ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 9(10), DOI: 10.17485/ijst/2016/v9i10/88903, March 2016* ISSN (Online) : 0974-5645

Selecting Multiview Point Similarity from Different Methods of Similarity Measure to Perform Document Comparison

S. Kalpana[1*] and S. Vigneshwari[2]

[1]Department of Computer Science and Engineering, Sathyabama University, Chennai - 600119, Tamil Nadu, India; kalpanakameshwaran@gmail.com
[2]Faculty of computing, Sathyabama University, Chennai - 600119, Tamil Nadu, India; vikiraju@gmail.com

Abstract

Objective: The main objective is to implement multi view point similarity to perform document comparisons that use the concept of clustering. **Methods/Analysis**: The main task of data mining is clustering which is used to group or select objects which are similar to one another. Data mining divides whole document into meaningful clusters and analyses data. There are many different types of clustering methods like hierarchical clustering, partitioned clustering and data grouping may be based on distance, viewpoints, Euclidean distance etc,, Of these, the current system uses single view point similarity. This type of single view point similarity has some disadvantages. The main disadvantage is it does not use full set of document data so that detailed comparison measures cannot be revealed. In the future system multi viewpoint similarity is used to overcome the above disadvantage. **Findings:** The multi view point similarity method is used to overcome the disadvantages mentioned under the analysis. This method compares similarity between the multiple documents in detailed manner. The documents have been compared line by line and show the similarity. Then we have enhanced the existing ECSMTP algorithm and it is named as ECSMTP (Enhanced Concept Based Similarity Measure for Text Processing). This algorithm categorizes data from selected documents along with weight age of document, and based on that it forms clusters and calculates the similarity measure. Further in this system different kind of documents were compared like text documents, word, PDF documents etc., but it is not in the existing system. User may select kind of document and comparisons can be made on the selected documents. Clusters were formed and these clusters were compared.

Keywords: Clustering, ECSMTP, Multiviewpoint, Pattern Recognition, Singleview Point

1. Introduction

Clustering usually performs unsupervised classification to form clusters. Mainly clustering algorithms are used in earthquake studies, land use, insurance, city planning, marketing and biometrics. It is normally said as multi-objective optimization problem. Several new clustering algorithms have been developed day by day. More than half a century, the simple k-mean algorithm remains as top 10 data mining algorithm, says recent study.

For information retrieval processes clustering algorithms are must. The clustering algorithm can be used to find similarity between the documents mainly for information retrieval. The existing systems use single view point similarity. The drawback of single view point similarity is that, the cluster cannot exhibit the complete set of relationship among objects. So, in the future system new measure called multi-view point similarity is used.

The existing system performs similarity based on only words. Although this idea is statistically significant, the dimensions formed by the vector model distances such as Euclidean distance were very high. But it is not needed by high dimension and sparse domains. The existing algorithms are used to detect patterns, based on the saved terms in the library or word dictionary. It does not concentrate particularly on context. In future system similarity measure

* *Author for correspondence*

is based on information. More meaningful clusters are formed and are proven to be correct. This type of clustering can be used in applications where text documents are to be searched or processed frequently. In the proposed system ECSMTP (Concept Based Similarity Measure for Text Processing) algorithm is used.

2. Review of The Related Work

[1]Mainly concentrates on cluster analysis and gives detailed reviews about the clusters and different clustering algorithms and their features.

[2]Discussed about some of the comparisons between single and multi-view point similarities. Mainly the author explained that the quality of clusters made by a clustering algorithm depends on the quality of similarity measure. Different types of clustering algorithms are discussed.

[3]Selected a mechanism called multi view point similarity measure with incremental clustering and compares it with the cosine similarity for efficient result .The author has done validity test on MVS matrix created earlier and based on the validity score clusters were formed for similarity identification.

[4]The author compares five different types of algorithms that were used for cluster formation and similarity identification and concluded that multi view point similarity measure is more effective , efficient and accurate.

[5]Had done theoretical analysis and empirical study to develop a high performance mechanism to identify the similarity between documents. He explains the cosine similarity measure and multi view point similarity in detail. By using the use case diagram he depicts the clear picture of how objects were compared. Finally he proves that multi view point similarity is more efficient.

[6]Performs the similarity identification by considering certain features. Based on the results the similarity measure values can be obtained from the clusters easily. The three cases considered are a) The feature appears in both documents, b) the feature appears in only one document, and c) the feature appears in none of the documents.

[7]Performs document clustering based on frequent concepts. They proposed frequent concept based clustering (FCDC) which considers frequent concepts than frequent terms which are the cases of other clustering algorithm. FCDC was found to be more accurate, scalable and effective when compared with existing clustering algorithms like Bisecting K-means, UPGMA and FIHC.

[8]Explain about document clustering in detail and provide overview about all the existing clustering algorithms. Applications about document clustering are discussed. Also internal and external quality measures have been used to evaluate the document clustering algorithms.

[9]Gives information about several clustering techniques and performs comparison between the clustering algorithms based on some aspects like performance, speed, usage, etc,. Comparison based on precision and recall values are also performed.

[10]Presents conceptual rule mining on text clusters which performs comparison between web documents which include markup languages and databases; which is not in other concept based technique.

[11]Use incremental clustering and perform comparison with traditional k-means method. By this approach the clusters formed by single object were removed and such type of cluster is said to be singleton. This process of removal of singleton is repeated again and again until no singleton is found.

[12] Performs parallel comparison between documents to retrieve particular information from a particular document in a database in a quick manner. They developed three algorithms NLP with Semantic Matching technique for mining, K-means for clustering and PFT-sim for parallel comparison. It is proved that systems that use these three algorithms are GUI based systems and user friendly.

[13]Uses multi view point similarity by considering the TF and TDF term frequencies and performs parsing, cumulation, document similarity and clustering using traditional algorithm. Chandrasekhar et al[14], mainly concentrated on web document comparison and uses two clustering criteria functions IR and IC.

[15]Compares MVSCIR and MVSC-IV All available clustering algorithms.The author mainly concentrates on comparing web documents using incremental clustering with multiviewpoint similarity.

[16]Computes similarity between the documents based on semantic similarity which considers concept similarity than necessary lexically related data terms and items. She considers multiple ontologies than single ontology to give effective result like Wordnet and MeSH. (WordNet-Online lexical reference system, MeSH-Medical related terms)

[17]Computes dissimilarity between the documents and uses multi view point with cosine similarity to identify the similarity /dissimilarity. Here dissimilar documents obtained by the result of K-mean are given to multi view point method which provides corresponding clusters. From that similarity measure can be identified.

[18]Particularly focused on studying and utilizing cluster overlapping phenomenon to design cluster merging criteria. The authors proposed a new way to compute the overlap rate in order to improve time efficiency.

[19]Explained about the enhancement of the document retrieval system with personalization and security.

[20] Discussed about mining the social media data with the help of cross ontologies.

3. Materials and Methods

The concept of the paper is to measure the similarity between the documents based on the concept and information along with traditional words and terms. The proposed system uses Enhanced Concept Based Similarity Measure for Text Processing (ECSMTP) algorithm. Based on the references[4,7] the algorithm takes following steps:

Step 1: Initially grasp any text, word or PDF document for comparison.

Step 2: The comparison between the documents is performed based on word similarity

Step 3: For this comparison clusters are formed by splitting up the document based on category and weightage.

Step 4: Under the mutilation, integration and filtration processes the documents are compared accurately.

Step 5: Then similarity measure is obtained.

Using ECSMTP algorithm the proposed system extracts the contents from the different type of documents, and creates multi view point from different clusters. These clusters are formed based on the type and weight age of documents which is shown in Figure 5, which is usually said to be mutilation process. After these filter based semantic can be measured and integration among the documents can be viewed in order to measure similarity.

3.1 Experimental Setup

Figure 1, explains that documents can be given as input to the proposed system. Then document contents can be extracted and location information about the document can be created which is said to be metadata.

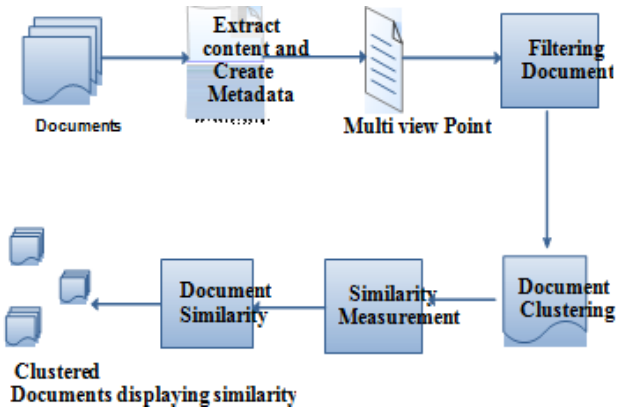

Figure 1. Experimental Setup.

Based on multi view point documents are filtered and based on weight and by using k- means clustering, the clusters are formed. From those clusters, similarity is successfully measured.

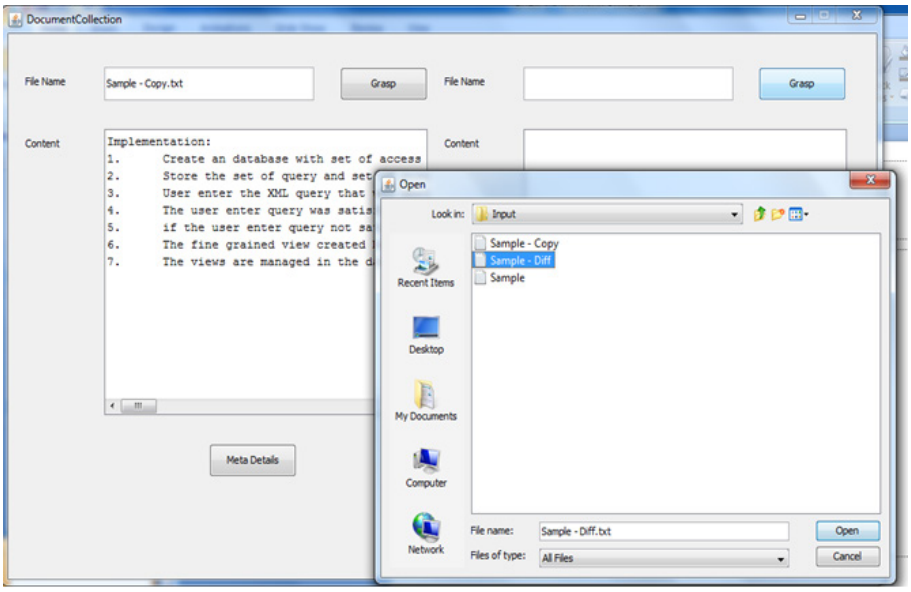

Figure 2. Grasping documents.

4. Discussion

The various steps involved in selecting the multi view point similarity is discussed based on the results obtained.

Figure 2, explains how the documents needed for comparison were extracted. It is done by clicking the grasp button.

Figure 3, explains that, upon clicking the similarity level button percentage of similarity between the documents can be displayed.

Figure 4, explains about how clusters were formed. In the proposed system it is based on word similarity and weight of each document is compared. By clicking on the fragment button those clusters were formed.

Figure 5, gives Meta data details about file or document on clicking Meta detail button.

Figure 6, explains how three different kinds of documents were compared. Three different documents compared are said to be document sets. The document set is created based on priority.

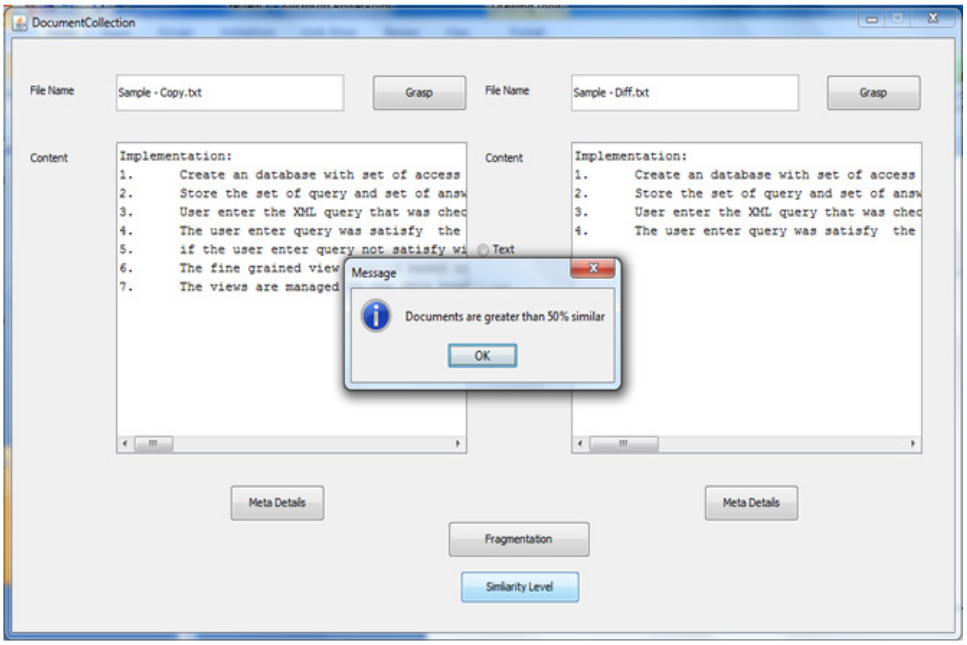

Figure 3. Similarity between documents.

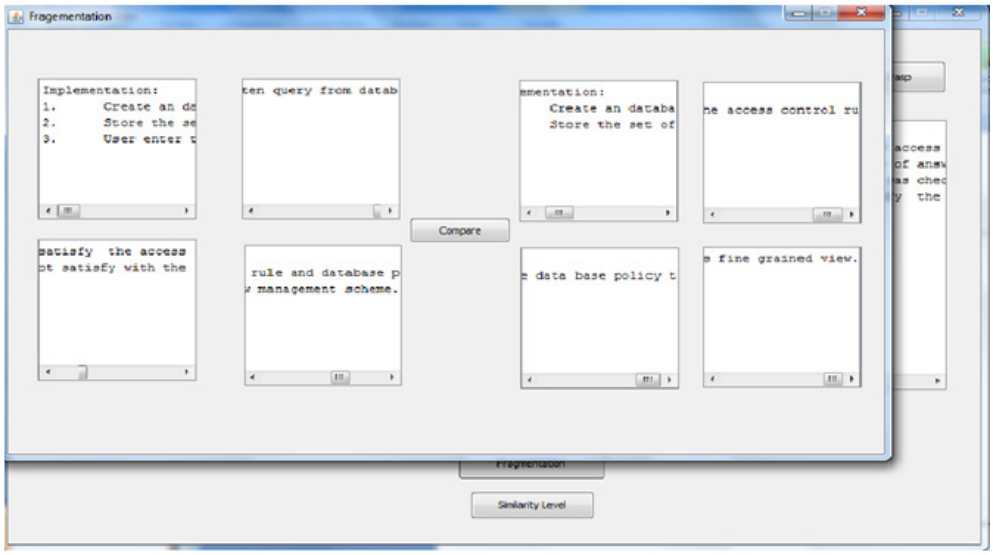

Figure 4. Cluster formation.

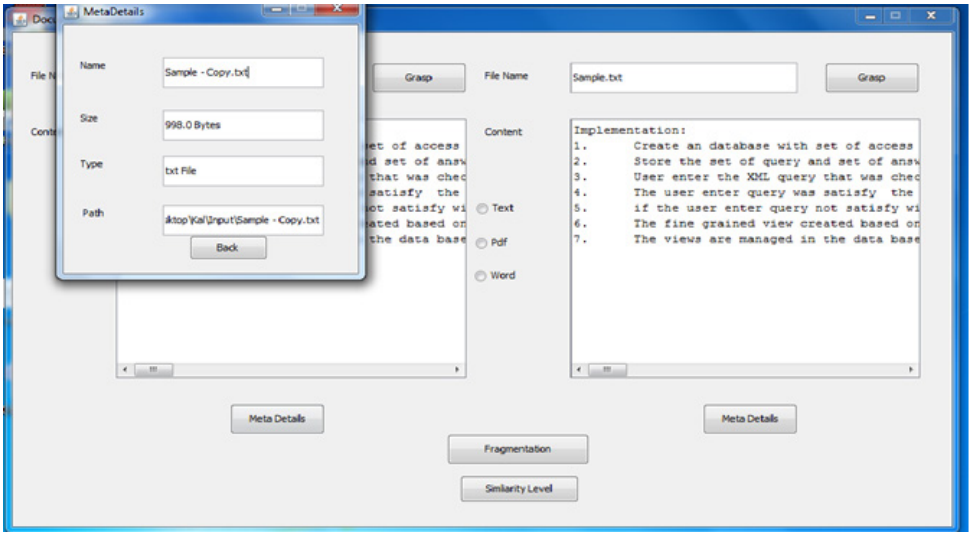

Figure 5. Meta data. Figure 5, gives Meta data details about file or document on clicking Meta detail button.

DOCUMENT SETS (ANY COMBINATION OF TEXT,PDF AND WORD)	DOC.SET1	DOC.SET2	DOC .SET3
	PRIORITY(>2→LOWAND >4→HIGH		
TEXT	2	4.1	4.4
PDF	4.3	3	5
WORD	5	5	2

Figure 6. Tabular Representation. Figure 6, explains how three different kinds of documents were compared. Three different documents compared are said to be document sets. The document set is created based on priority.

5. Conclusion

In this paper the concept of multi viewpoint based similarity measure for document comparison is analysed, which considers all possible sets of objects. Clusters are formed based on the object category and document weightage. ECSMTP algorithm is more accurate and produces appropriate similarity which is also accurate. The new measure aims at identifying more similarities with in the cluster and also between the clusters. The ECSMTP algorithm shows that it could afford significantly advanced clustering execution, than existing methods that use distinctive methods of similarity measure on a very large number of document data sets concealed by various assessment metrics.

6. References

1. Pande SR, Sambare SS, Thakre VM. Data Clustering Using Data Mining Techniques. International Journal of Advanced Research in Computer and Communication Engineering. 2012; 1(8):494–99.
2. Sriramoju SB. Multi View Point Measure for Achieving Highest Intra-Cluster Similarity. International Journal of Innovative Research in Computer and Communication Engineering. 2014; 2(3):3265–71.
3. Mohan Babu Chowdary S, Priyanka Chadalavada S. Multi View Point Similarity Measure with Incremental Clustering. International Journal of Computer Science Technology. 2013; 4(4):159–61.
4. Ramesh K, Vasumurthy C, Venkatesh D. High Quality Assessment of Similarity by Using Multiple View Points. International Journal of Emerging Technology in Computer Science and Electronics. 2014; 9(3):72–4.
5. Prasanna KAVL, Kumar V. Performance evaluation of multiview point-based similarity measure for data clustering. Journal of Global Research Computer Science. 2012; 3(11):21–26.
6. Lin Y-S, Jiang J-Y, Lee S-J. A Similarity Measure for Text Classification and Clustering. IEEE Transactions on Knowledge and Data Engineering. 2014; 26(7): 1575–90.
7. Baghel R, Dhir R. A Frequent Concept Based Document Clustering Algorithm. International Journal of Computer

Applications. 2010; 4(5):6–12.
8. Shah N, Maharajan S. Document Clustering: A Detailed Review. International Journal of Applied Information Systems. 2012; 4(5):30–38.
9. Prabha MS, Duraiswamy K, Sharmila MM. Analysis of Different Clustering Techniques in Data and Text Mining. International Journal of Computer Science Engineering. 2014; 3(2):107–16.
10. Navaneethakumar VM, Chandrasekar C. A Consistent Web Documents Based Text Clustering Using Concept Based Mining Model. IJCSI International Journal of Computer Science Issues. 2012; 9(4):365–70.
11. Warad VC, Baron Sam B. Incremental MVS based Clustering Method for Similarity Measurement. International Journal of Computer Science and Information Technologies. 2014; 5(2):1486–91.
12. Palsaniya PP, Dhanwani DC. Survey on Parallel Comparison of Text Document with Input Data Mining and VizSFP. International Journal of Science and Research. 2014; 3(11):1569–73.
13. Sindhudarshini S, Suresh GS. Assessment of Document Similarity with Clustering Using Multi View-Points. International Journal of Computer Science and Information Technologies. 2015; 6(3):2815–18.
14. Chandrasekhar S, Sasidhar K, Vajralu M.Study and Analysis of Multi-viewpoint clustering with similarity measures. International Journal of Emerging Technology and Advanced Engineering. 2012; 2(10):606–609.
15. Sailaja G, Dhanalakshmi B, Bharathis Y, Ramesh Reddy C. An Incremental clustering algorithm for multi–viewpoint similarity measure. International Journal of Emerging Trends Technology Computer Science. 2013; 2(1):208–12.
16. Jayasri D, Manimegalai D. Semantic Similarity Measures on Different Ontologies: Survey and a Proposal of Cross Ontology based Similarity Measure. International Journal of Science Research (IJSR). 2013; 2(2):455–61.
17. Mrinalini A, Rama Mohan Reddy A. Implementation of Multi View point method for similarity Measure in clustering the documents. International Journal of Advanced Research in Computer Science and Mangement Studies. 2014; 2(1):200–205.
18. Bhonde S, Chawan PM, Chauhan P. Multi-viewpoint Based Similarity Measure and Optimality Criteria for Document Clustering. International Journal of Advanced Research in Computer Science and Software Engineering. 2012; 2(6):232–36.
19. Vigneshwari S, Aramudhan M. Personalized cross ontological framework for secured document retrieval in the cloud. National Academy Science Letters-India. 2015; 38 (5):421–24.
20. Vigneshwari S, Aramudhan M.Social Information Retrieval Based on Semantic Annotation and Hashing upon the Multiple Ontologies. Indian Journal of Science and Technology. 2015; 8(2):103–107.