ISSN: 2321-7782 (Online)

Volume 2, Issue 1, January 2014

International Journal of Advance Research in Computer Science and Management Studies

Research Paper

Available online at: www.ijarcsms.com

Implementation of Multi View point method for similarity Measure in clustering the documents

Annavazula Mrinalini[1]
Computer Science and Engineering
SVU College of Engineering
Tirupathi, India

A.Rama Mohan Reddy[2]
Computer Science and Engineering
SVU College of Engineering
Tirupathi, India

Abstract: The clustering will have some clustering relationship between the documents or objects that we are applied on. In clustering the similarity is a measure to provide similarity between documents either explicitly or implicitly defined. In this paper we are using multi view point based similarity measure. In this method we will take more than one document as a reference between the documents. In traditional method we are using only one view point as a reference that is k-means algorithm for similarity between the documents. In the proposed method we used cosine with multi view point based similarity measure between the documents. The multi view will provide more information assessment than traditional method and reduce the irrelevant documentation. We are comparing k-means with multi view point similarity measure on different documents to verify the advantage of proposed method.

Keywords: Document clustering, Data Mining,Text mining, Tf -Idf, Similarity Measure, Cosine Similarity

I. INTRODUCTION

Data mining refers to extracting or "mining knowledge from large amounts of data. The term is actually a misnomer. Remember that the mining of gold from rocks or sand is referred to as gold mining rather than rock or sand mining .Thus, data mining should have been more appropriately named "knowledge mining from data," which is unfortunately somewhat long. "Knowledge mining", A shorter term may not reflect the emphasis on mining from large amounts of data. Nevertheless, mining is a vivid term characterizing the process that finds a small set of precious nuggets from a great deal of raw materials.

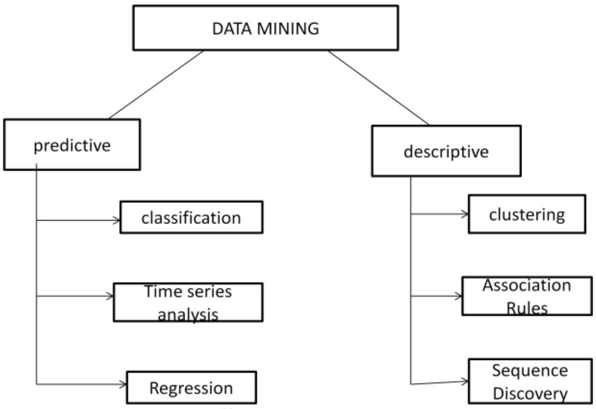


Fig.1 Different Data Mining Tasks

Fig.1 represents the different data mining tasks. This may be predictive or descriptive type. According to the work it's come under the descriptive type that is clustering.

Clustering is one of the most interesting and important topics in data mining. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group (called cluster) are more similar (in some sense or another) to each other than to those in other groups(clusters). It is a main task of explorative data mining. 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. Clustering can be considered the most important *unsupervised learning* problem; so, as every other problem of this kind, it deals with finding a *structure* in a collection of unlabeled data. A loose definition of clustering could be “the process of organizing objects into groups whose members are similar in some way”. A *cluster* is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

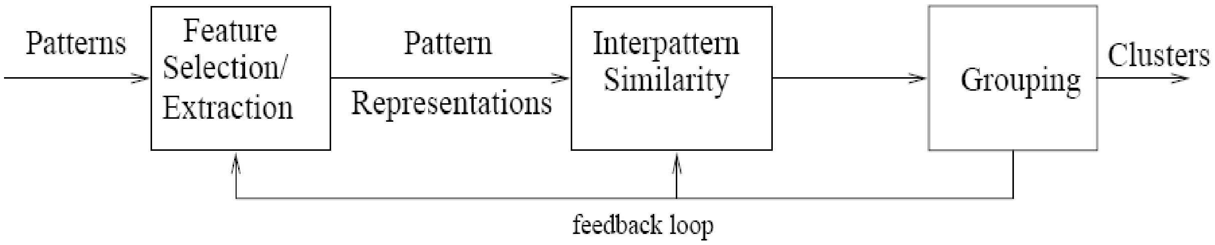


Fig.2 represents the process of clustering. In this paper we are collecting patterns as documents and giving input to the similarity measure and group the documents finally we will get output as clusters.

Cluster is therefore a collection of objects which are “similar” between them and are “dissimilar” to the objects belonging to other clusters.

Clustering has two criterions.

1. Distance based clustering
2. Conceptual clustering

Two or more objects belong to the same cluster if they are “close” according to a given distance. This is called distance-based clustering.

Another kind of clustering is conceptual clustering: two or more objects belong to the same cluster if this one defines a concept common to all that objects. In other words, objects are grouped according to their fit to descriptive concepts, not according to simple similarity measures.

II. Existing System

1) K-Means Algorithm

Existing System we are using K-means algorithm. k-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships.The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields. K-means clustering construct a partition of a database ***D*** of ***n*** objects into a set of ***k*** clusters. Given a *k*, find a partition of *k clusters* that optimizes the chosen partitioning criterion.

$$E=\sum_{i=1}^{k}\sum_{p\in C_i}\left\|p-m_i\right\|^2$$

K-means will use square and error method for calculating distance between the documents.

- Variations of K-Means
  - Initialisation (select the number of clusters, initial partitions)
  - Updating of centre

- Hill-climbing (trying to move an object to another cluster).

2) Steps in K-means Algorithm:

- Initially, the number of clusters must be known, or chosen, to be K say.
- The initial step is the choose a set of K instances as centres of the clusters. Often chosen such that the points are mutually "farthest apart", in some way.
- Next, the algorithm considers each instance and assigns it to the cluster which is closest.
- The cluster centroid are recalculated either after each instance assignment, or after the whole cycle of re-assignments.
- This process is iterated.

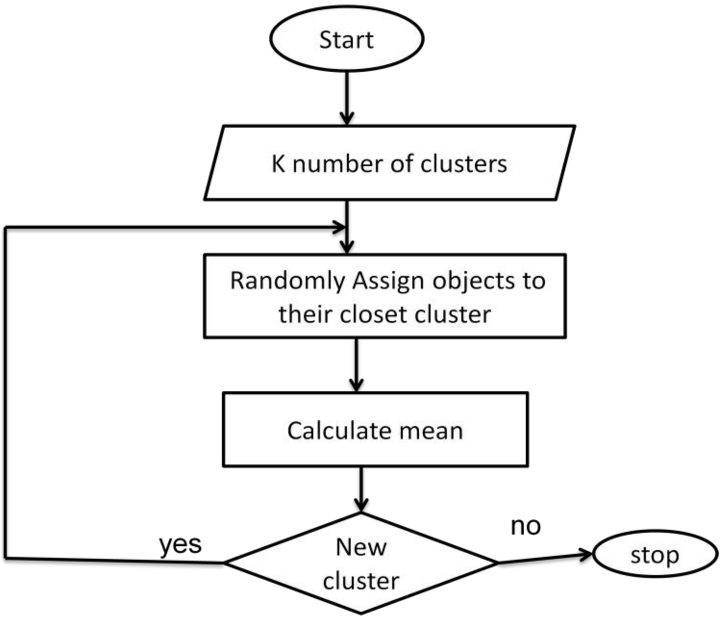


Fig.3 Flowchart of k-means Algorithm

Fig.3 Represents the Flowchart representation for k-means algorithm.

3) Limitations of k-means Algorithm:

- Document moved based on frequent Occurrence of next cluster.
- Cluster movement will quite complex when number of documents increased.
- Sometimes similarity process taking long period of time.
- It will not give a optimized partition process.

III. Proposed System

1) Cosine Similarity

The limitations of k-means algorithm we will go for cosine with multi view point based similarity measure. In The cosine similarity measure is also adapted as one of the variants of k-means known as "spherical k-means". In cosine similarity we are going to frame similarity vectors based on TF-IDF measure. Calculate cosine similarity measure by using following formula.

$$SIM(di,dj)= \frac{di.dj}{||di||\ ||dj||}$$

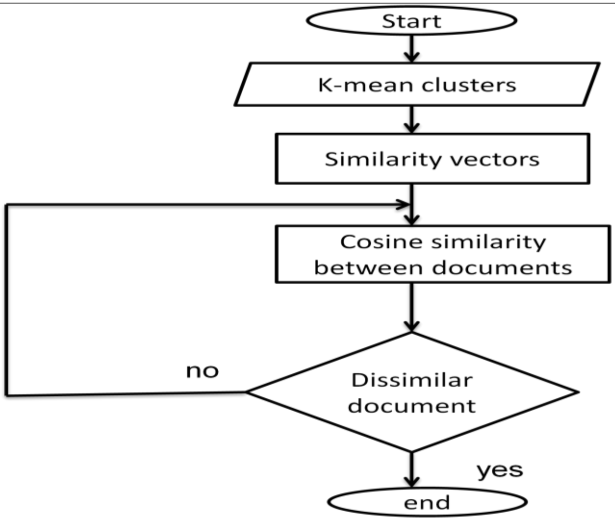


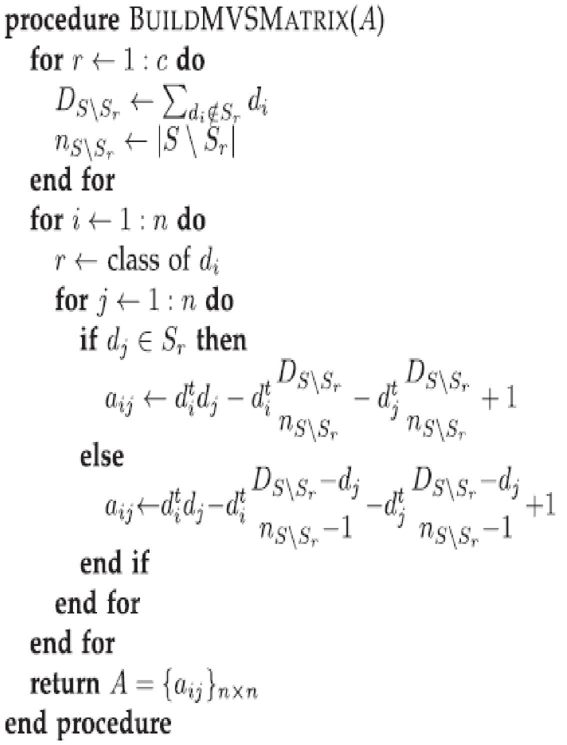
Fig.4 Flowchart of cosine similarity measure.

Fig.4 shows the step by procedure how cosine similarity will take place. The input for the cosine similarity is the clusters from the k-means algorithm give as an input and produce the documents which are not related to that cluster. So the output of the cosine similarity gives dissimilar documents.

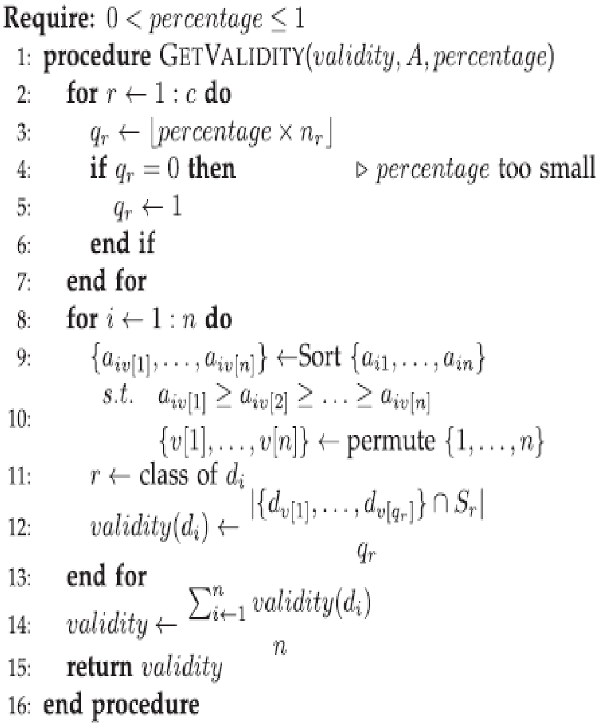
2) Multi View Point Method

The multi view point method will take output of the cosine similarity. The output of cosine similarity the dissimilar documents give input for the multi view point method. The output of the multi view point method will provide documents with their corresponding clusters. The multi view point method Reduce the irrelevant documentation.

3) Algorithm for Multi View Point Method



```
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
```



```
Require: $0 < percentage \leq 1$
procedure GetValidity($validity, A, percentage$)
  for $r \leftarrow 1 : c$ do
    $q_r \leftarrow \lfloor percentage \times n_r \rfloor$
    if $q_r = 0$ then                    ▷ $percentage$ too small
      $q_r \leftarrow 1$
    end if
  end for
  for $i \leftarrow 1 : n$ do
    $\{a_{iv[1]}, \ldots, a_{iv[n]}\} \leftarrow$ Sort $\{a_{i1}, \ldots, a_{in}\}$
      $s.t. \quad a_{iv[1]} \geq a_{iv[2]} \geq \ldots \geq a_{iv[n]}$
          $\{v[1], \ldots, v[n]\} \leftarrow$ permute $\{1, \ldots, n\}$
    $r \leftarrow$ class of $d_i$
    $validity(d_i) \leftarrow \frac{|\{d_{v[1]}, \ldots, d_{v[q_r]}\} \cap S_r|}{q_r}$
  end for
  $validity \leftarrow \frac{\sum_{i \leftarrow 1}^{n} validity(d_i)}{n}$
  return $validity$
end procedure
```

IV. Result

The multi view point method will provide more information assessment than the k-means algorithm and reduce the irrelevant documentation. Finally, the multi view provides more accurate result than k-means algorithm. (Fig.5)

$$Accuracy=\frac{1}{n}\ max\sum_{i=1}^{k} n_{i,\ q\ (i)}$$

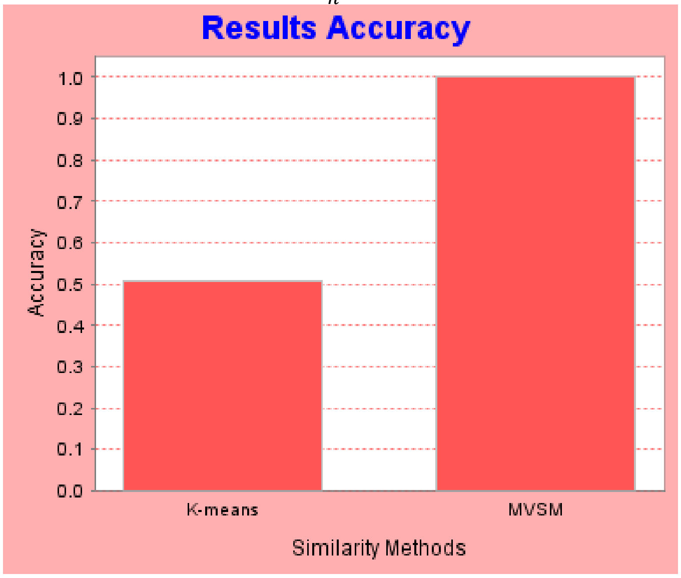


Fig.5 Accuracy Result

V. CONCLUSION

The multi view point similarity measure will provide good result in high dimensional domain. According to our work more number of documents comes under high dimensional domain. The irrelevant documentation is reduced here so that we can predict that the multi view will provide good result than the k-means algorithm. In future stemming can be used to reduce the load on each document.

ACKNOWLEDGEMENT

I wish to express my hearty gratitude and sincere regards to authors Duc Thang Nguyen, Lihui Chen, Senior Member, IEEE, and Chee Keong Chan and Jiawei Han and Micheline Kamber having provided me valuable information about the progress in my work, interms of their paper publications and text book.

References

1. 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.
2. I. Guyon, U.V. Luxburg, and R.C. Williamson, “Clustering:Science or Art?,” Proc. NIPS Workshop Clustering Theory, 2009.
3. S. Zhong, “Efficient Online Spherical K-means Clustering,” Proc.IEEE Int’l Joint Conf. Neural Networks (IJCNN), pp. 3180-3185, 2005.
4. H. Zha, X. He, C. Ding, H. Simon, and M. Gu, “Spectral Relaxationfor K-Means Clustering,” Proc. Neural Info. Processing Systems(NIPS), pp. 1057-1064, 2001.
5. I. Dhillon and D. Modha, “Concept Decompositions for LargeSparse Text Data Using Clustering,” Machine Learning, vol. 42,nos. 1/2, pp. 143-175, Jan. 2001.
6. W. Xu, X. Liu, and Y. Gong, “Document Clustering Based on Non-Negative Matrix Factorization,” Proc. 26th Ann. Int’l ACM SIGIR Conf. Research and Development in Informaion Retrieval, pp. 267-273,2003.
7. I.S. Dhillon, S. Mallela, and D.S. Modha, “Information-TheoreticCo-Clustering,” Proc. Ninth ACM SIGKDD Int’l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.
8. C.D. Manning, P. Raghavan, and H. Schu¨ tze, An Introduction toInformation Retrieval. Cambridge Univ. Press, 2009. C. Ding, X. He, H. Zha, M. Gu, and H. Simon, “A Min-Max CutAlgorithm for Graph Partitioning and Data Clustering,” Proc. IEEE Int’l Conf. Data Mining (ICDM), pp. 107-114, 2001.

AUTHOR(S) PROFILE

Annavazula Mrinalini received B.Tech degree in Computer Science and Information Technology from Madina Engineering College, JNTUA University, Anantapuram, A.P, India in 2010 and currently pursuing M.Tech, Computer Science and Engineering, final semester, from Sri Venkateswara University College of Engineering, TIRUPATI, A.P, India. Her interested areas are Data Mining, Software Engineering, and Software Architecture. She attended Two international Conferences and Two National Conferences during 2013 and 2014.

Dr. A. Rama Mohan Reddy working as a Professor of Computer Science and Engineering, Sri Venkateswara University, Tirupathi, A.P, India. He completed M.Tech (computer science) from NIT Warangal. He completed his Ph.D in the area of Software Architecture. He has more than 30 years of teaching experience. He published many papers in the peer-refereed journals and conferences. His interested areas are Software Engineering, Software Architecture, and Data Mining.