STAR
Research Journal

Available online at www.starresearchjournal.com (Star International Journal)

ENGINEERING

MULTI VIEWPOINT BASED SIMILARITY MEASURE IN P2P CLUSTERING USING PCP2P ALGORITHM

A.Priyadharshini[1], V. Kumar[2], R.Thiyagarajan[3] & S. Karthikeyan[4]

[1]*Assistant Professor, Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, Tamil Nadu, India.*
[2]*Professor & Head, Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, Tamil Nadu, India.*
[3]*Assistant Professor, Department of Electronics and Communication Engineering, Shreenivasa Engineering College, Dharmapuri, Tamil Nadu, India.*
[4]*Assistant Professor, Department of Computer Science and Engineering, Knowledge Institute of Technology, Salem, Tamil Nadu, India.*

Abstract

The Clustering methodologies should have some relationship among the cluster data objects. The pair of objects in similarity will be found implicitly or explicitly. Here we introduce a novel multi viewpoint-based similarity measure and two of its relative clustering methodologies. The Traditional dissimilarity/similarity measure is based on only a single viewpoint and that is the origin point. In our proposed methodology we introduce a different viewpoint concept even the object should not be in same cluster. Due to this we can achieve more similarity informative assessment. To prove this we are accomplishing the theoretical analysis and empirical study. We can use PCP2P algorithm for peer-to-peer clustering. The advantage of our proposed system by comparing with familiar clustering algorithms is high scalability for assigning documents to clusters.

Keywords: Clustering, P2P algorithm.

INTRODUCTION

Data mining (the analysis step of the "Knowledge Discovery in Databases" process, or KDD), an interdisciplinary subfield of computer science, is the computational process of discovering patterns in large data sets involving methods at the intersection of artificial intelligence, machine learning, statistics, and database systems. The overall goal of the data mining process is to extract information from a data set and transform it into an understandable structure for further use. Aside from the raw analysis step, it involves database and data management aspects, data pre-processing, model and inference considerations, interestingness metrics, complexity considerations, post-processing of discovered structures, visualization, and updating. The mining application has got rich focus due to its significance of classification algorithms. The comparison of classification algorithm is a complex and it is an open problem. First, the notion of the performance can be defined in many ways: accuracy, speed, cost, Reliability, etc. Second, an appropriate tool is necessary to quantify this performance. Third, a consistent method must be selected to compare with the measured values .It is the process of transforming raw data into actionable information that is nontrivial, previously unknown and is potentially valuable to the user.

Data mining techniques are used in a variety of fields including marketing and business intelligence, biotechnology, multimedia, and security. As a result, data mining algorithms have become increasingly complex, incorporating more functionality than in the past. Consequently, there is a need for faster execution of these algorithms, which creates ample opportunities for algorithmic and architectural optimizations. A number of data mining techniques have already been done on educational data mining to improve the performance of students like Regression, Genetic algorithm, Bays classification, k-means clustering, associate rules, prediction etc. Data mining techniques can be used in educational field to enhance our understanding of learning process to focus on identifying, extracting and evaluating variables related to the learning process of students.

RELATED WORK

Existing systems use the text clustering algorithms only in P2P system. In this method the clusters are documented without a central server. An important data mining technique is used here. And it's

useful for information retrieval and Challenging because of their network size, and high dimensionality of documents and cluster centroids. It enhances information retrieval efficiency and effectiveness. Disadvantages: Clustering is performed on a dedicated node only, It also fails to scale on highly distributed environments.

PROPOSED WORK

In the proposed system we use the PCP2P with MVS method. Due to this approximation to reduce the network and computational cost and Compare each document only with the *most promising* clusters. And it contains two selection processes, one is pre-filtering step and another one is full comparison step. In Pre-filtering step we find the candidate clusters for a document using an inverted index. And the Full comparison step use compact cluster summaries to exclude more candidate clusters. High scalability by using a probabilistic approach for assigning documents to clusters. Advantages: It offers guarantees for the correctness of each document assignment to a cluster, Experimental evaluation:1 million peers and 1 million documents demonstrate the scalability and effectiveness of the algorithm.

METHODOLOGIES

Cluster Documentation

Document clustering involves the use of descriptors and descriptor extraction. Descriptors are sets of words that describe the contents within the cluster. Document clustering is generally considered to be a centralized process. Examples of document clustering include web document clustering for search users. The application of document clustering can be categorized to two types, online and offline. Online applications are usually constrained by efficiency problems when compared offline applications. In general, there are two common algorithms. The first one is the hierarchical based algorithm, which includes single link, complete linkage, group average and Ward's method. By aggregating or dividing, documents can be clustered into hierarchical structure, which is suitable for browsing. However, such an algorithm usually suffers from efficiency problems. The other algorithm is developed using the K-means algorithm and its variants. These algorithms can further be classified as hard or soft clustering algorithms. Hard clustering computes a hard assignment – each document is a member of exactly one cluster. The assignment of soft clustering algorithms is soft – a document's assignment is a distribution over all clusters. In a soft assignment, a document has fractional membership in several clusters. Dimensionality reduction methods can be considered a subtype of soft clustering; for documents, these include latent semantic indexing (truncated singular value decomposition on term histograms) and topic models.

Other algorithms involve graph based clustering, ontology supported clustering and order sensitive clustering. Given a clustering, it can be beneficial to automatically derive human-readable labels for the clusters. Various methods exist for this purpose.

Cluster Assignment

Document assignment consists of two steps, pre-selection and full comparison. In the pre-selection step, the peer holding d retrieves selected cluster summaries from the DHT index, to identify the most relevant clusters. Pre-selection step already filters out most of the clusters. In the full comparison step, the peer computes similarity score estimates for d using the retrieved cluster summaries. Clusters with low similarity estimates are filtered out and the document is sent to the few remaining cluster holders for full similarity computation. Finally, d is assigned to the cluster with the highest similarity. This two-stage filtering algorithm reduces drastically the number of full comparisons (usually less than five comparisons per document, independent of the number of clusters). Both cluster indexing and document assignment is repeated periodically to compensate churn, and to maintain an up-to-date clustering solution. The algorithm enables controlling the tradeoff between the network cost and the clustering quality. In particular, the cluster indexing activity, as well as the pre-selection and full comparison steps, are configured using the results of a probabilistic analysis, thereby providing probabilistic guarantees that the resulting clustering solution.

MVS

All clustering methods have to assume some cluster relationship among the data objects that they are applied on. Similarity between a pair of objects can be defined either explicitly or implicitly. We introduce a novel multi viewpoint-based similarity measure and two related clustering methods. The major difference between a traditional dissimilarity/similarity measure and ours 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. We compare them with several well-known clustering algorithms that use other popular similarity measures on various document collections to verify the advantages of our proposal.

Cluster Indexing

Cluster holders are responsible for indexing the summaries of the clusters in the DHT. Particularly, each cluster holder periodically re computes the cluster centroid using the documents assigned to the cluster at the time. It also re computes a cluster summary and publishes it to the DHT index, using selected cluster terms as keys.

Each cluster holder selects random peers to serve as helper cluster holders, and replicates the cluster centroid to them. Their IP addresses are also included in the cluster summaries, so that peers can randomly choose

a helper for comparing their documents with the cluster centroid without going through the cluster holder. Communication between the master and helper cluster holders only takes place for updating the centroids; load balancing does not impede the system scalability.

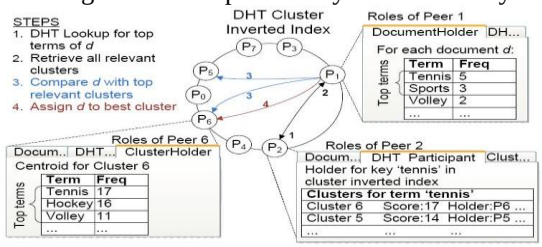

Fig. 1 Distributed Hash Table

PEER- PEER

A peer-to-peer(P2P) network is a type of decentralized and distributed architecture in which individual nodes in the network (called *"peers"*) act as both suppliers and consumers of resources, in contrast to centralized client–server model where client nodes request access to resources provided by central servers. Networks in which all computers have equal status are called peer-to-peer or P2P networks.

In a peer-to-peer network, tasks (such as searching for files or streaming audio/video) are shared amongst multiple interconnected peers who each make a portion of their resources (such as processing power, disk storage or network bandwidth) directly available to other network participants, without the need for centralized coordination by servers.

PCPP2P Algorithm

To optimize the fit between the data and the model using a probabilistic approach., cluster can be represented by a parametric distribution, like a *Gaussian* (continuous) or a *Poisson* (discrete),The entire data set is modeled by a *mixture* of these distributions.

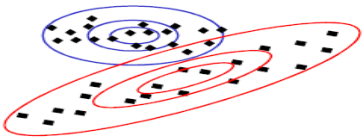

Fig: 2 Mixture of Multivariate Gaussian

PERFORMANCE EVALUATION OF MVSC

To verify the advantages of our proposed methods, we evaluate their performance in experiments on document data. The objective of this section is to compare MVSC-IR and MVSC-IV with the existing algorithms that also use specific similarity measures and criterion functions for document clustering. The similarity measures to be compared include Euclidean distance, cosine similarity, and extended Jacquard coefficient.

PERFORMANCE EVALUATION OF MVSC

To verify the advantages of our proposed methods, we evaluate their performance in experiments on document data. The objective of this section is to compare MVSC-IR and MVSC-IV with the existing algorithms that also use specific similarity measures and criterion functions for document clustering. The similarity measures to be compared include Euclidean distance, cosine similarity, and extended Jacquard coefficient.

RESULTS

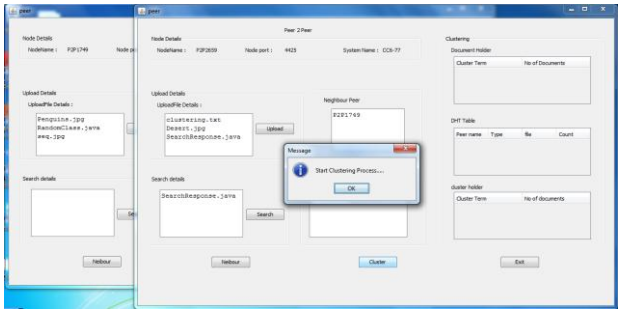

Fig.3. Uploading file and Clustering

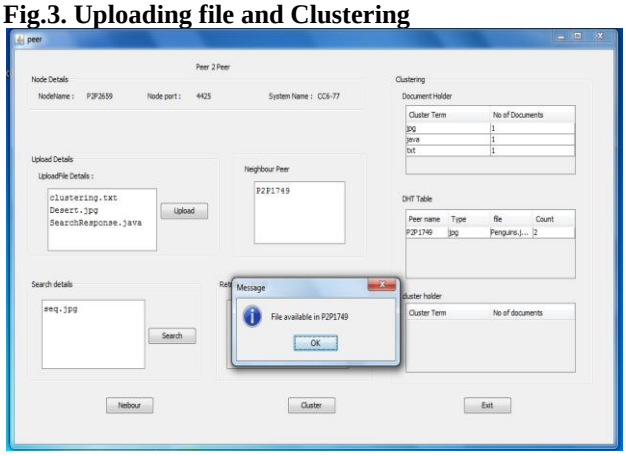

Fig. 4 Searching File

CONCLUSION

Multiviewpoint-based Similarity measuring method, named MVS. Theoretical analysis and empirical examples show that MVS is potentially more suitable for text documents than the popular cosine similarity. Based on MVS, two criterion functions, IR and IV, and their respective clustering algorithms, MVSC-IR and MVSC-IV, have been introduced. Compared with other state-of-the-art clustering methods that use different types of similarity measure, on a large number of document data sets and under different evaluation metrics, the proposed algorithms show that they could provide significantly improved clustering performance. The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints.

REFERENCES

[1] Ahmad.A and Dey.L (2007), "A Method to Compute Distance between Two Categorical Values of Same Attribute in Unsupervised Learning for Categorical Data Set,".

[2] Chim and X. Deng, "Efficient Phrase-Based Document Simiiarity for Ciustering," IEEE Trans. Knowledge and Data Eng., voi. 20, no. 9, pp. 1217-1229, Sept.2008.

[3] D. Lee and J. Lee, "Dynamic Dissimilarity Measure for Support Based Clustering," IEEE Trans. Knowledge and Data Eng., vol. 22,no. 6, pp. 900-905, June 2010.

[4] 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.

[5] I.M. Pelillo, "What Is a Cluster? Perspectives from Game Theory,"Proc. NIPS Workshop Clustering Theory, 2009.

[6] Lenco.D, Pensa.R.G, and Meo.R(2009), "Context-Based Distance Learning for Categorical Data Clustering," Proc. Eighth Int'l Symp. Intelligent Data Analysis (IDA), pp. 83-94.

[7] Lakkaraju.P, Gauch.S, and Speretta.M(2008), "Document Similarity Based on Concept Tree Distance," Proc. 19th ACM Conf. Hypertext and Hypermedia, pp. 127-132.