Volume 3, Issue 11, November 2013 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Clustering with Multi-view Point based Similarity Measure

Sobha Govathoti*, Aravind kumar.N, Prachi Junwale

Asst.Professor in CSE Dept
MLR Institute of Technology
JNTU Hyderabad , India

Abstract - 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. In this paper, 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.

Keywords— Data Mining, Clustering, Similarity Measure, Histograms, Parser.

I. 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 analysis, 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.

II. EXISTING SYSTEM

HFTC greedily picks the next frequent item set which represent the next cluster to minimize the overlapping between the documents that contain both the item set and some remaining item sets. Fig. 1 shows Web browser with many tabs open. In other words, the clustering result depends on the order of picking up the item sets, which in turns depends on the greedy heuristic. This method does not follow a sequential order of selecting clusters. Instead, we assign documents to the best cluster.

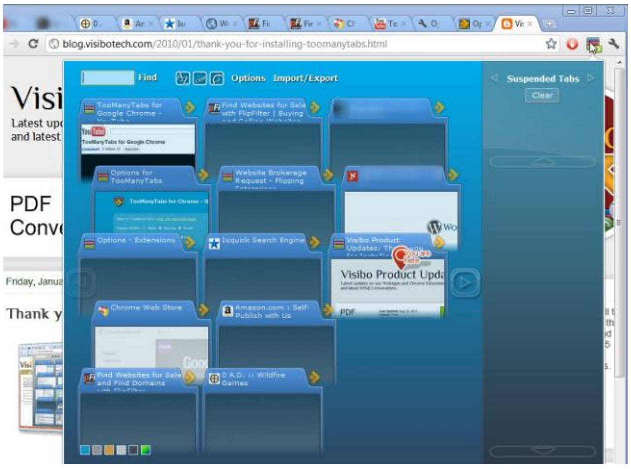
blog.visibotech.com/2010/01/thank-you-for-installing-toomanytabs.html
Find
Options
Import/Export
Suspended Tabs
Clear
TooManyTabs for Google Chrome
Find Websites for Sale with FlipFilter | Buying
Options for TooManyTabs
Website Brokerage Request - Flipping
Options - Extensions
Ixquick Search Engine
Visibo Product Updates
Visibo Product Upda
Chrome Web Store
Amazon.com : Self-Publish with Us
Find Websites for Sale and Find Domains
0 A.D. :: Wildfire Games
Visi
PDF
Conv
Friday, Janua
Thank y

Fig: 1 Web browser with many tabs open

III. PROPOSED SYSTEM

The main work is to develop a novel hierarchal algorithm for document clustering which provides maximum efficiency and performance. It is particularly focused in studying and making use of cluster overlapping phenomenon to design cluster merging criteria. Fig. 2 shows clustering of web pages into single window.

Proposing a new way to compute the overlap rate in order to improve time efficiency and "the veracity" is mainly concentrated. Based on the Hierarchical Clustering Method, the usage of Expectation-Maximization (EM) algorithm in the Gaussian Mixture Model to count the parameters and make the two sub-clusters combined when their overlap is the largest is narrated. Experiments in both public data and document clustering data show that this approach can improve the efficiency of clustering and save computing time.

Given a data set satisfying the distribution of a mixture of Gaussians, the degree of overlap between components affects the number of clusters "perceived" by a human operator or detected by a clustering algorithm. In other words, there may be a significant difference between intuitively defined clusters and the true clusters corresponding to the components in the mixture

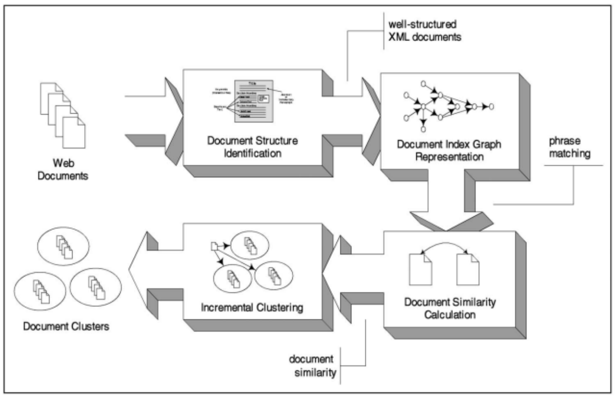
well-structured XML documents
Web Documents
Document Structure Identification
Document Index Graph Representation
phrase matching
Document Clusters
Incremental Clustering
Document Similarity Calculation
document similarity

Fig. 2.clustering of web pages into single window

IV. IMPLEMENTATION

Despite their popularity, most of the current unstructured P2P content distribution systems suffer from certain serious limitations. One such limitation is their simple, on demand mechanism for content discovery. Peers in these systems discover data items by circulating queries within the overlay network. A peer receiving a query responds back to the initiating node if it has any matching content. Upon processing a query, the recipient node removes it from its local buffers1. Thus, a query expires after it completes its circulation within the network. In other words, the network forgets the queries once they have completed their circulation. For clarity purposes, we call this the ad hoc query model, and we refer to the queries as ad hoc queries.

We present a scalable and effective middleware, called CoQUOS, for supporting continuous queries in unstructured overlay networks. Besides being independent of the overlay topology, Co QUOS preserves the simplicity and flexibility of the unstructured P2P network. Our design of the Co QUOS system is characterized by two novel techniques, namely cluster-resilient random walk algorithm for propagating the queries to various regions of the network and dynamic probability-based query registration scheme to ensure that the registrations are well distributed in the overlay. Further, we also develop effective and efficient schemes for providing resilience to the churn of the P2P network and for ensuring a fair distribution of the notification load among the peers. This paper studies the properties of our algorithms through theoretical analysis. We also report series of experiments evaluating the effectiveness and the costs of the proposed schemes.

DESIGN LAYOUT

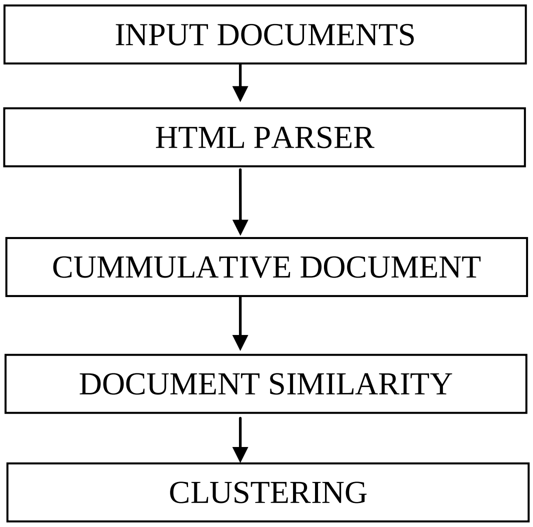

Fig. 3. Design Layout

(1) Initializing the weights parameters.
(2) Using the EM algorithm to estimate their means and covariance.
(3) Grouping the data to classes by the value of probability density to each class and calculating the weight of each class.
(4) Repeat the first step until the cluster number reaches the desired number or the largest OLR is smaller than the predefined threshold value. Go to step 3 and output the result. A distinctive element in this algorithm is to use the overlap rate to measure similarity between clusters.

Implementation is the stage of the project when the theoretical design is turned out into a working system. Thus it can be considered to be the most critical stage in achieving a successful new system and in giving the user, confidence that the new system will work and be effective.

The implementation stage involves careful planning, investigation of the existing system and it's constraints on implementation, designing of methods to achieve changeover and evaluation of changeover methods as shown in Fig. 3.

HTML Parser

- Parsing is the first step done when the document enters the process state.
- Parsing is defined as the separation or identification of meta tags in a HTML document.
- Here, the raw HTML file is read and it is parsed through all the nodes in the tree structure.

Cumulative Document

- The cumulative document is the sum of all the documents, containing meta-tags from all the documents.
- We find the references (to other pages) in the input base document and read other documents and then find references in them and so on.
- Thus in all the documents their meta-tags are identified, starting from the base document.

Document Similarity

- The similarity between two documents is found by the cosine-similarity measure technique.
- The weights in the cosine-similarity are found from the TF-IDF measure between the phrases (meta-tags) of the two documents.
- This is done by computing the term weights involved.
- TF = C / T
- IDF = D / DF.

  D → quotient of the total number of documents
  DF → number of times each word is found in the entire corpus
  C → quotient of no of times a word appears in each document
  T → total number of words in the document
- **TFIDF = TF * IDF**

Clustering

- Clustering is a division of data into groups of similar objects.
- Representing the data by fewer clusters necessarily loses certain fine details, but achieves simplification.

The similar documents are grouped together in a cluster, if their cosine similarity measure is less than a specified threshold

V. RESULTS

The clustering approach proposed here is an incremental dynamic method of building the clusters. An overlapped cluster model is adopted here. The key concept for the similarity histogram-based clustering method is to keep each cluster at a high degree of coherency at any time .Representation of the coherency of a cluster is called as Cluster Similarity Histogram.

Fig. 4 shows Cluster Similarity Histogram which is a concise statistical representation of the set of pair-wise document similarities distribution in the cluster. A number of bins in the histogram correspond to fixed similarity value intervals. Each bin contains the count of pair-wise document similarities in the corresponding interval. The below graph shows a typical cluster similarity histogram, where the distribution is almost a normal distribution. A perfect cluster would have a histogram where the similarities are all maximum, while a loose cluster would have a histogram where the similarities are all minimum.

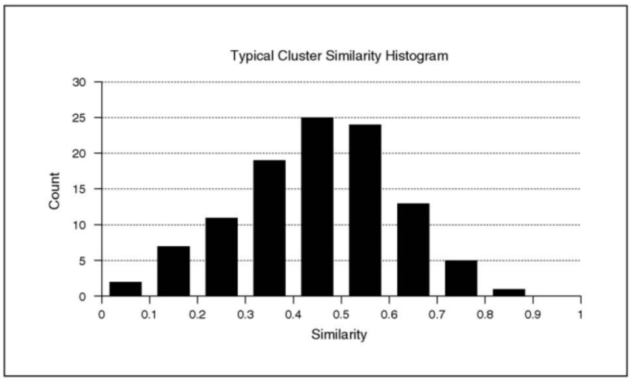

Fig. 4. Performance analysis

VI. CONCLUSIONS

Given a data set, the ideal scenario would be to have a given set of criteria to choose a proper clustering algorithm to apply. Choosing a clustering algorithm, however, can be a difficult task. Even ending just the most relevant approaches for a given data set is hard. Most of the algorithms generally assume some implicit structure in the data set. One of the most important elements is the nature of the data and the nature of the desired cluster. Another issue to keep in mind is the kind of input and tools that the algorithm requires. This report has a proposal of a new hierarchical clustering algorithm based on the overlap rate for cluster merging. The experience in general data sets and a document set indicates that the new method can decrease the time cost, reduce the space complexity and improve the accuracy of clustering. Specially, in the document clustering, the newly proposed algorithm measuring result show great advantages. The hierarchical document clustering algorithm provides a natural way of distinguishing clusters and implementing the basic requirement of clustering as high within-cluster similarity and between-cluster dissimilarity.

FUTURE WORKS

In the proposed model, selecting different dimensional space and frequency levels leads to different accuracy rate in the clustering results. How to extract the features reasonably will be investigated in the future work. There are a number of future research directions to extend and improve this work. One direction that this work might continue on is to improve on the accuracy of similarity calculation between documents by employing different similarity calculation strategies. Although the current scheme proved more accurate than traditional methods, there are still rooms for improvement.

ACKNOWLEDGEMENT

The authors express their deep gratitude to the Principal and the Management members of MLRIT for their encouragement and extensive support in preparing and publishing of this paper.

REFERENCES

[1] A. J. Cole & D. Wishart (1970), "An improved algorithm for the Jardine-Sibson method of generating overlapping clusters", *The Computer Journal* 13(2):156-163.

[2] D'andrade,R. 1978, "U-Statistic Hierarchical Clustering" *Psychometrika,* 4:58-67.

[3] S. C. Johnson. 1967, "Hierarchical Clustering Schemes" *Psychometrika,* 2:241-254.

[4] Shengrui Wang and Haojun Sun, "Measuring overlap-Rate for Cluster Merging in a Hierarchical Approach to Color Image Segmentation", *International Journal of Fuzzy Systems,*Vol.6,No.3,September 2004.

[5] Jeff A. Bilmes, " A Gentle Tutorial of the EM Algorithm and its Application to Parameter Estimation for Gaussian Mixture and Hidden Markov Models", *ICSI* TR-97-021, U.C. Berkeley, 1998

[6] E.M. Voorhees," Implementing agglomerative hierarchical clustering algorithms for use in document retrieval", *Information Processing and Management,* 22(6):465–476, 1986.

[7] Sun Da-fei,Chen Guo-li,Liu Wen-ju," The discussion of maximum likehood parameter estimation based on EM algorithm", *Journal of HeNan University*. 2002,32(4):35~41

[8] Khaled M. Hammouda, Mohamed S. Kamel , “efficient phrase-based document indexing for web document clustering” *, IEEE transactions on knowledge and data engineering,* October 2004

[9] Haojun sun, zhihui liu, lingjun kong, “A Document Clustering Method Based On Hierarchical Algorithm With Model Clustering”, *22nd international conference on advanced information networking and applications,*

[10] Shi zhong, joydeep ghosh, “Generative Model-Based Document Clustering: A Comparative Study”, *The University Of Texas.*