International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 2 Issue 12, December - 2013

Document Retrieval in Clustering using Multiple Viewpoints Similarity Measure

Ms. Nithya S Prasad (P.G scholar)
Dept. of Computer Science and Engg.
College Of Engineering Perumon
Kollam, Kerala, India

Mrs. Deepa K Daniel (Assistant Professor)
Dept. of Information Technology
College Of Engineering Perumon
Kollam, Kerala, India

Abstract— Clustering is the process of organizing objects into groups whose members are similar in some way. Similarity between objects can be measured either explicitly or implicitly. Traditional clustering methods use only a single viewpoint which is the origin. To construct a new concept of similarity, it is possible to use more than one point of reference. Thus may have a more accurate assessment of how close or distant a pair of points is, if look at them from many different viewpoints. This proposal is called Multiviewpoint based Similarity, or MVS. In MVS, the two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster. It is potentially more suitable for text documents. We compare the proposed method with well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal. Finally implement a novel approach for document retrieval using hierarchical agglomerative clustering based on multiviewpoint similarity measure.

Index Terms— Clustering, Document clustering, Text mining, Similarity measure.

I.INTRODUCTION

Clustering is a group of documents or items which are most related to each other. Clustering is the process of organizing objects into groups whose members are similar in some way. Cluster analysis[1] itself is not one specific algorithm, but the general task to be solved. It can be achieved by various algorithms that differ significantly in their notion of what constitutes a cluster and how to efficiently find them. Popular notions of clusters include groups with low distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings (including values such as the distance function to use, a density threshold or the number of expected clusters) depend on the individual data set and intended use of the results. Cluster analysis as such is not an automatic task, but an iterative process of knowledge discovery or interactive multi- objective optimization that involves trial and failure. It will often be necessary to modify preprocessing and parameters until the result achieves the desired properties. The subtle differences are often in the usage of the results: while in data mining, the resulting groups are the matter of interest, in automatic classification primarily their discriminative power is of interest.

Clustering can be divided into two categories: Partitional clustering (PC) and Hierarchical clustering (HC).

Partitional clustering (PC) algorithms relocate instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. To achieve global optimality in partition-based clustering, an exhaustive enumeration process of all possible partitions is required. Because this is not feasible, certain greedy heuristics are used in the form of iterative optimization. Namely, a relocation method iteratively relocates points between the *k* clusters.

Hierarchical clustering (HC) algorithms organize data into a hierarchical structure according to the proximity matrix. The results of HC are usually depicted by a binary tree or dendrogram. The root node of the dendrogram represents the whole data set and each leaf node is regarded as a data object. The height of the dendrogram usually expresses the distance between each pair of objects or clusters, or an object and a cluster. HC algorithms are mainly classified as agglomerative methods and divisive methods.

Agglomerative hierarchical clustering starts with N clusters and each of them includes exactly one object. Then clusters are successively merged until the desired cluster structure is obtained. Divisive hierarchical clustering proceeds in an opposite way. In the beginning, the entire data set belongs to a cluster and a procedure successively divides it until all clusters are singleton clusters. That is all

objects initially belong to one cluster. Then the cluster is divided into sub-clusters, which are successively divided into their own sub-clusters. This process continues until the desired cluster structure is obtained. Divisive clustering is not commonly used in practice.

The rest of this paper is organized as follows: Section II describes the previous work in related domains. Section III presents the proposed method i.e. document retrieval in clustering using MVS. The experimental results are given in Section IV. Finally, a short discussion on conclusions and future study are provide in Section V.

II.RELATED WORK

Clustering [2] is a useful technique that organizes a large quantity of unordered objects into a small number of meaningful and coherent clusters, thereby providing a basis for intuitive and informative navigation and browsing mechanism. Thus a cluster can be defined as a collection of objects which are 'similar' between them and are 'dissimilar' to the objects belonging to other clusters; and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set.

In data mining, clustering is of two types: Partitional clustering and Hierarchichal clustering. Partitional clustering [3] relocates instances by moving them from one cluster to another, starting from an initial partitioning. Such methods typically require that the number of clusters will be pre-set by the user. Hierarchical clustering [4] seeks to build a hierarchy of clusters. Strategies for hierarchical clustering generally fall into two types:
Agglomerative: This is a "bottom up" approach: each observation starts in its own cluster, and pairs of clusters are merged as one moves up the hierarchy.
Divisive: This is a "top down" approach: all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy.

Before clustering, a similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis, especially for a particular type of clustering algorithms. Recalling that closeness is quantified as the distance/similarity value and there are large number of distance/similarity computations which are required for estimating cluster assignment of new data objects. Therefore, understanding the effectiveness of different measures is of great importance in helping to choose the best one. In general, similarity/distance measures map the distance or similarity between the symbolic description of two objects into a single numeric value, which depends on two factors— the properties of the two objects and the measure itself.

Not every distance measure is a metric. To qualify as a metric, a measure d must satisfy the following four conditions.

Let x and y be any two objects in a set and d(x, y) be the distance between x and y.

1. The distance between any two points must be nonnegative, that is, $\mathbf{d(x,y) \geq 0}$.

2. The distance between two objects must be zero if and only if the two objects are identical, that is, d(x, y) = 0 if and only if x = y.

3. Distance must be symmetric, that is, distance from x to y is the same as the distance from y to x, ie. d(x, y) = d(y, x).

4. The measure must satisfy the triangle inequality i.e; $\mathbf{d(x,z) \leq d(x,y) + d(y,z)}$.

A wide variety of distance functions and similarity measures have been used for clustering, such as Euclidean distance, cosine similarity, etc.

Euclidean distance [5] is a standard metric for geometrical problems. It is the ordinary distance between two points and can be easily measured with a ruler in two- or three-dimensional space. Euclidean distance is widely used in clustering problems, including clustering text. It satisfies all the above four conditions and therefore is a true metric. It is also the default distance measure used with the K-means algorithm.

Measuring distance between text documents, given two documents $\mathbf{d}_a$ and $\mathbf{d}_b$ represented by their term vectors $\vec{\mathbf{t}}_a$ and $\vec{\mathbf{t}}_b$ respectively, the Euclidean distance of the two documents is defined as:

$$D_E(\vec{t_a}, \vec{t_b}) = (\sum_{t=1}^{m} |w_{t,a} - w_{t,b}|^2)^{1/2} \qquad \text{Eqn (2.1)}$$

where the term set is T = {t1, . . . , tm}. As mentioned previously, tfidf value is used as term weights, that is $w_{t,a} = tfidf(d_{a,t})$.

When documents are represented as term vectors, the similarity of two documents corresponds to the correlation between the vectors. This is quantified as the cosine of the angle between vectors, that is, the so-called cosine similarity. Cosine similarity [6] is one of the most popular similarity measure applied to text documents, such as in numerous information retrieval applications and clustering too.

Given two documents $\vec{t_a}$ and $\vec{t_b}$, their cosine similarity is

$$SIM_C(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}| \times |\vec{t_b}|} \qquad \text{Eqn (2.2)}$$

where $\vec{t_a}$ and $\vec{t_b}$ are m-dimensional vectors over the term set T = {t1, . . . , t_m}. Each dimension represents a term with its weight in the document, which is non-negative. As a result, the cosine similarity is non-negative and bounded between [0,1].

An important property of the cosine similarity is its independence of document length. For example, combining two identical copies of a document d to get a new pseudo document d', the cosine similarity between d and d' is 1, which means that these two documents are regarded to be identical. Meanwhile, given another document l, d and d' will have the same similarity value to l, that is, $sim(\vec{t_d}, \vec{t_l}) = sim(\vec{t_{d'}}, \vec{t_l})$. In other words, documents with the same composition but different totals will be treated identically. Strictly speaking, this does not satisfy the second condition of a metric, because after all the combination of two copies is a different object from the original document. However, in practice, when the term vectors are normalized to a unit length such as 1, and in this case the representation of d and d' is the same.

The Jaccard coefficient [7], which is sometimes referred to as the Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects. For text document, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two document but are not the shared terms. The formal definition is:

$$SIM_J(\vec{t_a}, \vec{t_b}) = \frac{\vec{t_a} \cdot \vec{t_b}}{|\vec{t_a}|^2 + |\vec{t_b}|^2 - \vec{t_a} \cdot \vec{t_b}} \qquad \text{Eqn (2.3)}$$

The Jaccard coefficient is a similarity measure and ranges between 0 and 1. It is 1 when the $\vec{t_a} = \vec{t_b}$ and 0 when $\vec{t_a}$ and $\vec{t_b}$ are disjoint, where 1 means the two objects are the same and 0 means they are completely different. The corresponding distance measure is defined by $w_t = \Pi_1 \times w_{t,a} + \Pi_2 \times w_{t,b}$ $D_J = 1 - SIM_J$, and it will use D_J instead.

It is possible to use more than just one point of reference. Thus may have a more accurate assessment of how close or distant a pair of points is, if look at them from many different viewpoints. The two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster. This proposal is called Multiviewpoint-based Similarity, or MVS [8]. The similarity of two documents d_i and d_j, given that they are in the same cluster is defined as the average of similarities measured relatively from the views of all other documents outside that cluster. From this point onwards the proposed similarity measure between two document vectors d_i and d_j will be denoted as $MVS(d_i, d_j / d_i, d_j \in S_r)$ or occasionally $MVS(d_i, d_j)$ for short.

The similarity between two points d_i and d_j inside cluster S_r, viewed from a point d_h outside this cluster, is equal to the product of the cosine of the angle between d_i and d_j looking from d_h and the euclidean distances from d_h to these two points. This definition is based on the assumption that d_h is not in the same cluster with d_i and d_j.

$$MVS(d_i, d_j | d_i, d_j \in S_r) = \sum_{d_h} \cos(d_i - d_h, d_j - d_h) \|d_i - d_h\| \|d_j - d_h\| \qquad \text{Eqn (2.4)}$$

The smaller the distances $|d_i - d_h|$ and $|d_j - d_h|$ are, the higher the chance that d_h is in fact in the same cluster with d_i and d_j. The overall similarity

between d_i and d_j is determined by taking average over all the viewpoints not belonging to cluster S_r.

$$\mathrm{MVS}(d_i, d_j \, d_i, d_j \in S_r) = \frac{1}{n-n_r} \sum_{d_h} \cos(d_i - d_h, d_j - d_h) \| d_i - d_h \| \| d_j - d_h \| \quad \text{Eqn (2.5)}$$

It can be seen that this method offers more informative assessment of similarity than the single origin point-based similarity measure.

III. PROPOSED METHOD

Clustering is one of the most interesting and important topics in data mining. The aim of clustering is to find intrinsic structures in data, and organize them into meaningful subgroups for further study and analysis. There have been many clustering algorithms published every year. The existing clustering includes:

- The internal structure of the data will be find and organize them into a meaningful groups.
- It greedily picks the next frequent item set in the next cluster.
- The clustering result depends on the order of picking up the item sets.
- Cosine similarity is used to find out the dissimilar document object in the cluster.
- Existing system proposed a multiviewpoint algorithm for move the dissimilar document object from one cluster to another cluster.
- The second similarity measures similarity between the dissimilar document object and the other cluster group's document objects.

The drawbacks of existing clustering are:

- Document is moved based on frequent occurrence of next cluster.
- Low efficiency and performance.
- Cluster movement will quite complexity. Sometimes the similarity process takes a long period of time.
- Clustering accuracy is low.
- Returns an unstructured set of clusters.
- Large number of scanning.
- Reduce the clustering quality.

Hierarchical Agglomerative Clustering based on multiviewpoint similarity measure includes:

- Propose a new method to group the documents into cluster.
- The documents are collected and perform the preprocessing step for stemming and stop word removing.
- Multiviewpoint based similarity calculation is used for measuring similarity between data objects.
- Similarity measures will depend on the text mining
- With the proposed similarity measure MVS, we then implement Hierarchical Clustering Algorithm which forms the document groups.
- For this, use cosine similarity for find out the dissimilar document object in the cluster.
- Propose multiviewpoint similarity measure for move the dissimilar document object from one cluster to another cluster.
- The second similarity measures similarity between the dissimilar document object and the other cluster group's document objects.
- From the clustered objects, the document retrieval can be done based on the query.
- Ranking is provided for the Clusters with respect to the query matching result.

Hierarchical multiviewpoint similarity uses correlation similarity and cosine similarity to measure the similarity between objects in the same cluster and dissimilarity between objects in the different cluster groups. The architecture of the Hierarchical MVS is shown in the Fig 3.1.

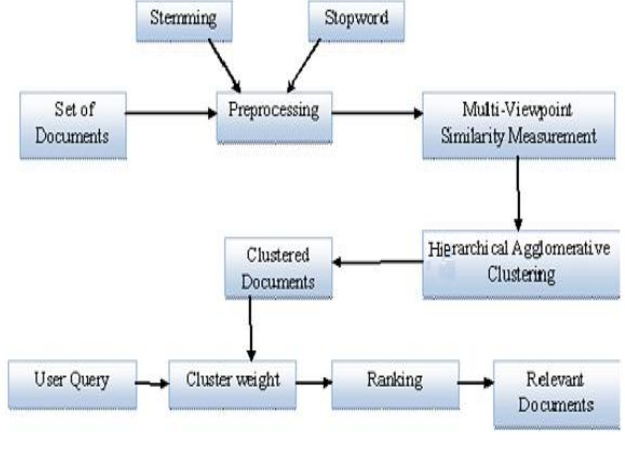

Fig 3.1: Architecture of Hierarchical Multiviewpoint Similarity

The set of documents are taken as input, then each block performs the operations on the documents to form the final hierarchical clustering. Preprocessing is done in two steps i.e removal of stopwords and stemming. Stop-words [19] are very common words that do not provide any useful information to us, such as "and", "the", "which", "is", etc.. It is often useful to get rid of these words otherwise they might mislead the clustering process by including frequent terms that are not informative to us. Word stemming [10] is the process of converting different forms of a word into one canonical form. Words like "compute", "computing", "computer" are all changed to a single word "compute". This is necessary to avoid treating different variations of a word distinctly.

To implement similarity measure, it is possible to use more than just one point of reference. Thus may have a more accurate assessment of how close or distant a pair of points is, if look at them from many different viewpoints. The two objects to be measured must be in the same cluster, while the points from where to establish this measurement must be outside of the cluster. This proposal is called Multiviewpoint-based Similarity, or MVS. The similarity of two documents d_i and d_j, given that they are in the same cluster is defined as the average of similarities measured relatively from the views of all other documents outside that cluster. From this point onwards the proposed similarity measure between two document vectors d_i and d_j will be denoted as $MVS(d_i, d_j / d_i, d_j \in S_r)$ or occasionally $MVS(d_i, d_j)$ for short.

The similarity between two points d_i and d_j inside cluster S_r, viewed from a point d_h outside this cluster, is equal to the product of the cosine of the angle between d_i and d_j looking from d_h and the euclidean distances from d_h to these two points. This definition is based on the assumption that d_h is not in the same cluster with d_i and d_j.

$$MVS(d_i, d_j | d_i, d_j \in S_r) = \sum_{d_h} \cos(d_i - d_h, d_j - d_h) \| d_i - d_h \| \| d_j - d_h \| \quad \text{Eqn (3.1)}$$

The overall similarity between d_i and d_j is determined by taking average over all the viewpoints not belonging to cluster S_r.

$$MVS(d_i, d_j d_i, d_j \in S_r) = \frac{1}{n - n_r} \sum_{d_h} \cos(d_i - d_h, d_j - d_h) \| d_i - d_h \| \| d_j - d_h \| \quad \text{Eqn (3.2)}$$

It can be seen that multiviewpoint offers more informative assessment of similarity than the single origin point-based similarity measure. With the multiviewpoint similarity measure, we implement hierarchical clustering which forms the document groups. For MVS, use cosine similarity for find out the dissimilar document object in the cluster. For this analyze the content of the each document and compare with the other document content. Similarity measures will depend on the text mining. Remove the dissimilar document from the cluster group and declare that the document as an outlier for the cluster group. To remove the document from the cluster group get the details of the dissimilar object likely name, location, current cluster id, etc. Compute the correlation similarity for each document with this outlier document. Related cluster group of the object will be predicted by using the incremental mining algorithm.

In the incremental optimization algorithm, we have two major steps Initialization and Refinement. At Initialization, k arbitrary documents are chosen to be the seeds from which primary partitions are formed. Refinement is a process that consists of a number of iterations. In each iteration, the n documents are randomly visited one by one. Each document is verified, if its move to another cluster results in progress of the objective function. Then the document is moved to the cluster that leads to the highest improvement. If no cluster is better than the current cluster, the document is not moved. The clustering process terminates when iteration completes without any documents being moved to new clusters. The incremental clustering algorithm updates instantly whenever each document is moved to new cluster. Since every move results increases the objective function value, convergence to a local optimum is guaranteed.

From the clustered objects, the document retrieval can be done based on the query. The query is preprocessed then it is matched with the documents in the clusters. Ranking is provided for the clusters with respect to the query matching result. The most relevant cluster to the query will be retrieved with this approach. From this, more informative assessment of similarity could be achieved between the documents.

Merits of Hierarchical multiviewpoint similarity include:

- Cosine Similarity measures will provide the dissimilar document object.

- Cluster overlapping phenomenon used to design cluster merging.
- Multiviewpoint is used to select the most relevant cluster of the other clusters.
- Reduce the irrelevant document in the cluster.
- Provide more accuracy of the result.
- High clustering accuracy.
- Clustering quality is increased.
- Performs both clustering and document retrieval.
- Retrieval accuracy is high.

IV. RESULTS AND DISCUSSION

In this section, the experiments and the performance results of the proposed method is described. To implement a novel approach for document retrieval using Hierarchical Agglomerative Clustering based on multi-view point similarity measure. The documents are collected and perform the preprocessing step for stemming and stop word removing. With the multiviewpoint similarity measure, we implement hierarchical clustering which forms the document groups. For MVS, use cosine similarity for find out the dissimilar document object in the cluster. For this analyze the content of the each document and compare with the other document content. Similarity measures will depend on the text mining. Remove the dissimilar document from the cluster group and declare that the document as an outlier for the cluster group. To remove the document from the cluster group get the details of the dissimilar object likely name, location, current cluster id, etc. Compute the correlation similarity for each document with this outlier document. Related Cluster group of the object will be predicted by using the incremental mining algorithm. From the clustered objects, the document retrieval can be done based on the query. The query is preprocessed then it is matched with the documents in the clusters. Ranking is provided for the Clusters with respect to the query matching result. The most relevant Cluster to the query will be retrieved with this approach.

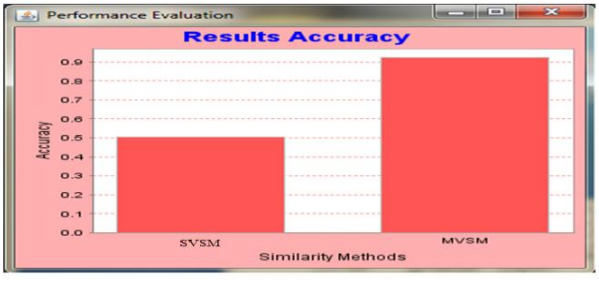

Fig 4.1: Results accuracy

The figure shows the accuracy comparison between the single viewpoint similarity measure (SVSM) and the multiviewpoint similarity measure (MVSM). From the figure, the SVSM shows an accuracy of just 50% where the MVSM shows an accuracy of about 95% which is far ahead from SVSM.

V. CONCLUSION

The paper proposes a novel approach for document retrieval using hierarchical agglomerative clustering based on multi-view point similarity measure rather than single viewpoint. Thus have a more accurate assessment of how close or distant a pair of points is. It is potentially more suitable for text documents. The documents are collected and perform the preprocessing step for stemming and stop word removing. The preprocessing step reduces the irrelevant document in the cluster and provides more accuracy of the result. From the clustered objects, the document retrieval can be done based on the query. The query is preprocessed then it is matched with the documents in the clusters. Ranking is provided for the Clusters with respect to the query matching result. The most relevant Cluster to the query will be retrieved with this approach.

The key contribution of this paper is the fundamental concept of hierarchical clustering from multiple view points. Future based on the same concept using different alternative measures and use other methods to combine the relative similarities according to the different viewpoints or do not use average but have other methods to combine the relative similarities according to the different viewpoints.

VI. REFERENCES

[1] Osmar R Zaiane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang: On Data Clustering Analysis: Scalability, Constraints and Validation.
[2] A K Jain, M N Murty, P J Flynn: Data Clustering: A Review, ACM Computing Surveys, Vol.31, No. 3, September 1999.
[3] I K Ravichandra Rao: Data Mining and Clustering Techniques, DRTC Workshop on Semantic Web 8th – 10th December, 2003, DRTC, Bangalore.
[4] S Susmitha, A Isabella: Hierarchical Viewpoint based clustering, International Journal of Computer Science and Technology, IJCST Vol. 4, Issue 1, Jan - March 2013.
[5] N Sandhya, Y SriLalitha, Dr. A Govardhan, Dr. K Anuradha: Analysis of Similarity Measures for Text Clustering.

[6] Lior Rokac, Oded Maimon: Clustering methods from Datamining and Knowledge Discovery Handbook, pp 325.
[7] Anna Huang: Similarity Measures for Text Document Clustering, Proceedings of Conference on NewZealand Computer Science Research Student Conference, NZCSRSC-2008, Christchurch, New Zealand, April 2008.
[8] Duc Thang Nguyen, Lihui Chen and Chee Keong Chan: Clustering with Multiviewpoint-Based Similarity Measure, IEEE transactions on Knowledge and Data Engineering, Vol. 24, No.6, June 2012, pp 988-1001.
[9] http://en.wikipedia.org/wiki/stopword removal.
[10] Porter M F: "An algorithm for suffix stripping" Program, Vol.14, pp. 130-137. 1980.