Review of Document Clustering Methods and Similarity Measurement Methods

Ms. P. S. Chaudhari
SAE,Kondhwa
Pune,India
e-mail:priyankadhirajbendale@gmail.com

Prof.L. J. Sankpal
SAE,Kondhwa
Pune,India
e-mail:ljsankpal.sae@sinhgad.edu

Abstract— The process of document clustering is nothing but the data mining method used for grouping of same items together. The clustering approach is aiming to identify the required structures from the input data as well as arranging them into the subgroup. There are many clustering algorithms presented by various authors already such as k-means clustering, Agglomerative Clustering, EM methods, direct clustering methods etc. Each clustering method is having its own advantages and disadvantages. The similarity measure used to measure the similarity between two clusters of input dataset derived from different or similar algorithms. In this paper we are first presenting the review over the document clustering and its different methods, and then later we taking the review of similarity measure method. The techniques of similarity measurements discussed in this paper are used for single viewpoint only. Finally, the limitations of this method are introduced.

Keywords: clustering, documents, data mining, similarity measurement, k-means, single viewpoint, multi viewpoint

*****

I. Introduction

Since from last decade, the use of e-documents or digital documents over the Internet is tremendously increasing through various activities from end users like search, retrieval, information management, social networking use etc. Thus designing of techniques in order to arrange large volume of unstructured data into the subgroups of meaningful information it important for such approaches indexing, automated metadata generation, filtering, word sense disambiguation, any application requiring document organization, and population of hierarchical catalogues of web resources [1], [2]. Therefore, the document clustering is most widely used approach for real time applications. We have also studied many pattern-mining methods in [3] those are frequently used for text mining.

The mechanism of clustering is method of data mining which makes the group of similar objects from the input dataset. Now days the clustering is most frequently used in data mining due its wide range of application usage. The main aim of clustering is find out structures in the dataset and classify those structures into meaningful subgroups for further analysis. There are many clustering methods presented by various researchers, and still more research is going on for their improvements. The clustering algorithms which are having better performance and efficiently work in most of applications is preferred most as compared to other clustering algorithms. Some algorithms givens better performances but resulted into high complexity. K-means clustering algorithm is best among all other existing algorithms due to its efficiency, easy to use, and faster in case of larger systems [4].

The clustering problem is addressed by using the method of optimization. Clustering is treated as optimization process. By using this optimization process, the optimal partition is extracted by optimizing the similarity measure among the data items [5]. In general, the similarity measurement metrics are defined those are used to measure the distance. The true intrinsic structure of data is defined properly by using such similarity measures, which in turns included into the clustering criterion function. Therefore the efficiency, effectiveness, processing speed of any clustering method is depends on use of similarity measurement metrics [6]. This performance dependency of clustering algorithms motivates the research on similarity measurement metrics.

This review paper is motivated by two factors, one is concepts of document clustering and its different methods, and similarity measurement used to evaluate the performances of clustering algorithms. In below sections, first in section II, we are presenting review of document clustering and different clustering algorithm. Then in next section III, we are presenting the overview of different similarity measurement techniques based on single viewpoint. In section IV, we point out limitations of these existing methods.

II. REVIEW DOCUMENT CLUSTERING AND ITS METHODS

A. Introduction:

A cluster is a collection of objects which are 'similar' between them and are 'dissimilar' to the objects belonging to other clusters [4]; and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set. There are several categories of clustering algorithms. In this paper, we will be focusing on algorithms that are exclusive in that the clusters may not overlap. Some of the algorithms are hierarchical and probabilistic. A hierarchical algorithm-clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations, it reaches the final clusters wanted.

M or Expectation Maximization is an example of this type of clustering algorithm. In [5], Pen et al. utilized cluster analysis composed of two methods. In Method I, a majority voting committee with three results generates the final analysis result. The performance measure of the classification is decided by majority vote of the committee. If more than two of the committee members give the same classification result, then the clustering analysis for that observation is successful; otherwise, the analysis fails.

Kalton et al. [6] did clustering and after letting the algorithm create its own clusters, added a step. After the clustering was completed, each member of a class was assigned the value of the cluster's majority population. The authors noted that the approach loses detail, but allowed them to evaluate each clustering algorithm against the "correct" clusters. In below section we are discussing the different methods of document clustering.

B. Document Clustering Methods

In this subsection, we review some of the clustering techniques related to this study.

1) Partitioned Clustering

Partitioned clustering approaches disjoint hierarchical document or non-cluster is a collection of documents in a predefined number of clustering algorithms partition attempted a flat Partitioned are divided or reallocation of run and pass a single methods. Most of them are single-pass way of walking is usually a reallocation method at the beginning of the data produced in the first division.

Partitioned clustering algorithms compute a k-way clustering of a set of documents either directly or via a sequence of repeated bisections. A direct clustering method generally follows k-calculated. Initially k cluster a set of documents to act as seed collection is selected by then each document for its resemblance to the seeds of k is computed. Then this is its most similar to seed clusters assigned to the initial k-way clustering is then repeatedly forms the refined so that it desired clustering Optimizes the test function a k-way partitioning via recursively repeated bisections 2-way clustering (i.e., Bisections) to calculate the above algorithm is obtained by applying. Initially, the documents are divided into two groups, and then one of these groups has been selected and is bisected, and so on. K-1 this process continues the times leading to these cluster k bisections has performed so that each of the resulting clustering solutions special two-way optimizes the test function.

M. Steinbach, G. Karypis, and V. Kumar [9] the overall k-way clustering solution will not necessarily be at local minima with respect to the criterion function. The key step in this algorithm is the method used to select which cluster to bisect next. This approach led to reasonably good and balanced clustering solutions [9] as all of our experiments, we select the largest cluster, and comprehensive experiments show that presented in [10], and clustering solutions gained through repeatedly bisections comparable or better than those via direct clustering produced did. They each step to resolve a simple optimization problem, as also the computational requirements are very small. For this reason, in all of our experiments we use partitioned clustering approach to compute the solutions.

Y. Zhao, G. Karypis [10] one of the differences between partitioned and agglomerative clustering algorithms are the fact that the former do not generate an agglomerative tree. Agglomerative trees are very useful as they contain information on how the different documents are related to each other, at different levels of granularity. One-way of inducing an agglomerative tree from a partitioned clustering solution is to do it in such a way so that it preserves the already computed k-way clustering. First, of all each, one of us groups documents relating to an agglomerative tree and then we are a discovered partition ally agglomerative tree. whose leaves cluster by building a coalition of these trees, this approach ensures that the overall clustering solution tree induced by k way clustering solution calculated by partitioned algorithm similar to both of these trees are constructed so that those same criteria Partitioned clustering solution function was used to optimize..

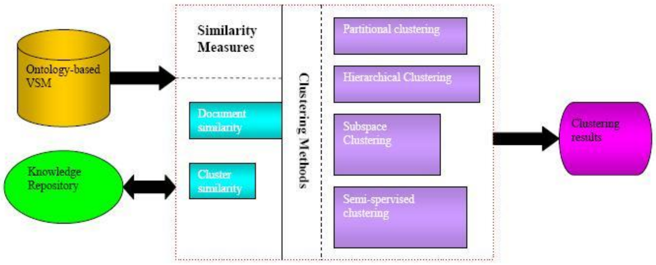

Figure 1: Clustering Methods

2) Hierarchical Clustering Algorithms

Hierarchical clustering approaches attempt to create a hierarchical decomposition of the given document collection thus achieving a hierarchical structure. Hierarchical methods are usually classified into Agglomerative and Divisive methods depending on how the hierarchy is constructed.

Agglomerative methods start with an initial clustering of the term space, where all documents are considered to represent a separate cluster. The closest clusters using a given inter-cluster similarity measure are then merged continuously until only one cluster or a predefined number of clusters remain.

$$\vec{\mu}(C) = \frac{1}{|C|}\sum_{\vec{x}\in C}\vec{x}$$

Simple Agglomerative Clustering Algorithms are

1. Compute the similarity between all pairs of clusters i.e. calculates a similarity matrix whose it entry gives the similarity between the ith and jet clusters.

2. Merge the most similar (closest) two clusters.

3. Update the similarity matrix to reflect the pair wise similarity between the new cluster and
The original clusters.

4. Repeat steps two and three until only a single cluster remains. Divisive clustering algorithms start with a single cluster containing all documents. It then continuously divides clusters until all documents are contained in their own cluster or a redefined number of clusters are found.

Agglomerative algorithms are usually classified according to the inter-cluster similarity measure they use. The most popular of these are single-link, complete-link and group average. In the single link method, the distance between clusters is the minimum distance between any pair of elements drawn from these clusters (one from each), in the complete link, it is the maximum distance and in the average link it is correspondingly an average distance.

a) Advantages of hierarchical clustering

1. Embedded flexibility regarding the level of granularity.

2. Ease of handling any forms of similarity or distance.
Applicability to any attributes type.

b) Disadvantages of hierarchical clustering Algorithm is for K Means vagueness of termination criteria.

Most hierarchal algorithm does not revisit once constructed clusters with the purpose of improvement.

C. K-Means

K-means clustering algorithm is the most important flat. Objective function means your K cluster centers, a cluster Center is defined as mean or centroid µ K-means cluster in the c: a model of the center of gravity in the cluster centroid as an area where the average squared distance from objects is to reduce ideally, the cluster must not overlap. Residual amount of squares (RSS), summed over all vectors each of its centroid vector squared distance how well members of your groups represent a measure of centroids

$$\mathrm{RSS_i} = \sum_{\vec{x}\in C_i} \| \vec{x} - \vec{\mu}(C_i) \|^2$$

$$\mathrm{RSS} = \sum_{i=1}^{K} RSS_i$$

K-means can start by selecting as initial cluster centers K randomly chosen objects, namely the seeds. It then moves the cluster centers around in space in order to minimize RSS.

1. Reassigning objects in the cluster with the closest centroid.
2. Recomposing each centroid based on the current members of its cluster.

We can use one of the following terms conditions as stopping criterion

1. A fixed number of iterations have been completed.
2. Centroids µi do not change between iterations.
3. Terminate when RSS falls below a pre-established threshold.

```
procedure KMEANS(X,K)
    {s1, s2, · · · , sk}   SelectRandomSeeds(K,X)
    for i ←1,K do
        µ(Ci) ← si
    end for
  repeat
      $\min_{k\sim x_n--\sim\mu(C_k)k} C_k = C_k [ \{\sim x_n\}$
      for all $C_k$ do
          $\mu(C_k) = 1$
      end for
  until stopping criterion is met
end procedure
```

D. Expectation Maximization

The EM algorithm fall within a subcategory of the flat clustering algorithms, called Model-based clustering. The model-based clustering assumes that data were generated by a model and then tries to recover the original model from the data. This model then defines clusters and the cluster membership of data.

III. REVIEW OF SIMILARITY MEASUREMENT METHODS

In this section, we are discussing the different methods of similarity measurement used on single viewpoint.

A. Euclidean Distance

It is common distance between two points and can be without difficulty measured with a ruler in two or three-dimensional space. Euclidean distance one of the important popular measures:

Dist (di, dj) = ||di − dj ||

$$min \sum_{r=1}^{K} \sum_{di\in Sr} ||di - cr||^2$$

Particularly, similarity of two documents vector di and dj, Sim (di, dj), is defined as the cosine of angle between them. For unit vectors, this equals to their inner product:

$$\mathrm{sim(di,dj) = cos(di,dj) = di^t\ dj}$$

B. Cosine Similarity

When documents are represented in terms of vectors, vector commonality of the correlation between two documents corresponds a sparse and high dimensional space, the cosine similarity is widely used is also a popular similarity score text mining and information retrieval.

$$\max \sum_{r=1}^{k} \sum_{di\in Sr} \frac{di^t\ Cr}{||Cr||}$$

C. Jaccard Coefficient

It sometimes referred to as the divided by the union of the objects. For text documents, the Jaccard coefficient compares the sum weight of shared terms to the sum weight of terms that are present in either of the two documents but are not the shared terms.

$$\mathrm{Sim_{eJacc}[Ui, Uj]} = \frac{\mathrm{Ui^t\ Uj}}{\mathrm{||Ui||^2 + ||Uj||^2 - Ui^t\ Uj}}$$

D. Pearson Correlation Measure

It provides a method for clustering a set of objects into the set of objects into the best possible number of clusters, without specifying that number in proceed.

$$S(x_i, x_j) = \frac{(X_i - \overline{x_I})^T\ (x_j - \overline{X_I})}{\|X_i - \overline{x_I}\| \|x_j - X_I\|}$$

Where $\overline{x_I}$ denotes the average feature value of x over all dimensions.

IV. LIMITATIONS

From this study we identified below are listed limitations of existing clustering algorithms as well as similarity measurement methods.

- The clustering algorithms are working efficient on single viewpoint based documents, however those failed to work with multi viewpoint based documents.
- As the clustering algorithms fails, the similarity measurement metrics fails to predict the distances accurately.

V. CONCLUSION AND FUTURE WORK

In this review paper, we have presented the survey on document clustering technique and different algorithms presented by various researchers. The similarity measure is the metrics, which is used to measure the distance between the two clustering data items. From this survey study, we conclude that similarity measures playing important role in defining the success or failure of any particular clustering approach. Thus for the evaluation of any clustering method we need to have efficient similarity measurement. The methods, which we have studied in this paper, are only efficient in case of single viewpoint based clustering methods, but now days most of datasets are sparse as well as high dimensional forms. On such datasets, existing similarity measures failed. Thus, the future work is based on designing of efficient Multi viewpoint-Based Similarity Measure method.

REFERENCES

[1] K.Sathiyakumari, G.Manimekalai, V.Preamsudha, “A Survey on Various Approaches in Document Clustering”, G Manimekalai et al, Int. J. Comp. Tech. Appl., Vol 2 (5), 1534-1539.

[2] Ashish Moon, T. Raju, A Survey on Document Clustering with Similarity Measures, International Journal of Advanced Research in Computer Science and Software Engineering 3(11), November - 2013, pp. 599-601.

[3] Mr. Uday P. Kulkarni, Mr. Hemant B. Mahajan, Pattern Discovery for Text Mining Procedure Using Taxonomy, International Journal of Engineering &Extended Technologies Research (IJEETR), Vol. 2, Issue 9, September – 2013.

[4] K. M. Hammouda and M. S. Kamel. Efficient phrase based document indexing for web document clustering. IEEE Transactions on knowledge and data engineering, 16(10):1279{1296, 2004.

[5] George A. Miller. Wordnet: a lexical database for English. Common ACM, 38(11):39{41, 1995.

[6] Alessandro Moschitti and Roberto Basili. Complex linguistic features for text classification: A comprehensive study. In ECIR '04: 27th European conference on IR research, pages 181{196, Sunderland, UK, April 2004.

[7] X. Wu, V. Kumar, J.R. Quinlan, J. Ghosh, Q. Yang, H. Motoda, G.J. McLachlan, A. Ng, B. Liu, P.S. Yu, Z.-H. Zhou, M. Steinbach, D.J. Hand, and D. Steinberg, “Top 10 Algorithms in Data Mining,” Knowledge Information Systems, vol. 14, no. 1, pp. 1-37, 2007.

[8] Guyon, U.V. Luxburg, and R.C. Williamson, “Clustering: Science or Art?” Proc. NIPS Workshop Clustering Theory, 2009.

[9] M. Steinbach, G. Karypis, and V. Kumar. A comparison of document clustering techniques. In KDD Workshop on Text Mining, 2000.

[10] Y. Zhao, G. Karypis. Criterion Functions for Document Clustering: Experiments and Analysis. Technical Report #01-34, University of Minnesota, MN 2001.