ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

Improving the efficiency of search of large amounts of data using Suffix Tree Clustering

S.Sreedevi[1], K.Satyanarayana Murthy[2]

1 Final M.Tech(CST) Student ,Dept of Computer Science and Engineering, Baba Institute of Technology and Sciences (JNTUK) Visakhapatnam Dist ,Andhra pradesh, India.

2 Associate professor, Dept of Computer Science and Engineering, Baba Institute of Technology and Sciences (JNTUK) Visakhapatnam Dist ,Andhra pradesh, India.

Abstract: Data Mining as a process of finding new, useful knowledge from data using different techniques. Using these techniques we getting faster and better search of large amounts of data that we facing every day. Clustering of data is one of the techniques that are used in data mining. In this paper we explore clustering algorithms and take suffix tree clustering algorithm for the best of them. We Implement an application that use this algorithm in the process of clustering, and search of clustered documents. 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. Suffix tree document model and Suffix Tree Clustering (STC) algorithm first were analysised and implemented. STC is a linear time clustering algorithm, which is based on identifying phrases that are common to groups of documents. A phrase is an ordered sequence of one or more words. STC algorithm is different from the other kind of clustering algorithms. It is a data structure which contains all the suffixes of a given string, so as to run many important string operations more efficiently. This algorithm not treats documents as a collection of words but as a string of words.

Keywords: Suffix Tree Clustering (STC), Data Mining, Document search, Suffix Tree.

Introduction:

Document retrieval systems typically present search results in a ranked list, ordered by their estimated relevance to the query. The similarity based relevancy is estimated between the text of a document and the query. Such ranking schemes work well when users can formulate a well-defined query for their searches. The users of search engines often formulate very short queries that often retrieve large numbers of documents. Based on such a condensed representation of the user's search interest, it is impossible for the search engine to identify the specific documents that are of interest to the users. Moreover, many webmasters now actively work to influence rankings. These problems are exacerbated when the users are unfamiliar with the topic they are querying about, when they are novices at performing searches, or when the search engine's database contains a large number of documents. These conditions commonly exist for Web search engine users. Therefore the vast majority of the retrieved documents are often of no interest to the user, such searches are termed as low precision searches. The low precision of the search engines coupled with the ranked list presentation force users to sift through a large number of documents and make it hard for them to find the information they are looking for. As low

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

precision searches are inevitable, tools must be provided to help users "cope" with (and make use of) these large document sets.

Clustering and Its Models

Cluster analysis or clustering is the task of assigning a set of objects into groups (called clusters) so that the objects in the same cluster are more similar (in some sense or another) to each other than to those in other clusters. The main task of clustering is explorative data mining, and the common technique for statistical data analysis used in many fields, like including machine learning, pattern recognition, image analysis, information retrieval, and bioinformatics. The notion of a "cluster" varies between algorithms and is one of the many decisions to take when choosing the appropriate algorithm for a particular problem. The terminology of a cluster seems obvious: a group of data objects. The clusters found by different algorithms vary significantly in their properties, and understanding the cluster models is key to understanding the differences between the various algorithms.

Typical cluster models include:

Connectivity models: for example hierarchical clustering builds models based on distance connectivity.

Centroid models: for example the k-means algorithm represents each cluster by a single mean vector.

Distribution models: clusters are modeled using statistical distributions, such as multivariate normal distributions used by the Expectation-maximization algorithm.

Density models: for example DBSCAN and OPTICS defines clusters as connected dense regions in the data space.

Subspace models: in Biclustering (also known as Coclustering or two-mode-clustering), clusters are modeled with both cluster members and relevant attributes.

Group models: some algorithms (unfortunately) do not provide a refined model for their results and just provide the grouping information.

Graph-based models: a clique, i.e., a subset of nodes in a graph such that every two nodes in the subset are connected by an edge can be considered as a prototypical form of cluster. Relaxations of the complete connectivity requirement (a fraction of the edges can be missing) are known as quasicliques[4-6].

A "clustering" is essentially a set of such clusters, usually containing all objects in the data set. Additionally, it may specify the relationship of the clusters to each other, for example a hierarchy of clusters embedded in each other.

Clustering can be roughly distinguished in:

Hard clustering: each object belongs to a cluster or not Soft clustering (also: fuzzy clustering): each object belongs to each cluster to a certain degree (e.g. a likelihood of belonging to the cluster)

There are also finer distinctions possible, for example:

Strict partitioning clustering: here each object belongs to exactly one cluster

Strict partitioning clustering with outliers: objects can also belong to no cluster, and are considered outliers.

Overlapping clustering (also: alternative clustering, multiview clustering): while usually a hard clustering, objects may belong to more than one cluster.

Hierarchical clustering: objects that belong to a child cluster also belong to the parent cluster

Subspace clustering: while an overlapping clustering, within a uniquely defined subspace, clusters are not expected to overlap [7-8].

Document Clustering

Document clustering (or Text clustering) is automatic document organization, topic extraction and fast information retrieval or filtering. It is closely related to data clustering. A web search engine often returns thousands of pages in response to a broad query, making it difficult for users to browse or to identify relevant information. Clustering methods can be used to automatically group the retrieved documents into a list of meaningful categories, as is achieved by Enterprise Search engines. Document clustering

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

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 link, 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 Kmeans algorithm and its variants. Usually, it is of greater efficiency, but less accurate than the hierarchical algorithm [9] and [10].

Related Work:

Lan Yu et. al. In this paper, we use one of the data mining algorithms (clustering) to analyze PHS test data with the help of SQL Server. The purposes of experiments are to discover how to construct a training set and how to set parameters of Microsoft clustering algorithm. Some valuable conclusions are achieved. A good training set should have sufficient records, but a huge number of records could degrade the performance of clustering. That is the reason why the accuracy of case 2 is better than that of case 1. And using all data to compose a training set as case 1 of this experiment is not practical. When the number of the training set is determinate, the more records a ranking has, the higher accuracy of this ranking obtained [2].

M. Asif Naeem et. al. proposed a novel clustering algorithm called Hierarchical Particle Swarm Optimization (HPSO) data clustering. The proposed algorithm exploits the swarm intelligence of cooperating agents in a decentralized environment. The experimental results were compared with benchmark clustering techniques, which include K-means, PSO clustering, Hierarchical Agglomerative clustering (HAC) and Density-Based Spatial Clustering of Applications with Noise (DBSCAN). The results are evidence of the effectiveness of Swarm based clustering and the capability to perform clustering in a hierarchical agglomerative manner. The approach to tackling the problem of hierarchical agglomerative clustering through hierarchical particle swarm optimization based clustering is inspired by the collective intelligent behavior of swarms. HPSO-clustering has the properties of both partitioned based data clustering and hierarchical data clustering. Experimental results verify the performance of PSO against PSO, traditional hierarchical agglomerative clustering and K-means clustering.

Yun Ling et. al. utilized correspondence analysis algorithm to process matrix decomposition and then make use of Bayesian approach for co-clustering. They find that utilizing the two methods synthetically is very significative to solve actual problems. Experiments on synthetic and real world data demonstrate the efficiency and effectiveness of the algorithm. They proposed a CA method for fast co-clustering on large data. The learned knowledge is useful when synthesizing these two methods. Because of the small size of the approximation matrices, it has runtime complexity equal to orders of magnitude faster than the runtime complexity of the previous co-clustering algorithms. Due to its low complexity and simple implementation the work presented will make a broad application [3].

M. Sudharshan et. al. propose an innovative clustering technique called the Rapid Clustering Method (RCM), which uses Subtractive Clustering combined with Fuzzy CMeans clustering along with a histogram sampling technique to provide quick and effective results for large sized datasets. Rapid Clustering Method can be used to cluster the dataset and analyze the characteristics in a social network. It can also be used to enhance the cross-selling practices using quantitative association rule mining. With the help of

ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

RCM they can obtain rapid results and identify structural changes much more quickly. They considered the problem of clustering data over time and proposed a rapid clustering technique so that it can be used to generate quicker results for social network analysis [4].

Existing System

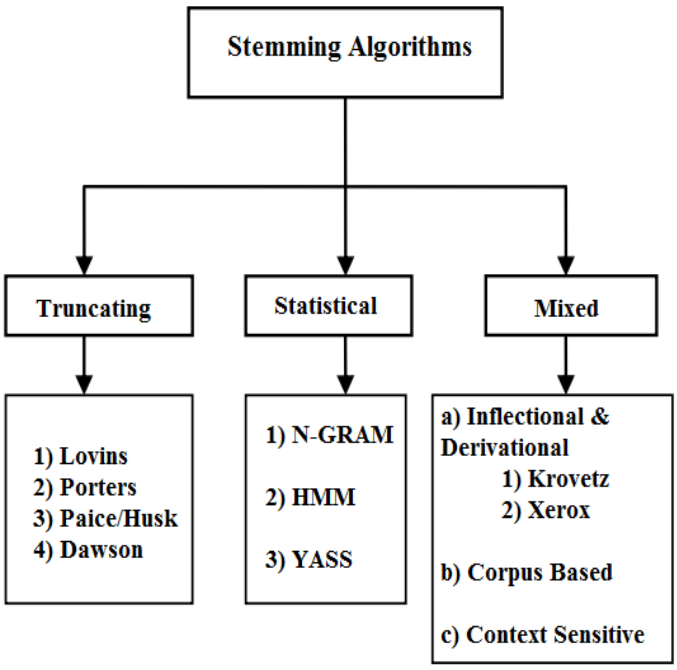
There is no shortage of clustering algorithms, and recently a new wave of excitement has spread across the machine learning community mainly because of the important development of spectral methods. At the same time, there is also growing interest around fundamental questions pertaining to the very nature of the clustering problem. Yet, despite the tremendous progress in the field, the clustering problem remains elusive and a satisfactory answer even to the most basic questions is still to come.

Upon scrutinizing the relevant literature on the subject, it becomes apparent that the vast majority of the existing approaches deal with a very specific version of the problem, which asks for partitioning the input data into coherent classes. In fact, almost invariably, the problem of clustering is defined as a partitioning problem, and even the classical distinction between hierarchical and partitional algorithms seems to suggest the idea that partitioning data is, in essence, what clustering is all about (as hierarchies are but nested partitions). This is unfortunate, because it has drawn the community's attention away from different, and more general, variants of the problem and has led people to neglect underdeveloped foundational issues. As J. Hartigan clearly put it more than a decade ago: "We pay too much attention to the details of algorithms. We must begin to subordinate engineering to philosophy.

Proposed System

We propose Suffix tree document model and Suffix Tree Clustering (STC) algorithm first were proposed and use in. STC is a linear time clustering algorithm (linear in the size of the document set), which is based on identifying phrases that are common to groups of documents. A phrase is an ordered sequence of one or more words. STC algorithm is different from the other kind of clustering algorithms. It is a data structure which contains all the suffixes of a given string, so as to run many important string operations more efficiently. This algorithm not treats documents as a collection of words but as a string of words. On that way thus operates using the proximity information between words. STC use suffix tree structure to efficiently identify sets of documents that share common phrases and terms, and uses this information to create clusters and to concisely present their contents to the users. STC meanly includes four logical steps: first, document "cleaning"; secondly, constructing a generalized suffix tree; thirdly, identifying base clusters; the last step is to combine these base clusters into clusters.

In the next step words stemmers are using. A stemming algorithm is a process of linguistic normalization, in which the variant forms of a word are reduced to a 16 common form. Stemming algorithms can be classified in three groups: truncating methods, statistical methods, and mixed methods. Each of these groups has a typical way of finding the stems of the word variants.



ISSN No: 2348-4845
International Journal & Magazine of Engineering, Technology, Management and Research
A Peer Reviewed Open Access International Journal

- **Cleaning**
- **Constructing generalized suffix tree**
- **Identifying Base Clusters**
- **Combining base clusters into clusters**

Advantages

- Suffix tree clustering is one of the most important algorithm that is in use in the process of clustering.
- This algorithm have linear complexity O(n). Linear complexity puts this algorithm in the top of clustering algorithms.
- Response time in the process of clustering is minimal compared with other clustering algorithms.
- This low complexity return fast search in suffix tree.

Because of that reasons this algorithm is used in online clustering, and in web document clustering. In the process of searching the first ten results shall be taken as the best. Application created by authors tests other side of suffix tree clustering too.

Conclusion:

Document clustering has initially been investigated in Information Retrieval mainly as a means of improving the performance of search engines by pre-clustering the entire corpus. The cluster hypopaper stated that similar documents will tend to be relevant to the same queries, thus the automatic detection of clusters of similar documents can improve recall by effectively broadening a search request. In our paper implement a Suffix Tree Document Clustering. Suffix tree clustering is one of the most important algorithm that is in use in the process of clustering. Linear complexity puts this algorithm in the top of clustering algorithms. Response time in the process of clustering is minimal compared with other clustering algorithms. This low complexity return fast search in suffix tree. Because of that reasons this algorithm is used in online clustering, and in web document clustering. Our proposed Suffix Tree Document Clustering mechanism have four steps, these are (1) Cleaning (2) Constructing generalized suffix tree (3) Identifying Base Clusters (4)Combining base clusters into clusters. We conclude that our proposed technique is better perform based on document clustering than k-means clustering. Test result show that suffix tree clustering is good algorithm for all data types, and all documents types.

References:

[1] Milos Ilic, Petar Spalevic & Mladen Veinovic, Suffix Tree Clustering - Data mining algorithm, ERK'2014, Portorož, B:15-18

[2] David Pettinger and Giuseppe Di Fatta, "Scalability of Efficient Parallel K-Means", IEEE e-Science 2009 Workshops, pp. 96-101.

[3] Yun Ling and Hangzhou, "Fast Co-clustering Using Matrix Decomposition", IEEE 2009 Asia-Pacific Conference on Information Processing, pp. 201-204.

[4] J. Prabhu and M. Sudharshan and M. Saravanan and G.Prasad, "Augmenting Rapid Clustering Method for Social Network Analysis", 2010 International Conference on Advances in Social Networks Analysis and Mining, pp. 407-408.

[5] F. Yang, T. Sun, C. Zhang, An efficient hybrid data clustering method based on K-harmonic means, and Particle Swarm Optimization, Expert Systems with Applications 2009, pp. 9847–9852.

[6] Y.-T. Kao, E. Zahara, I.-W. Kao, A hybridized approach to data clustering, Expert Systems with Applications 2008, pp. 1754-1762.

[7] Madjid Khalilian, Farsad Zamani Boroujeni, Norwati Mustapha, Md. Nasir Sulaiman, "K-Means Divide and Conquer Clustering", IEEE 2009, International Conference on Computer and Automation Engineering, pp. 306-309.

[8] Lan Yu, "Applying Clustering to Data Analysis of Physical Healthy Standard", 2010 Seventh International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2010), pp. 2766-2768.

[9] Vignesh T. Ravi and Gagan Agrawal, "Performance Issues in Parallelizing Data-Intensive Applications on a Multi-core Cluster", 9th IEEE/ACM

International Symposium on Cluster Computing and the Grid, pp. 308-315.
[10] M.Mounika and Dr. Sudarson Jena , Improving Effectiveness by Clustering of BioMedical Documents Using Semi supervised Clustering Method, IJMETMR, Volume No: 2 (2015), Issue No: 5 (May) http://www.ijmetmr.com/olmay2015/MMounika-DrSudarsonJena-8.pdf

Authors:

S.SREEDEVI was born in Visakhapatnam, Andhra Pradesh, India. She received B.Tech in C.S.E from JNTU University, Hyderabad, and Andhra Pradesh, India. After having 6 years she did different private jobs in NGO Management and HR Consultant. She presently Pursuing M.Tech (CST) department of Computer Science Engineering from Baba Institute of Technology and Sciences, Visakhapatnam. Her Research interest in Data Mining and Knowledge Discovery.

K.SATYANARAYANA MURTHY is an Associate Professor in the Department of Computer Science and Engineering in Baba Institute of Technology and Sciences. He completed his M.Tech in Computer Science in Gayatri Vidya Parishad College of Engineering, Visakhapatnam. He has 6 yrs Experience in teaching Computer Science and Engineering related Subjects. He can be reached at murthy8542.mtech@gmail.com