International Journal of Computer Applications Technology and Research
Volume 3– Issue 1, 5 - 9, 2013, ISSN: 2319–8656

Clustering that depends upon Various Viewpoints Similarity Measure

V. Venkaiah,
Department of Computer Science and Engineering,
TKR College of Engineering and Technology
Hyderabad, A.P-500 097, India

A. Pramod Reddy
Department of Computer Science and Engineering,
TKR College of Engineering and Technology
Hyderabad, A.P-500 097, India

P.V.S. Srinivas
Department of Computer Science and Engineering
TKR College of Engineering and Technology
Hyderabad, A.P-500 097, India

Abstract: Between the data objects there must be cluster relationship which was assumed by clustering methods which they are applied on. Explicitly or implicitly will be defined based upon common things from a pair of objects. In this paper, we explored a novel based upon similarity measure which is multiviewpoint and also couple of clustering methods which are related. There is difference and also it is a major drawback in traditional dissimilarity/similarity measure to ours. The previous method only utilizes a single viewpoint, and then it is treated as origin, while our proposed system deals with a lot various viewpoints. Same cluster objects will not be assumed for measuring. Assessment of similarity will be gained by utilizing multiple viewpoints. Our system is supported with the help of conducting theoretical analysis and empirical study. For document clustering based on this new measure two criterion functions are explored. We measures the similarities and dissimilarities of them with a lot various clustering algorithms that utilizes another well known similarity measures on different document collections to check our proposal advantages.

Keywords: Clustering, Multiviewpoint, Similarity measure, Singleviewpoint, Two criterion functions.

1. INTRODUCTION

In data mining clustering takes key role when compared to other important ones. The main goal of clustering is to get intrinsic structures in information, and arrange them into perfect subgroups for future work and analysis. There are a lot clustering algorithms comes in every academic year. Those have been proved in various research areas, and created utilizing completely various techniques and approaches.

According to present researches, even after 5 decades after inventing, the simple algorithm k-means has its goal and still in top 10 algorithms list of data mining. Partitional clustering algorithm is the frequently utilized in reality. Not only that there is another one that recent survey states that k-means is the mostly utilized algorithm that users used in their practicals in the similar fields selected to utilize. There is no need to say, k-means has a little more drawbacks than the basic drawbacks, like cluster size sensitiveness initialization and its working capability is the most worst when in lot domains than rest of different state-of-the-art algorithms.

It is based on similarity measure from multi view points. Can process large amounts of data. It improves the clustering performances. Can be applied documents clustering. Can be applied on web document clustering. Can be applied on social web sites clustering. An optimization process is most commonly used approach for the clustering crisis. Optimizing invented optimal partition with the help of particular similarity function among data.

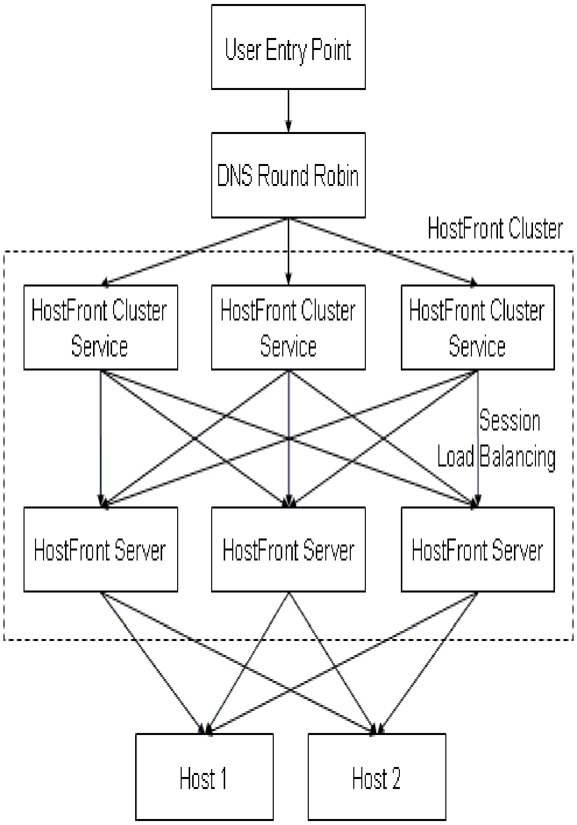


Fig: Clustering with multiple view point

Basically, there is a chance intrinsic structure of information could be accurately declared with the help of similarity formula embedded and declared in the function of clustering criterion. Based upon similarity measure on data or information the clustering algorithms effectiveness will be calculated. K-means contains objective function sum-of-squared-error that utilizes Euclidean distance. In the case of high-dimensional domains which are spherical k-means, text documents which utilizes cosine similarity (CS) as measure rather than euclidean distance cause its more efficient.

2. PROBLEM STATEMENT

Existing system:

The previous system deals with concepts like Phrase-Based clustering, Concept-Based clustering, Conceptual Tree similarity clustering. These are containing such drawbacks that make system failure.

There are some issues pertinent to most of the existing feature clustering methods.

- First, the parameter k, indicating the desired number of extracted features, has to be specified in advance. This gives a burden to the user, since trial-and-error has to be done until the appropriate number of extracted features is found.
- Second, when calculating similarities, the variance of the underlying cluster is not considered. Intuitively, the distribution of the data in a cluster is an important factor in the calculation of similarity.
- Third, all words in a cluster have the same degree of contribution to the resulting extracted feature. Sometimes, it may be better if more similar words are allowed to have bigger degrees of contribution. Our feature clustering algorithm is proposed to deal with these issues.

Proposed System:

Because of the nature of similarity plays an important role in clustering, there is a need of a method for calculating common data between objects. Proposed a novel clustering similarity measure based on multi view point.

3. SYSTEM DEVELOPMENT

Multiviewpoint-based similarity

The cosine similarity will be explored in the following form without changing its meaning.

$$Sim(d_i, d_j) = \cos(d_i - 0, d_j - 0) = (d_i - 0)^t (d_j - 0),$$

Where vector 0 indicates the point of origin. Based upon this formula, the calculation picks 0 as one and also point of reference. The common thing of couple of documents di and dj is defined with respect to the angle between the couple of points when the view is from origin.

Two Clustering Criterion Functions IR and IV

We now created our clustering criterion functions. The initial function, known as IR, is the cluster size-weighted those are sum of average of normal pair wise common things of documents which are in the one cluster only. Initially, we explored the sum by function F in a normal form.

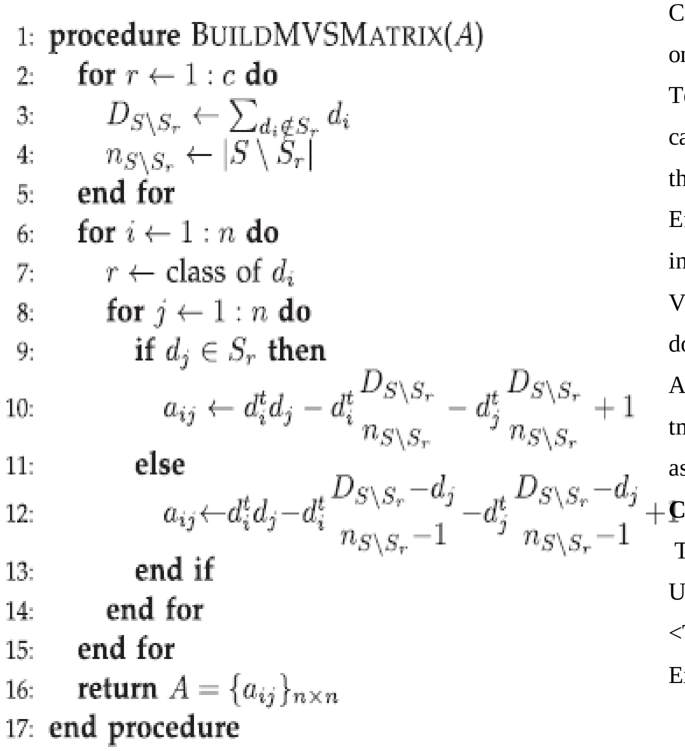
$$F=\sum_{r=1}^{k} n_r \left[\frac{1}{n_r^2} \sum_{d_i,d_j \in S_r} Sim(d_i, d_j) \right].$$

Optimization Algorithm and Complexity

We have given clustering framework of ours by Multiviewpoint-based Similarity Clustering. At the same time, we are having MVSC-IV and MVSC-IR, those are with MVSC IR and IV criterion function, respectively. The key aim is to do clustering by optimizing on document IV and IR. Cause of this sake, the algorithm incremental k-way a sequential flow of k-means is used. Taking that the IV expression depends upon only on nr and Dr, r = 1; . . . ; k, IV will be in a general form.

$$I_V=\sum_{r=1}^{k} I_r(n_r, D_r),$$

Where Ir(nr,Dr) is the objective value of r cluster. The similar is used on IR. The incremental optimization algorithm in general form, which contains couple of important steps those are refinement and initialization, are explored.



```
procedure BUILDMVSMATRIX($A$)
    for $r \leftarrow 1 : c$ do
        $D_{S\setminus S_r} \leftarrow \sum_{d_i \notin S_r} d_i$
        $n_{S\setminus S_r} \leftarrow |S \setminus S_r|$
    end for
    for $i \leftarrow 1 : n$ do
        $r \leftarrow$ class of $d_i$
        for $j \leftarrow 1 : n$ do
            if $d_j \in S_r$ then
                $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S\setminus S_r}}{n_{S\setminus S_r}} - d_j^t \frac{D_{S\setminus S_r}}{n_{S\setminus S_r}} + 1$
            else
                $a_{ij} \leftarrow d_i^t d_j - d_i^t \frac{D_{S\setminus S_r} - d_j}{n_{S\setminus S_r} - 1} - d_j^t \frac{D_{S\setminus S_r} - d_j}{n_{S\setminus S_r} - 1} + 1$
            end if
        end for
    end for
    return $A = \{a_{ij}\}_{n\times n}$
end procedure
```

Algorithm: MVS Similarity matrix construction

4. RELATED WORK

Hard clustering: Each document belongs to exactly one cluster. More common and easier to do Soft clustering: A document can belong to more than one cluster. Makes more sense for applications like creating browsable hierarchies.

Cosine similarity measurement

Assign Boolean values to a vector describing the attributes of a database element, then measure vector similarities with the Cosine Similarity Metric. Cosine similarity is a measure of similarity between two vectors by measuring the cosine of the angle between them. The result of the Cosine function is equal to 1 when the angle is 0, and it is less than 1 when the angle is of any other value. As the angle between the vectors shortens the cosine angle approaches 1, meaning that the two vectors are getting closer, meaning that the similarity of whatever is represented by the vectors increases. Assign Numeric values to non-numerical items, and then use one of the standard clustering algorithms. Then use one of the standard clustering algorithms like, hierarchical clustering, agglomerative ("bottom-up") or, divisive ("top-down"), Partitional clustering, Exclusive Clustering, Overlapping Clustering, Probabilistic Clustering , Text Clustering, Text clustering is one of the fundamental functions in text mining.

Text clustering is to divide a collection of text documents into different category groups so that documents in the same category group describe the same topic, such as classic music or history or romantic story. Efficiently and automatically grouping documents with similar content into the same cluster.

Vector space model is an algebraic model for representing text documents (and any objects, in general) as vectors of identifiers.

A text document is represented as a vector of terms <t1, t2, …, ti, …, tm>.Each term ti represents a word. A set of documents are represented as a set of vectors, that can be written as a matrix.

Collection Reader

Transform raw document collection into a common format, e.g., XML Use tags to mark off sections of each document, such as, <TOPIC>, <TITLE>, <ABSTRACT>,<BODY> Extract useful sections easily Example:

“Instead of direct prediction of a continuous output variable, the method discretizes the variable by kMeans clustering and solves the resultant classification problem.”

Detagger

Find the special tags in document “,”,”.”. Filter away tags. “Instead of direct prediction of a continuous output variable the method discretizes the variable by kMeans clustering and solves the resultant classification problem”.

Removing Stopwords

Stopwords

Function words and connectives. Appear in a large number of documents and have little use in describing the characteristics of documents.

Example

Removing Stopwords

Stopwords: “of”, “a”, “by”, “and” , “the”, “instead”

Example: “direct prediction continuous output variable method discretizes variable kMeans clustering solves resultant classification problem”

- **Stemming**

Remove inflections that convey parts of speech, tense.

Techniques: Morphological analysis (e.g., Porter’s algorithm), Dictionary lookup (e.g., WordNet), Stems: “prediction --->predict”, “discretizes --->discretize”, “kMeans ---> kMean”, “clustering --> cluster”, “solves ---> solve”, “classification ---> classify”.

Example sentence: “direct predict continuous output variable method discretize variable kMean cluster solve resultant classify problem”

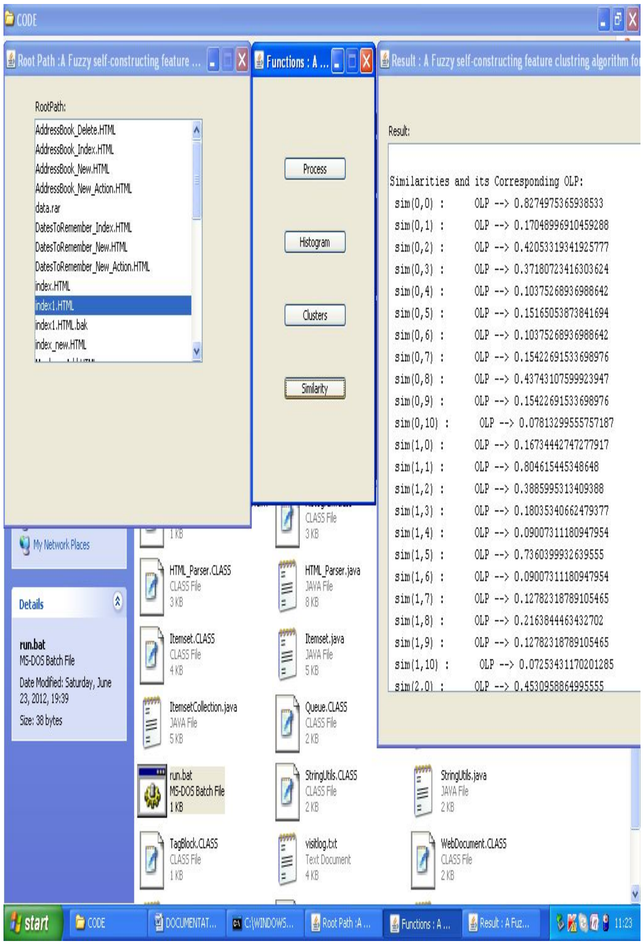
K-Means

In some case documents are taken as real-valued vectors. Select K random docs {s1, s2,… sK} as seeds. Until clustering converges or other stopping criterion: For each doc di: Assign di to the cluster cj such that dist (xi, sj) is minimal. (Update the seeds to the centroid of each cluster) For each cluster cj sj = μ(cj). A fixed number of iterations. Doc partition unchanged. Centroid positions don’t change. A state in which clusters don’t change. K-means is a special case of a general procedure known as the Expectation Maximization (EM) algorithm. EM is known to converge. Number of iterations could be large.

Result: Whole corpus analysis/navigation. Better user interface: search without typing. For improving recall in search applications. Better search results (like pseudo RF). For better navigation of search results. Effective “user recall” will be higher. For speeding up vector space retrieval. Cluster-based retrieval gives faster search. Cluster hypothesis - Documents in the same cluster behave similarly with respect to relevance to information needs. Therefore, to improve search recall: Cluster docs in corpus a priori. When a query matches a doc D, also return other docs in the cluster containing D. Hope if we do this: The query “car” will also return docs containing automobile. Because clustering grouped together docs containing car with those containing automobile.

5. CONCLUSION

In this paper, based upon similarity measuring we explored a Multiviewpoint- method, known as MVS. MVS is potentially perfect based upon theoretical analysis and empirical examples. MVS is perfect than the cosine similarity for text documents. There are couple of clustering algorithms that impacts some criterion functions, those are IR and IV.



```
Similarities and its Corresponding OLP:
 sim(0,0) :      OLP --> 0.8274975365938533
 sim(0,1) :      OLP --> 0.17048996910459288
 sim(0,2) :      OLP --> 0.42053319341925777
 sim(0,3) :      OLP --> 0.37180723416303624
 sim(0,4) :      OLP --> 0.10375268936988642
 sim(0,5) :      OLP --> 0.15165053873841694
 sim(0,6) :      OLP --> 0.10375268936988642
 sim(0,7) :      OLP --> 0.15422691533698976
 sim(0,8) :      OLP --> 0.43743107599923947
 sim(0,9) :      OLP --> 0.15422691533698976
 sim(0,10) :      OLP --> 0.07813299555757187
 sim(1,0) :      OLP --> 0.16734442747277917
 sim(1,1) :      OLP --> 0.804615445348648
 sim(1,2) :      OLP --> 0.3885995313409388
 sim(1,3) :      OLP --> 0.18035340662479377
 sim(1,4) :      OLP --> 0.09007311180947954
 sim(1,5) :      OLP --> 0.7360399932639555
 sim(1,6) :      OLP --> 0.09007311180947954
 sim(1,7) :      OLP --> 0.12782318789105465
 sim(1,8) :      OLP --> 0.2163844463432702
 sim(1,9) :      OLP --> 0.12782318789105465
 sim(1,10) :      OLP --> 0.07253431170201285
 sim(2,0) :      OLP --> 0.4530958864995555
```

MVSC-IR and MVSC-IV, are explored newely. State-of-the-art clustering methods utilizes various ones of similarity measure, on a huge count sets of document data and under various evaluation metrics, the explored new algorithms proven that they are capable of providing better clustering operating capability significantly. The main thing of this is the basic concept of common measure from many viewpoints. Upcoming methods capable of utilizing of the one principle, but for the relative similarity mentioned alternative forms, or wont utilize average but contains different methods to club the relative common things based upon various viewpoints.

Apart from this paper keeping interest on documents partition clustering. In upcoming days, for hierarchical clustering algorithms it may be capable to use the proposed criterion functions. At last, we have explored the MVS application and its text data clustering algorithms. It is an interesting part to show the working on various types of sparse and also on dimensional data which was high.

6. REFERENCES

[1]http://people.csail.mit.edu/jrennie/20Newsgroups/, 2010.

[2]Http://kdd.ics.uci.edu/databases/reuters21578/reuters21578.html. 2010.

[3] H. Kim, P. Howland, and H. Park, “Dimension Reduction in Text Classification with Support Vector Machines,” J. Machine Learning Research, vol. 6, pp. 37-53, 2005.

[4] F. Sebastiani, “Machine Learning in Automated Text Categorization,” ACM Computing Surveys, vol. 34, no. 1, pp. 1-47, 2002.

[5] B.Y. Ricardo and R.N. Berthier, Modern Information Retrieval. Addison Wesley Longman, 1999.

[6] A.L. Blum and P. Langley, “Selection of Relevant Features and Examples in Machine Learning,” Artificial Intelligence, vol. 97, nos. 1/2, pp. 245-271, 1997.

[7] E.F. Combarro, E. Montan˜ e´s, I. Dı´az, J. Ranilla, and R. Mones, “Introducing a Family of Linear Measures for Feature Selection in Text Categorization,” IEEE Trans. Knowledge and Data Eng., vol. 17, no. 9, pp. 1223-1232, Sept. 2005.

[8] K. Daphne and M. Sahami, “Toward Optimal Feature Selection,” Proc. 13th Int’l Conf. Machine Learning, pp. 284-292, 1996.

[9] R.Kohavi and G.John, “Wrappers for Feature Subset Selection,” Artificial Intelligence, vol. 97, no. 1-2, pp. 273-324, 1997.

[10] Y. Yang and J.O. Pedersen, “A Comparative Study on Feature Selection in Text Categorization,” Proc. 14th Int’l Conf. Machine Learning, pp. 412-420, 1997.

[11] D.D. Lewis, “Feature Selection and Feature Extraction for Text Categorization,” Proc. Workshop Speech and Natural Language, pp. 212-217, 1992.

[12] H. Li, T. Jiang, and K. Zang, “Efficient and Robust Feature Extraction by Maximum Margin Criterion,” T. Sebastian, S. Lawrence, and S. Bernhard eds. Advances in Neural Information Processing System, pp. 97-104, Springer, 2004.