clusters.
3. Update the similarity matrix to show the pair wise similarity between the new cluster and the original clusters.
4. Steps 2 and 3 repeat until only a single cluster remains.

B) Divisive Clustering:

While in divisive clustering it considered all data points as one clusters and recursively splits it into most appropriate clusters until some stopping criteria is achieved the process will continue. In divisive approach, one cluster containing all data objects. In every step, cluster is split into smaller clusters, until some coverage's conditions holds. Agglomerative algorithms are very popular clustering algorithm.

Divisive clustering algorithms contain all documents in one cluster. It then continuously divides clusters until all documents are grouped in predefined number of clusters. Agglomerative algorithms are usually further classified according to the inter-cluster similarity measure they use. They are single-link, complete-link and group average.

3.4 SIMILARITY MEASURE:

Analysis of Clustering methods are depends on measurements of the similarity between a pair of objects. To calculate similarity between a pair of objects involve three major steps:

A) The one is selection of the variables; it is used to characterize the objects,
B)For these variables the selection of a weighting scheme
C) Other is the selection of a similarity coefficient to evaluate the degree of resemblance between the two attribute vectors [11].

For the accurate of clustering requires an exact definition of the closeness between a pair of objects, it may in terms of either the pair-wise similarity or distance. A large variety of similarity or distance measures have been proposed and widely applied, such as cosine similarity, Jaccard correlation coefficient, Euclidean distance, and relative entropy [10].

Few similarity measures that are widely used [10]:

I. **Euclidean Distance:** It is used for geometrical problems as standard matrices. It is the just a simple distance between two points and basically can be easily measured with a ruler. It is also used in K-mean algorithm as default distance measure used in K-means algorithm.

II. **Cosine Similarity:** Cosine similarity is the similarity of two documents corresponds to the correlation between the vectors. This is evaluated as the cosine of the angle between vectors, that is, cosine similarity.

III. **Jaccard Coefficient:** The Jaccard coefficient compares the sum weight of shared terms with the sum weight of terms that are present in either of the two documents but are not the shared terms.

Two principle ways have been proposed to calculate the similarity between two documents *di* and *dj*:
The first method is calculating similarity measure between two document usingcosine functionon the commonly-used (Salton, 1989):

$$\cos(di, dj)_{=} di^{t}dj/(\|di\|\ \|dj\|) \qquad (3.1)$$

Since the document vectors are of unit length, it simplifies to$di^{t}dj$.

The second method calculated the similaritymeasure between the documents using the Euclidean distance *dis :*

$$(di, dj) = \|di - dj\|. \ (3.2)$$

Note that apart from this one measures similarity and the other measures distance, these measures are quite close to each other because the document vectors are of unit length.

IV. CONCIUSION

In this paper, Clustering Approach with Multi-viewpoint based on similarity measure ispresented. Theoretical this analysis show that Multi-viewpoint based similarity measure is probable more suitable as compare to existing single point dissimilarity/ similarity measure for text document. This survey initiated with a brief introduction about clustering in data mining and explored various research papers related to text document clustering. More research works have to be carried out based on semantic to improve the quality of text document clustering.

V. ACKNOWLEDGEMENT

I would like to give my sincere gratitude to my guide Prof.Rahuldubeywho encouraged and guided me throughout this paper. I also thanks to PG CO-ORDINATOR, Prof.Rajeevkumargupta for their valuable guidance and for allowing me to use the college resources and facilities provided by them.

VI. REFRENCES

[1] Han, J. and Kamber, M., *Data Mining: Concepts and Techniques, 2nd ed*., Morgan Kaufmann Publishers, 2006.

[2] Kaufman, L. and Rousseeuw, P., *Finding Groups in Data*, Wiley, New York, NY, 1990

[4] Zhao, Y., and Karypis, G., "Criterion functions for document clustering: Experiments and analysis", *Technical Report*, Department of Computer Science, University of Minnesota, 2001.

[5] Zhao, Y., and Karypis, G., "Evaluation of hierarchical clustering algorithms for document datasets", *International Conference on Information and Knowledge Management*, McLean, Virginia, United States, pp.515-524, 2002.

[6] Zhao, Y. and Karypis, G., "Hierarchical Clustering Algorithms for Document Datasets", *Data Mining and Knowledge Discovery* [C].10(2), pp.141-168, 2005.

[7] Steinbach, M., Karypis, G. and Kumar, V., "A Comparison of Document Clustering Techniques", *KDD Workshop on Text Mining*, 2000.

[8] Pankaj Jajoo, "Document Clustering," Masters' Thesis, IIT Kharagpur, 2008

[9] RekhaBaghel and Dr. RenuDhir, "A Frequent Concepts Based Document Clustering Algorithm," International JournalofComputer Applications, vol. 4, No.5, pp. 0975 – 8887, Jul. 2010

[10] A. Huang, "Similarity measures for text document clustering," In *Proc. of the Sixth New Zealand Computer Science Research Student Conference NZCSRSC*, pp. 49—56, 2008.

[11]Priti B. Kudal,Prof. M.M.Naoghare,‖A Review of Modern Document Clustering Techniques‖,International Journal of Science & Research(IJSR), Volume 3 Issue 10, October 2014

[12]Di WU1, Yan ZENG2 and Yin-chuanQU,Text document clustering based on density k-mean 2016 International Conference on Computer, Mechatronics and Electronic Engineering (CMEE 2016)