IJETST- Volume||01||Issue||04||Pages 520-528||June||ISSN 2348-9480 2014

Open access Journal International journal of Emerging Trends in Science and Technology

Study on Framework for Efficient Document Clustering on Centralized System

Authors

Divyashree G[1], Mrs. AncyThomas[2]

[1]Sapthagiri College of Engineering, Bangalore, India

[2]Assistant Professor, Dept of CS&E, Sapthagiri College of Engineering Bangalore, India

E-Mail: *divyashreeraju@gmail.com*

Abstract

The word "cluster" is used broadly in computer networking to refer to a number of different implementations of shared computing resources. Typically, a cluster integrates the resources of two or more computing devices (that could otherwise function separately) together for some common purpose. In this paper we have presented brief introduction about cluster. We also presented document cluster, similarity measure and cosine similarity.

Keywords: *-Cluster, Similarity Measure.*

INTRODUCTION

The concept of similarity is fundamentally important in almost every scientific field. For example, in mathematics, geometric methods for assessing similarity are used in studies of congruence and homothetic as well as in allied fields such as trigonometry. Topological methods are applied in fields such as semantics. Graph theory is widely used for assessing cladistic similarities in taxonomy. Fuzzy set theory has also developed its own measures of similarity, which find application in areas such as management, medicine and meteorology. An important problem in molecular biology is to measure the sequence similarity of pairs of proteins.

A review or even a listing of all the uses of similarity is impossible. Instead, this article focuses on perceived similarity. The degree to which people perceive two things as similar fundamentally affects their rational thought and behavior. Negotiations between politicians or corporate executives may be viewed as a process of data collection and assessment of the similarity of hypothesized and real motivators. The appreciation of a fine fragrance can be understood in the same way. Similarity is a core element in achieving an understanding of variables that motivate behavior and mediate affect.

Not surprisingly, similarity has also played a fundamentally important role in psychological experiments and theories. For example, in many experiments people are asked to make direct or indirect judgments about the similarity of pairs of objects. A variety of experimental techniques are used in these studies, but the most common are to ask subjects whether the objects are the same or different, or to ask them to produce a number, between say 1 and 7, that matches their feelings

about how similar the objects appear (e.g., with 1 meaning very dissimilar and 7 meaning very similar). The concept of similarity also plays a crucial but less direct role in the modeling of many other psychological tasks. This is especially true in theories of the recognition, identification, and categorization of objects, where a common assumption is that the greater the similarity between a pair of objects, the more likely one will be confused with the other. Similarity also plays a key role in the modeling of preference and liking for products or brands, as well as motivations for product consumption.

A computer cluster consists of a set of loosely connected or tightly connected computers that work together so that in many respects they can be viewed as a single system.

The components of a cluster are usually connected to each other through fast local area networks ("LAN"), with each node (computer used as a server) running its own instance of an operating system. Computer clusters emerged as a result of convergence of a number of computing trends including the availability of low cost microprocessors, high speed networks, and software for high performance distributed computing.

Clusters are usually deployed to improve performance and availability over that of a single computer, while typically being much more cost-effective than single computers of comparable speed or availability.

Computer clusters have a wide range of applicability and deployment, ranging from small business clusters with a handful of nodes to some of the fastest supercomputers in the world such as IBM's Sequoia.

The desire to get more computing power and better reliability by orchestrating a number of low cost commercial off-the-shelf computers has given rise to a variety of architectures and configurations.

CLUSTERING

Clustering is useful in several exploratory pattern-analysis, grouping, decision making, and machine-learning situations; including data mining, document retrieval, image segmentation, and pattern classification. However, in many such problems, there is little prior information (e.g., statistical models) available about the data, and the decision-maker must make as few assumptions about the data as possible. It is under these restrictions that clustering methodology is particularly appropriate for the exploration of interrelationships among the data points to make an assessment (perhaps preliminary) of their structure. The term "clustering" is used in several research communities to describe methods for grouping of unlabeled data. These communities have different terminologies and assumptions for the components of the clustering process and the context in which clustering is used. Thus, we face a dilemma regarding the scope of this survey. The production of a truly comprehensive survey would be a monumental task given the sheer mass of literature in this area. The accessibility of the survey might also be questionable given the need to reconcile very different vocabularies and assumptions regarding clustering in the various communities.

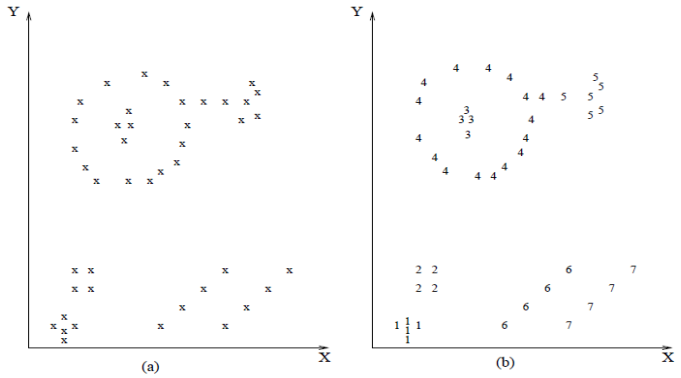

Figure.1 Data Clustering

COMPONENTS OF CLUSTER TASK

Typical pattern clustering activity involves the following steps:-

(1) Pattern representation (optionally including feature extraction and/or selection),

(2) Definition of a pattern proximity measure appropriate to the data domain,

(3) Clustering or grouping,

(4) Data abstraction (if needed), and

(5) Assessment of output (if needed).

Figure 2 depicts a typical sequencing of the first three of these steps, including a feedback path where the grouping process output could affect subsequent feature extraction and similarity computations.

Pattern representation refers to the number of classes, the number of available patterns, and the number, type, and scale of the features available to the clustering algorithm. Some of this information may not be controllable by the practitioner. Feature selection is the process of identifying the most effective subset of the original features to use in clustering. Feature extraction is the use of one or more transformations of the input features to produce new salient features. Either or both of these techniques can be used to obtain an appropriate set of features to use in clustering.

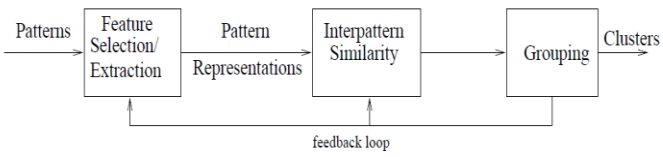

Figure 2 Stages in clustering

CLUSTERING METHOD [6]

Cluster is a collection of objects which are 'similar' between them and are 'dissimilar' to the objects belonging to other clusters [1]; and a clustering algorithm aims to find a natural structure or relationship in an unlabeled data set. There are several categories of clustering algorithms.

Some of the algorithms are hierarchical and probabilistic. A hierarchical algorithm clustering algorithm is based on the union between the two nearest clusters. The beginning condition is realized by setting every datum as a cluster. After a few iterations, it reaches the final clusters wanted. The final category of probabilistic algorithms is focused around model matching using probabilities as opposed to distances to decide clusters. EM or Expectation Maximization is an example of this type of clustering algorithm. In [2], Pen et al. utilized cluster analysis composed of 2 methods. In Method I, a majority voting committee with 3 results generates the final analysis result. The performance measure of the classification is decided by majority vote of the committee. If more than 2 of the committee members give the same classification result, then the clustering analysis for that observation is successful; otherwise, the analysis fails. Kalton et al. [3] did clustering and after letting the algorithm create its own clusters, added a step. After the clustering was completed each member of a class was assigned the value of the cluster's majority population. The authors noted that the approach loses detail, but allowed them to evaluate each clustering algorithm against the "correct" clusters. Pattern proximity is usually measured by distance function defined on pairs of patterns. A variety of distance measures are in use in the various communities [4, 5, 6]. A simple distance measure like Euclidean distance can often be used to reflect dissimilarity between two patterns, whereas other similarity measures can be used to characterize the conceptual similarity between patterns [7]. The grouping step can be performed in a number of ways. The output clustering (or clustering's) can be hard (a partition of the data into groups) or fuzzy (where each pattern has a variable degree of membership in each of the output clusters). Hierarchical clustering algorithms produce

a nested series of partitions based on a criterion for merging or splitting clusters based on similarity. Partition clustering algorithms identify the partition that optimizes (usually locally) a clustering criterion. Data abstraction is the process of extracting a simple and compact representation of a data set. Here, simplicity is either from the perspective of automatic analysis (so that a machine can perform further processing efficiently) or it is human-oriented (so that the representation obtained is easy to comprehend and intuitively appealing). In the clustering context, a typical data abstraction is a compact description of each cluster, usually in terms of cluster prototypes or representative patterns such as the centroid [6].

How is the output of a clustering algorithm evaluated? What characterizes a 'good' clustering result and a 'poor' one? All clustering algorithms will, when presented with data, produce clusters — regardless of whether the data contain clusters or not. If the data does contain clusters, some clustering algorithms may obtain 'better' clusters than others. The assessment of a clustering procedure's output, then, has several facets. One is actually an assessment of the data domain rather than the clustering algorithm itself— data which do not contain clusters should not be processed by a clustering algorithm. The study of cluster tendency, wherein the input data are examined to see if there is any merit to a cluster analysis prior to one being performed, is a relatively inactive research area, and will not be considered further in this survey. The interested reader is referred to [8] and [9] for information. How is the output of a clustering algorithm evaluated? What characterizes a 'good' clustering result and a 'poor' one? All clustering algorithms will, when presented with data, produce clusters — regardless of whether the data contain clusters or not. If the data does contain clusters, some clustering algorithms may obtain 'better' clusters than others. The assessment of a clustering procedure's output, then, has several facets. One is actually an assessment of the data domain rather than the clustering algorithm itself— data which do not contain clusters should not be processed by a clustering algorithm. The study of cluster tendency, wherein the input data are examined to see if there is any merit to a cluster analysis prior to one being performed, is a relatively inactive research area, and will not be considered further in this survey. The interested reader is referred to [8] and Cheng [9] for information.

Cluster validity analysis, by contrast, is the assessment of a clustering procedure's output. Often this analysis uses a specific criterion of optimality; however, these criteria are usually arrived at subjectively. Hence, little in the way of 'gold standards' exist in clustering except in well-prescribed subdomains. Validity assessments are objective [10] and are performed to determine whether the output is meaningful. A clustering structure is valid if it cannot reasonably have occurred by chance or as an artifact of a clustering algorithm. When statistical approaches to clustering are used, validation is accomplished by carefully applying statistical methods and testing hypotheses. There are three types of validation studies. An external assessment of validity compares the recovered structure to an a priori structure. An internal examination of validity tries to determine if the structure is intrinsically appropriate for the data. A relative test compares two structures and measures their relative merit.

INTRODUCTION OF DOCUMENT CLUSTER

Document clustering groups similar documents that to form a coherent cluster, while documents that are different have separated apart into different clusters. Clustering of text documents plays a vital role in efficient Document Organization, Summarization, Topic Extraction and Information Retrieval. Initially used for improving the precision or recall

in an Information Retrieval System .more recently, clustering has been proposed for use in browsing a collection of documents or in organizing the results returned by a search engine in response to user's query or help users quickly identify and focus on the relevant set of results. However, the definition of a pair of documents being similar or different is not always clear and normally varies with the actual problem setting. For example, when clustering research papers, two documents are regarded as similar if they share similar thematic topics. When clustering is employed on web sites, we are usually more interested in clustering the component pages according to the type of information that is presented in the page. For instance, when dealing with universities" web sites, we may want to separate professors" home pages from students" home pages, and pages for courses from pages for research projects. This kind of clustering can benefit further analysis and utilize of the dataset such as information retrieval and information extraction, by grouping similar types of information sources together. Accurate clustering requires a precise definition of the closeness between a pair of objects, in terms of either the pair wised similarity or distance. A variety of similarity or distance measures have been proposed and widely applied, such as Euclidean distance, cosine similarity, Jaccard coefficient, Pearson correlation coefficient. Given the diversity of similarity and distance measures available, their effectiveness in text document clustering is still not clear. The traditional dissimilarity or similarity measure and ours is that the former uses only a single viewpoint, which is the origin by using a specific measure. By using this measure less informative assessment of similarity could be achieved. We propose a Multiviewpoint-based Similarity measuring method, named MVS. MVS is potentially more suitable for text documents than the popular cosine similarity. The key contribution of this paper is the fundamental concept of similarity measure from multiple viewpoints. Multiview point similarity measure for document clustering which provides maximum efficiency and performance. Scope of this MVS is measure similarity and dissimilarity between objects which are present in different clusters. Two criterion functions for document clustering are proposed based on this new measure. Which are fast and scalable like k-means, but are also capable of providing high-quality and consistent performance? It provides best and more accuracy in results.

DOCUMENT CLUSTERING

Given a set S of n documents, we would like to partition them into a pre-determined number of k subsets S_1, S_2…. S_K, such that the documents assigned to each subset are more similar to each other than the documents assigned to different subsets. 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.
- Each document in a corpus corresponds to an m-dimensional vector d, where „m" is the total number of terms.
- Document vectors are often subjected to some weighting schemes, such as the standard Term Frequency-Inverse Document Frequency (TF-IDF), and normalized to have unit length.

A. Document Pre-Processing Steps:-

- Tokenization: A document is treated as a string (or bag of words), and then partitioned into a list of tokens.

- Removing stop words: Stop words are frequently occurring, insignificant words. This step eliminates the stop words.
- Stemming word: This step is the process of conflating tokens to their root form.

B. Document Representation:-

Generating N-distinct words from the corpora and call them as index terms (or the vocabulary). The document collection is then represented as a N-dimensional vector in term space. Computing Term weights Term Frequency. Inverse Document Frequency. Compute the TF-IDF weighting.

C. TFIDF Analysis:-

By taking into account these two factors: term frequency (TF) and inverse document frequency (IDF) it is possible to assign weights to search results and therefore ordering them statistically. Put another way a search result's score Ranking is the product of TF and IDF: TFIDF = TF * IDF where:

- ✓ TF = C / T where C = number of times a given word appears in a document and T = total number of words in a document.
- ✓ Document IDF = D / DF where D = total number of documents in a corpus, and DF = total number of documents containing a given word.

SIMILARITY MEASURE

A similarity/distance measure must be determined. The measure reflects the degree of closeness or separation of the target objects and should correspond to the characteristics that are believed to distinguish the clusters embedded in the data. In many cases, these characteristics are dependent on the data or the problem context at hand, and there is no measure that is universally best for all kinds of clustering problems. Moreover, choosing an appropriate similarity measure is also crucial for cluster analysis, especially for a particular type of clustering algorithms. For example, the density-based clustering algorithms, such as DBScan [11], rely heavily on the similarity computation. Density-based clustering finds clusters as dense areas in the data set, and the density of a given point is in turn estimated as the closeness of the corresponding data object to its neighboring objects. Recalling that closeness is quantified as the distance/similarity value, we can see that large number of distance/similarity computations are required for finding dense areas and estimate cluster assignment of new data objects. Therefore, understanding the effectiveness of different measures is of great importance in helping to choose the best one. The following are the different similarity measures

D. Cosine Similarity

Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgment of orientation and not magnitude: two vectors with the same orientation have a Cosine similarity of 1, two vectors at 90° have a similarity of 0, and two vectors diametrically opposed have a similarity of -1, independent of their magnitude. Cosine similarity is particularly used in positive space, where the outcome is neatly bounded in [0, 1].

$$\cos(\theta) = \frac{A \cdot B}{\|A\|\|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} (A_i)^2} \times \sqrt{\sum_{i=1}^{n} (B_i)^2}}$$

Where, A- term count value in the document1 and B- term count value in the document2. The value of the cosine similarity lies between 0-1.the value 0 represents the documents are not similar and 1 represents the documents are similar.

E. Jaccard coefficient

The Jaccard coefficient, also known as Tanimoto coefficient, measures similarity as the intersection divided by the union of the objects.

$$T(A,B) = \frac{A \cdot B}{\|A\|^2 + \|B\|^2 - A \cdot B}$$

Where, A- term count value in the document1 and B- term count value in the document2.

F. Pearson coefficient

$$r = \frac{\sum_i (x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_i (x_i - \overline{x})^2} \sqrt{\sum_i (y_i - \overline{y})^2}}$$

RELATED WORK

Martin Eisenhardt et al., [1] have presented an algorithm for the distributed clustering of documents. The authors have shown that algorithm is capable of handling real-world-sized problems and that distribution creates—even in their P2P setting—speed ups in comparison to centralized processing. For data sets such as text (many features, many documents with respect to the number of desired clusters), the savings in transfer time alone justify their distributed approach.

Souptik Datta et al., [2] have considered the problem of K-Means clustering on data distributed over a large, dynamic network, the data or the network itself may change. The authors assume the network to be peer-to-peer (Does not have any special servers). Centralizing all the data to a single machine to run a centralized K-Means is not an attractive option.

Michael Steinbach et al., [3] have compared the two main approaches to document clustering, agglomerative hierarchical clustering and K-means. Hierarchical clustering is often portrayed as the better quality clustering approach, but is limited because of its quadratic time complexity. In contrast, K-means and its variants have a time complexity which is linear in the number of documents, but are thought to produce inferior clusters. Sometimes K-means and agglomerative hierarchical approaches are combined so as to "get the best of both worlds." However, the results indicate that the bisecting K-means technique is better than the standard K-means approach and as good as or better than the hierarchical approaches that they tested for a variety of cluster evaluation metrics.

Geaorge Forman et al., [4] have illustrated in this paper that even for relatively small problem sizes, it can be more cost effective to cluster the data in place using an exact distributed algorithm to cluster the data in place using an exact distributed algorithm than to collect the data in one central location for clustering.

Souptik Datta et al., [5] have demonstrated an overview of distributed data mining applications and algorithms for peer-to-peer environments. It describes both exact and approximate distributed data mining algorithms that work in a decentralized manner. It illustrates these approaches for the problem of computing and monitoring clusters in the data residing at the different nodes of a peer-to-peer network.

Ion Stoica et al., [6] have presented Chord, a distributed lookup protocol that addresses this problem. Chord provides support for just one operation: given a key, it maps the key onto a node. Data location can be easily implemented on top of Chord by associating a key with each data item, and storing the key/data item pair at the node to which the key maps. Chord adapts efficiently as nodes join and leave the system, and can answer queries even if the system is continuously changing.

Karl Aberer et al., [7] have implemented P-Grid in Java and are currently in the final test phase. More information about P-Grid may be found on

the project's web page at http://www.pgrid. org. and It takes advantage of the resulting emergent properties for improving various services including routing, updates and identity management. One may also benefit from self-organizing principles when dealing with higher-level abstractions such as trust or global semantic interoperability

Antony Rowstron and Peter Druschel [8] have presented the design and evaluation of Pastry, a scalable, distributed object location and routing substrate for wide-area peer-to-peer applications. Pastry performs application-level routing and object location in a potentially very large overlay network of nodes connected via the Internet. It can be used to support a variety of peer-to-peer applications, including global data storage, data sharing, and group communication and naming.

Souptik Datta et al., [9] have offered an overview of distributed data mining applications and algorithms for peer-to-peer environments. It describes both exact and approximate distributed data mining algorithms that work in a decentralized manner. It illustrates these approaches for the problem of computing and monitoring clusters in the data residing at the different nodes of a peer-to-peer network.

CONCLUSION

In this paper we have summarized the research work that has done related to the clustering document clustering and similarity measure. Group of independent servers (usually in close proximity to one another) interconnected through a dedicated network to work as one centralized data processing resource. Clusters are capable of performing multiple complex instructions by distributing workload across all connected servers. Clustering improves the system's availability to users, its aggregate performance, and overall tolerance to faults and component failures.

REFERENCES

[1] M. Matteucci, "A Tutorial on Clustering Algorithms",*Available:http://home.dei.polimi.it/matteucc/Clustering/tutorial_html/, 2008*

[2] Y. Pen, G. Kou, Y. Shi, and Z. Chen, "Improving Clustering Analysis for Credit Card Accounts Classification," LNCS 3516, pp. 548-553, 2005

[3] A. Kalton, K. Wagstaff, and J. Yoo, "Generalized Clustering, Supervised Learning, and Data Assignment," *Proceedings of the Seventh International Conference on Knowledge Discovery and Data Mining, ACM Press, 2001.*

[4] M.R.ANDERBERG, "Cluster Analysis for Applications", *Academic Press, Inc., New York, NY, 1973*

[5] A. K. Jain, R.C.Dubes, , "Algorithms for Clustering Data", *Prentice-Hall advanced reference series. Prentice-Hall, Inc., Upper Saddle River, NJ, 1988*

[6] E.Diday, J.C.Simon, J. C, "Clustering analysis", *In Digital Pattern Recognition, K. S. Fu, Ed. Springer-Verlag, Secaucus, NJ, 47–94, 1976*

[7] R.Michalski, R.E.Stepp, E.Diday, "Automated construction of classifications: conceptual clustering versus numerical taxonomy", *IEEE Trans. Pattern Anal. Mach. Intell. PAMI-5, 5 (Sept.),pp. 396–409, 1983*

[8] R.C.Dubes, "How many clusters are best?—an experiment", *Pattern Recogn. 20, 6,pp. 645–663, 1987*

[9] C.H.Cheng, "A branch-and-bound clustering algorithm", *IEEE Trans. Syst. Man Cybern, Vol.25,pp. 895–898, 1995*

[10] R.C. Dubber, "Cluster analysis and related issues", *In Handbook of Pattern Recognition & Computer Vision, C. H. Chen, L. F.* Pau, and P. S. P. Wang, Eds. World Scientific Publishing Co., Inc., River Edge, NJ, 3–32, 1993

[11] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu. "A density-based algorithm for discovering clusters in large spatial databases with noise", In Proceedings of 2nd International Conference on KDD, 1996.

[12] M. Eisenhardt, W. Muller, and A. Henrich, "Classifying documents by distributed P2P clustering." in INFORMATIK, 2003

[13] S. Datta, C. R. Giannella, and H. Kargupta, "Kmeans Clustering over a Large, Dynamic Network," Proc. SIAM Int'l Conf. Data Mining (SDM), 2006

[14] M. Steinbach, G. Karypis, and V. Kumar, "A Comparison of Document Clustering Techniques," Proc. KDD Workshop Text Mining, 2000

[15] G. Forman and B. Zhang, "Distributed Data Clustering Can Be Efficient and Exact," SIGKDD Explorations Newsletter, vol. 2, no. 2, pp. 34-38, 2000

[16] S. Datta, K. Bhaduri, C. R. Giannella, R. Wolff and H. Kargupta, " Distributed data mining in Peer-to-Peer network's", IEEE Internet Computing, vol.10 , no. 4, pp. 18-26, July 2006

[17] I. Stoica, R. Morris, D. Karger, F. Kaashoek, and H. Balakrishnan, "Chord: A Scalable Peer-to-Peer Lookup Service for Internet Applications," Proc. SIGCOMM, 2001

[18] Aberer, Karl, et al. "P-Grid: a self-organizing structured P2P system." ACM SIGMOD Record 32.3, pp.29-33, 2003

[19] Rowstron, Antony, and Peter Druschel. "Pastry: Scalable, decentralized object location, and routing for large-scale peer-to-peer systems." Middleware 2001. Springer Berlin Heidelberg, 2001

[20] Datta, S., Bhaduri, K., Giannella, C., Wolff, R., & Kargupta, H., "Distributed data mining in peer-to-peer network's", Internet Computing, IEEE, Vol. 10(4), pp. 18-26, 2006