WWW.IJRPR.COM

International Journal of Research Publication and Reviews

Journal homepage: www.ijrpr.com ISSN 2582-7421

MULTI-VIEWPOINT BASED SIMILARITY MEASURE FOR DOCUMENT CLUSTERING

[a]Dr.S.Hemalatha, [b]M.Manikandan

[a]Associate Professor, Department of Computer Science, Karpagam Academy of Higher Education
[b]PG Student, Department of Computer Science, Karpagam Academy of Higher Education

A B S T R A C T

Clustering is a process of partitioning a set of data (or object) in a set of meaningful sub-classes, called clusters. Cluster is a collection of data that are similar to one another and thus can be treated collectively as one group..All clustering methods have to assume some cluster relationship among the data objects that they are applied on. This paper introduces a novel multiview point-based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and this paper is that the former uses only a single viewpoint, which is the origin, while the latter utilizes many different viewpoints, which are objects assumed to not be in the same cluster with the two objects being measured. Using multiple viewpoints, more informative assessment of similarity could be achieved. Theoretical analysis and empirical study are conducted to support this claim. Two criterion functions for document clustering are proposed based on this new measure. This paper gives the detailed study on comparing them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of this proposal.

Keywords: Document clustering, text mining, similarity measure, clustering methods.

1. Introduction

Clustering is the classification of objects into different groups, or more precisely, the partitioning of a data set into subsets (clusters), so that the data in each subset (ideally) share some common trait - often proximity according to some defined distance measure. Data clustering is a common technique for statistical data, which is used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics. The computational task of classifying the data set into k clusters is often referred to as k-clustering. Besides the term data clustering (or just clustering), there are a number of terms with similar meanings, including cluster analysis, automatic classification, numerical taxonomy, botryology and typological analysis. Document clustering aims to group, in an unsupervised way, a given document set into clusters such that documents within each cluster are more similar between each other than those in different clusters. It is an enabling technique for a wide range of information retrieval tasks such as efficient organization, browsing and summarization of large volumes of text documents. Cluster analysis aims to organize a collection of patterns into clusters based on similarity. Clustering has its root in many fields, such as mathematics, computer science, statistics, biology, and economics. In different application domains, a variety of clustering techniques have been developed, depending on the methods used to represent data, the measures of similarity between data objects, and the techniques for grouping data objects into clusters.

Document clustering techniques mostly rely on single term analysis of the document data set, such as the Vector Space Model. To achieve more accurate document clustering, more informative features including phrases and their weights are particularly important in such scenarios. Document clustering is particularly useful in many applications such as automatic categorization of documents, grouping search engine results, building taxonomy of documents, and others. For this Hierarchical Clustering method provides a better improvement in achieving the result. Our project presents two key parts of successful

Hierarchical document clustering. The first part is a document index model, the Document Index Graph, which allows for incremental construction of the index of the document set with an emphasis on efficiency, rather than relying on single-term indexes only. It provides efficient phrase matching that is used to judge the similarity between documents. This model is flexible in that it could revert to a compact representation of the vector space model if we choose not to index phrases. The second part is an incremental document clustering algorithm based on maximizing the tightness of clusters by carefully watching the pair-wise document similarity distribution inside clusters. Both the phases are based upon two algorithmic models called Gaussian Mixture Model and Expectation Maximization. The combination of these two components creates an underlying model for robust and accurate document similarity calculation that leads to much improved results in Web document clustering over traditional methods.

2. Methodology

Distance Measure

An important step in any clustering is to select a distance measure, which will determine how the similarity of two elements is calculated. This will influence the shape of the clusters, as some elements may be close to one another according to one distance and further away according to another. For example, in a 2-dimensional space, the distance between the point (x=1, y=0) and the origin (x=0, y=0) is always 1 according to the usual norms, but the distance between the point (x=1, y=1) and the origin can be 2, $\sqrt{2}$ or 1 if you take respectively the 1-norm, 2-norm or infinity-norm distance.

- The Euclidean distance (also called distance as the crow flies or 2-norm distance). A review of cluster analysis in health psychology research found that the most common distance measure in published studies in that research area is the Euclidean distance or the squared Euclidean distance.
- The Manhattan distance (also called taxicab norm or 1-norm)
- The maximum norm
- The Mahalanobis distance corrects data for different scales and correlations in the variables
- The angle between two vectors can be used as a distance measure when clustering high dimensional data. See Inner product space.
- The Hamming distance (sometimes edit distance) measures the minimum number of substitutions required to change one member into another.

Document clustering has been studied intensively because of its wide applicability in areas such as web mining, search engines, information retrieval, and topological analysis. Unlike in document classification, in document clustering no labeled documents are provided. Although standard clustering techniques such as k-means can be applied to document clustering, they usually do not satisfy the special requirements for clustering documents: high dimensionality, high volume of data, ease for browsing, and meaningful cluster labels. In addition, many existing document clustering algorithms require the user to specify the number of clusters as an input parameter and are not robust enough to handle different types of document sets in a real-world environment. For example, in some document sets the cluster size varies from few to thousands of documents. This variation tremendously reduces the clustering accuracy for some of the state-of-the art algorithms. *Frequent Itemset-based Hierarchical Clustering (FIHC)*, for document clustering based on the idea of *frequent itemsets* proposed by Agrawalet. Al. The intuition of our clustering criterion is that there are some frequent itemsets for each cluster (topic) in the document set, and different clusters share few frequent itemsets. A frequent itemset is a set of words that occur together in some minimum fraction of documents in a cluster. Therefore, a frequent itemset describes something common to many documents in a cluster. In this technique use frequent itemsets to construct clusters and to organize clusters into a topic hierarchy. Here are the features of this approach.

- Reduced dimensionality.This approach use only the frequent items that occur in some minimum fraction of documents in document vectors, which drastically reduces the dimensionality of the document set. Experiments show that clustering with reduced dimensionality is significantly more efficient and scalable. This decision is consistent with the study from linguistics (Longman Lancaster Corpus) that only 3000 words are required to cover 80% of the written text in English and the result is coherent with the Zipf's law and the findings in Mladenic et al. and Yang et al.

- *High clustering accuracy*. Experimental results show that the proposed approach FIHC outperforms best documents clustering algorithms in terms of accuracy. It is robust even when applied to large and complicated document sets.
- Number of clusters as an optional input parameter. Many existing clustering algorithms require the user to specify the desired number of clusters as an input parameter. FIHC treats it only as an optional input parameter. Close to optimal clustering quality can be achieved even when this value is unknown.

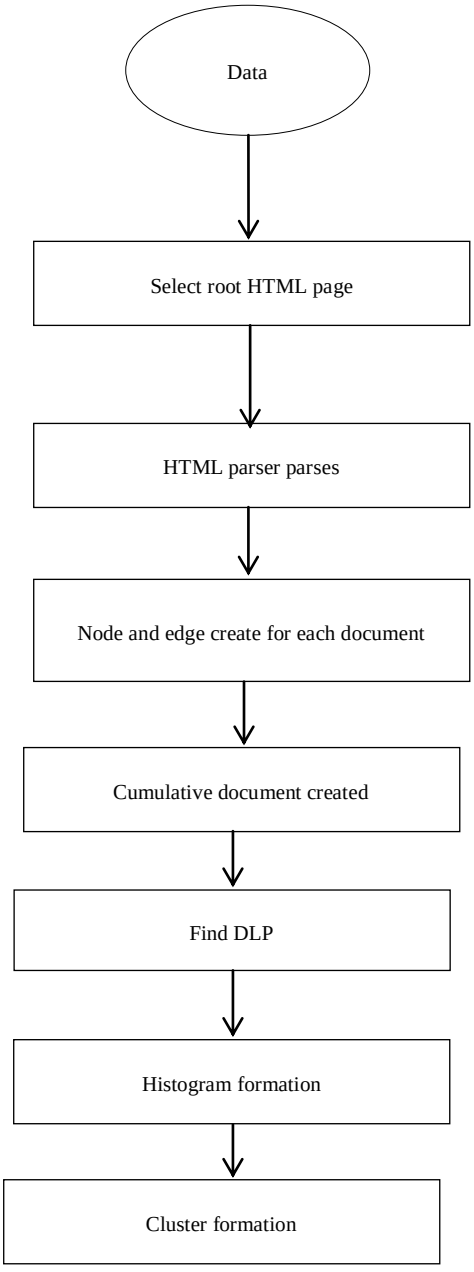

Fig1 System architecture

3. Conclusion

This paper makes a study on new similarity measure known as MVS (Multi-Viewpoint based Similarity). When it is compared with cosine similarity, MVS is more useful for finding the similarity of text documents. . IR and IV are the two criterion functions proposed based on MVS. The respective clustering algorithms are also introduced. The proposed

scheme is tested with large datasets with various evolution metrics. The results reveal that the clustering algorithm provides performance that is better than much state – of – the - art clustering algorithms.

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] A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, "Clustering on the Unit Hyper sphere Using Von Mises-Fisher Distributions,"J. Machine Learning Research, vol. 6, pp. 1345-1382, Sept. 2005.
[4] 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 Information Retrieval, pp. 267-273,2003.
[5] I.S. Dhillon, S. Mallela, and D.S. Modha, "Information-Theoretic Co-Clustering," Proc. Ninth ACM SIGKDD Int'l Conf. Knowledge Discovery and Data Mining (KDD), pp. 89-98, 2003.
[6]A. Strehl, J. Ghosh, and R. Mooney, "Impact of Similarity Measures on Web-Page Clustering," Proc. 17th Nat'l Conf. Artificial Intelligence: Workshop of Artificial Intelligence for Web Search (AAAI), pp. 58-64, July 2000
[7] I. Dhillon and D. Modha, "Concept Decompositions for Large Sparse Text Data Using Clustering," Machine Learning, vol. 42,nos. 1/2, pp. 143-175, Jan. 2001.
[8]. A. Banerjee, I. Dhillon, J. Ghosh, and S. Sra, ―Clustering on the unit hypersphere using von Mises-Fisher distributions,‖ *J. Mach. Learn. Res.*,vol. 6, pp. 1345–1382, Sep 2005.
[9] W. Xu, X. Liu, and Y. Gong, ―Document clustering based on Nonnegative matrix factorization,‖ in *SIGIR*, 2003, pp. 267–273.
[10] M Praneesh and Jaya R Kumar. Article: Novel Approach for Color based Comic Image Segmentation for Extraction of Text using Modify Fuzzy Possiblistic C-Means Clustering Algorithm. IJCA Special Issue on Information Processing and Remote Computing IPRC(1):16-18, August 2012. Published by Foundation of Computer Science, New York, USA.