International Journal of Contemporary Research in Computer Science and Technology (IJCRCST) e-ISSN: 2395-5325
Volume2, Issue 7 (July '2016)

THE ANALYSIS AND REVIEW FOR HIERARCHICAL AGGLOMERATIVE CLUSTERING TECHNIQUES IN VARIOUS WSN APPLICATIONS

S.Aravindhan,
Research Scholar,
PG & Research Department of Computer Science,
Sri Vijay Vidyalaya College of Arts & Science,
Dharmapuri,Tamilnadu,India.

Dr.D.Maruthanayagam,
Assistant Professor,
PG & Research Department of Computer Science,
Sri Vijay Vidyalaya College of Arts & Science,
Dharmapuri,Tamilnadu,India.

Abstract: Clustering is the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but these objects are very dissimilar to the objects that are in other clusters. Clustering methods are mainly divided into two groups: hierarchical and partitioning methods. Agglomerative group will be a "bottom up" approach and each observation start in its own cluster, and pairs of clusters are merged as one move up the hierarchy. Divisive group will be a "top down" approach and all observations start in one cluster, and splits are performed recursively as one moves down the hierarchy. In this paper, analysis and review about hierarchical clustering which is mainly focus on hierarchical agglomerative clustering (HAC) technique. Then we also presents detail description of some hierarchical agglomerative clustering technique related previous research papers for understanding the importance of hierarchical clustering concepts in various research areas.

Keywords: Clustering, Agglomerative, Divisive, WSN and Hierarchical Clustering

I. INTRODUCTION

Data mining allows us to extract knowledge from our historical data and predict outcomes of our future situations. Clustering is an important data mining task. It can be described as the process of organizing objects into groups whose members are similar in some way. Clustering can also be define as the process of grouping the data into classes or clusters, so that objects within a cluster have high similarity in comparison to one another but are very dissimilar to objects in other clusters. Mainly clustering can be done by two methods: Hierarchical and Partitioning method [1]. In data mining hierarchical clustering works by grouping data objects into a tree of cluster. Hierarchical clustering methods can be further classified into **agglomerative** and **divisive** hierarchical clustering. This classification depends on whether the hierarchical decomposition is formed in a bottom-up or top-down fashion. Hierarchical techniques produce a nested sequence of partitions, with a single, all inclusive cluster at the top and singleton clusters of individual objects at the bottom. Each intermediate level can be viewed as combining two clusters from the next lower level or splitting a cluster from the next higher level. The result of a hierarchical clustering algorithm can be graphically displayed as tree, called a dendrogram. This tree graphically displays the merging process and the intermediate clusters. This graphical structure shows how points can be merged into a single cluster. Hierarchical methods suffer from the fact that once we have performed either merge or split step, it can never be undone. This inflexibility is useful in that it leads to smaller computation costs by not having to worry about a combinatorial number of different choices. However, such techniques cannot correct mistaken decisions that once have taken. There are two approaches that can help in improving the quality of hierarchical clustering: (1) Firstly to perform careful analysis of object linkages at each hierarchical partitioning or (2) By integrating hierarchical agglomeration and other approaches by first using a hierarchical agglomerative algorithm to group objects into micro-clusters, and then performing macro-clustering on the micro clusters using another clustering method such as iterative relocation [2].

There are two basic approaches to generating a hierarchical clustering:

- **Agglomerative** Start with the points as individual clusters and, at each step, merges the closest pair of clusters. This requires defining the notion of cluster proximity. Agglomerative techniques are most popular, and most of this section will be spent describing them.
- **Divisive** Start with one, all-inclusive cluster and, at each step, split a cluster until only singleton clusters of individual points remain. In this case, we need to decide which cluster to split at each step and how to do the splitting.

Agglomerative Clustering: One of the earliest and most widely used clustering strategies is agglomerative clustering. The history of agglomerative clustering goes back at least to the 1950s (see for example [3, 4]). Later, biological taxonomy became one of the driving forces of cluster analysis. In [5] the authors, who were the first biologists using computers to classify organisms, discuss several agglomerative clustering methods. Agglomerative clustering is a bottom-up clustering process. At the beginning, every input object forms its own cluster. In each subsequent step, the two 'closest' clusters will be merged until only one cluster remains. This clustering process creates a hierarchy of

clusters, such that for any two different clusters A and B from possibly different levels of the hierarchy we either have $A \cap B = \emptyset$, $A \subset B$, or $B \subset A$. Such a hierarchy is useful in many applications, for example, when one is interested in hereditary properties of the clusters (as in some bioinformatics applications) or if the exact number of clusters is a priori unknown. In order to define the agglomerative strategy properly, we have to specify a distance measure between clusters. Given a distance function between data objects, the following distance measures between clusters are frequently used. In the single linkage strategy, the distance between two clusters is defined as the distance between their closest pair of data objects. It is not hard to see that this strategy is equivalent to computing a minimum spanning tree of the graph induced by the distance function using Kruskal's algorithm. In case of the complete linkage strategy, the distance between two clusters is defined as the distance between their furthest pair of data objects. In the average linkage strategy the distance is defined as the average distance between data objects from the two clusters.

Algorithm: Agglomerative hierarchical clustering is a clustering algorithm that builds a cluster hierarchy from the bottom-up. It starts by adding a cluster for each of the data points to be clustered, followed by iterative pair-wise merging of clusters until only one cluster is left at the top of the hierarchy. The choice of clusters to merge at the each iteration is decided based on a distance metric.
The agglomerative hierarchical clustering algorithm is implemented as follows:

1. Based on the linkage criteria, set the distance function.
2. Parse the input file and read in the items to cluster.
3. Initialize the cluster array and set the appropriate item counts.
4. Calculate the Euclidean distances between item pairs.
5. For each item in the input
    1. Add a leaf node that represents a root cluster.
    2. Update the distance from root cluster to its neighbours.
6. While there are more than one cluster.
    1. Find a pair of clusters with the minimum distance.
    2. Merge them into a root cluster.
    3. Update the distance from root cluster to its neighbours.

II. MATERIALS AND METHODS

Karuna Katariya, *et al.* [8] Web Usage Mining used to extract knowledge from WWW. Nowadays interaction of user towards web data is growing, web usage mining is significant in effective website management, adaptive website creation, support services, personalization, and network traffic flow analysis and user trend analysis and user's profile also helps to promote website in ranking. Agglomerative clustering is a most flexible method and it is also used for clustering the web data in web usage mining, there are do not need the number of clusters as an input. Agglomerative have many drawbacks such as initial error propagation, dimensionality, complexity and data set size issues. In this paper introduced solution for data set size problem that helpful for information retrieve from large web data, web log data files are as an input for agglomerative clustering algorithms and output is efficient clustering that will be used further for information extraction in web usage mining.

Subaira.A.S, *et al.* [9] in information system, security has remained one hard line area for computers as well as networks. In information protection, Intrusion Detection System (IDS) is used to safeguard the data confidentiality, integrity and system availability from various types of attacks. Data mining is an efficient artifice applied to intrusion detection to ascertain a new outline from the massive network data as well as it used to reduce the strain of the manual compilations of the normal and abnormal behavior patterns. This piece of writing reviews the present state of data mining clustering techniques to implement an intrusion detection system such as, Partitioning methods, Hierarchical methods, Model based clustering methods and their various types.

Fathi H. Saad, *et al*. [10] Extensive amount of data stored in medical documents require developing methods that help users to find what they are looking for effectively by organizing large amounts of information into a small number of meaningful clusters. The produced clusters contain groups of objects which are more similar to each other than to the members of any other group. Thus, the aim of high-quality document clustering algorithms is to determine a set of clusters in which the inter-cluster similarity is minimized and intra-cluster similarity is maximized. The most important feature in many clustering algorithms is treating the clustering problem as an optimization process, that is, maximizing or minimizing a particular clustering criterion function defined over the whole clustering solution. The only real difference between agglomerative algorithms is how they choose which clusters to merge. The main purpose of this paper is to compare different agglomerative algorithms based on the evaluation of the clusters quality produced by different hierarchical agglomerative clustering algorithms using different criterion functions for the problem of clustering medical documents. The experimental results showed that the agglomerative algorithm that uses I1 as its criterion function for choosing which clusters to merge produced better clusters quality than the other criterion functions in term of entropy and purity as external measures.

Jia Li, *et al.* [11] consider the problem of clustering under the constraint that data points in the same cluster are connected according to a pre-existed graph. This constraint can be efficiently addressed by an agglomerative clustering approach, which exploits to construct a new fully automatic segmentation algorithm for color photographs. For image segmentation, if the pixel grid with eight neighbor connectivity is imposed as the graph, each group of pixels generated by this clustering method is ensured to be a geometrically connected region in the image, a desirable trait for many subsequent operations. To achieve scalability for images with large sizes, the segmentation algorithm combines the top-down k-means clustering with the bottom-

up agglomerative clustering method. Also find that it is advantageous to conduct clustering at multiple stages through which the similarity measure is adjusted. Experimental results with comparison to other widely used and state-of-the-art segmentation methods show that the new algorithm achieves higher accuracy at much faster speed. A software package is provided for public access.

Doug Beeferman, *et al.* [12] this paper introduces a technique for mining a collection of user transactions with an internet search engine to discover clusters of similar queries and similar URLs. The information exploit is "click through data": each record consists of a user's query to a search engine along with the URL which the user selected from among the dataset as a bipartite graph, with the vertices on one side corresponding to queries and on the other side to URLs. One can apply an agglomerative clustering algorithm to the graph's vertices to identify related queries and URLs. One noteworthy feature of the proposed algorithm is that it is "content-ignorant"-the algorithm makes no use of the actual content of the queries or URLs, but only how they co-occur within the click through data. This paper describes how to enlist the discovered clusters to assist users in web search, and measure the effectiveness of the discovered clusters in the Lycos search engine.

Mohammad Saiful Islam Mamun, *et al.* [13] in this paper proposed a hierarchical architectural design based intrusion detection system that fits the current demands and restrictions of wireless ad hoc sensor network. In this proposed intrusion detection system architecture followed clustering mechanism to build a four level hierarchical network which enhances network scalability to large geographical area and use both anomaly and misuse detection techniques for intrusion detection. This paper, introduce policy based detection mechanism as well as intrusion response together with GSM cell concept for intrusion detection architecture.

Wang Lei, *et al.* [14] A new distributed algorithm of data compression based on hierarchical cluster model for sensor networks is proposed, the basic ideas of which are as follows, firstly the whole sensor network is mapped into a kind of hierarchical clusters model, and then different wavelet transform models are used to commit data compression in inner and super clusters respectively, according to the relative regularity of sensor nodes deployed in the inner clusters, and the relative irregularity of sensor nodes deployed in super cluster. Theoretical analyses and simulation results show that, the above new methods have good performance of approximation, and can compress data and reduce the amount of data efficiently. So, it can prolong the lifetime of the whole sensor network to a greater degree.

Jae-Yoon Jung, *et al.* [15] Business process is collection of standardized and structured tasks inducing value creation of a company. Nowadays, it is recognized as one of significant intangible business assets to achieve competitive advantages. This paper, introduce a novel approach to business process analysis, which has more and more significance as process aware in-formation systems that are spreading widely over a lot of companies. In this paper, a methodology of business process clustering based on process similarity is proposed. The purpose of business process clustering is to analyze accumulated process models in order to assist new process design or process reengineering. The proposed methodology exploits structural similarity metrics of business processes. This work also illustrated the methodology with example processes inducing the hierarchical merged models from the process clusters.

Chung-Horng Lung, *et al.* [16] Wireless Sensor Networks (WSNs) have a wide range of applications that base on the collaborative effort of a number of sensor nodes. Cluster-based network architecture can enhance network self-control capability and resource efficiency, and prolong the whole network lifetime. Thus, finding an effective and efficient way to generate clusters is an important topic in WSNs. existing clustering approaches may not be flexible enough to cope with various factors or have higher communication overhead. To achieve the goal, tailor the HAC (Hierarchical Agglomerative Clustering) algorithm for WSNs. HAC is a well-known approach and has been successfully applied to many disciplines. HAC uses simple numerical methods to make clustering decisions. In addition, HAC provides flexibility with respect to input data type (e.g., location data or connectivity information) and weight assignment to different factors (e.g., connections or power strength).

Sadaaki Miyamoto, *et al.* [17] Although data clustering is relatively uninvestigated in rough set studies, there are much room for applying clustering and related techniques to this field. In this paper mainly focus on generalization of agglomerative clustering to information systems. A poset-valued hierarchical clustering is defined and the combination of traditional agglomerative clustering and lattice diagram of attributes in an information system is considered. Inner product spaces are available to information systems by using kernel functions in support vector machines. Different algorithms for generalized agglomerative clustering using the inner product are described.

Latifur Khan, *et al.* [18] this paper use Support Vector Machines (SVM) for classification. The SVM is one of the most successful classification algorithms in the data mining area, but it's long training time limits its use. This paper presents a study for enhancing the training time of SVM, specifically when dealing with large data sets, using hierarchical clustering analysis. Use the Dynamically Growing Self-Organizing Tree (DGSOT) algorithm for clustering because it has proved to overcome the drawbacks of traditional hierarchical clustering algorithms (e.g., hierarchical agglomerative clustering). Clustering analysis helps find the boundary points, which are the most qualified data points to train SVM, between two classes. In this paper, present a new approach of combination of SVM and DGSOT, which starts with an initial training set and expands it gradually using the clustering structure produced by the DGSOT algorithm.

Ying Zhao, *et al.* [19] Fast and high-quality document clustering algorithms play an important role in providing intuitive navigation and browsing mechanisms by organizing large amounts of information into a small number of meaningful clusters. In particular, clustering algorithms that

build meaningful hierarchies out of large document collections are ideal tools for their interactive visualization and exploration as they provide data-views that are consistent, predictable, and at different levels of granularity. This paper focuses on document clustering algorithms that build such hierarchical solutions and (i) presents a comprehensive study of partitional and agglomerative algorithms that use different criterion functions and merging schemes, and (ii) presents a new class of clustering algorithms called *constrained agglomerative algorithms*, which combine features from both partitional and agglomerative approaches that allows them to reduce the early-stage errors made by agglomerative methods and hence improve the quality of clustering solutions. The experimental evaluation shows that, contrary to the common belief, partitional algorithms always lead to better solutions than agglomerative algorithms; making them ideal for clustering large document collections due to not only their relatively low computational requirements, but also higher clustering quality. Furthermore, the constrained agglomerative methods consistently lead to better solutions than agglomerative methods alone and for many cases they outperform partitional methods, as well.

Deng Cai, *et al.* [20] Consider the problem of clustering web image search results. Generally, the image search results returned by an image search engine contain multiple topics. Organizing the results into different semantic clusters facilitates users' browsing. In this paper, propose a hierarchical clustering method using visual, textual and link analysis. By using a vision-based page segmentation algorithm, a web page is partitioned into blocks, and the textual and link information of an image can be accurately extracted from the block containing that image. By using block-level link analysis techniques, an image graph can be constructed. Then apply spectral techniques to find a Euclidean embedding of the images which respects the graph structure. Thus for each image, have three kinds of representations, i.e. visual feature based representation, textual feature based representation and graph based representation. Using spectral clustering techniques, we can cluster the search results into different semantic clusters. An image search example illustrates the potential of these techniques.

Jason W.beckstead , *et al.* [21] This paper is a pedagogical piece on hierarchical cluster analysis, a method for investigating the structure underlying data. Such methods are useful for finding similar groups of cases in data sets when it is not known a priori how many groups are present. The paper is laid out as follows: First, a brief history and overview of the methods is presented; second, an illustrative example with a small hypothetical data set is used to clarify fundamental concepts; third, hierarchical cluster analysis is applied to a data set from the author's own program of research to illustrate one way in which the methods may be employed in nursing research; fourth, the limitations of the methods are discussed; and finally, a list of suggested readings, at varying levels of detail, are provided for the interested researcher.

Santhana Krishnamachari, *et al.* [22] Image retrieval systems that compare the query image exhaustively with each individual image in the database are not scalable to large databases. A scalable search system should ensure that the search time does not increase linearly with the number of images in the database. In this paper, present a clustering based indexing technique, where the images in the database are grouped into clusters of images with similar color content using a hierarchical clustering algorithm. At search time the query image is not compared with all the images in the database, but only with a small subset. Experiments show that this clustering based approach offers a superior response time with high retrieval accuracy. Experiments with different database sizes indicate that for a given retrieval accuracy the search time does not increase linearly with the database size.

Chung-Horng Lung, *et al* [23] in wireless sensor networks (WSNs), hierarchical network structures have the advantage of providing scalable and resource efficient solutions. To find an efficient way to generate clusters, this paper adapts the well-understood hierarchical agglomerative clustering (HAC) algorithm by proposing a distributed HAC (DHAC) algorithm. With simple six-step clustering, DHAC provides a bottom-up clustering approach by grouping similar nodes together before the cluster head (CH) is selected. DHAC can accommodate both quantitative and qualitative information types in clustering, while offering flexible combinations using four commonly used HAC algorithm methods, SLINK, CLINK, UPGMA, and WPGMA. With automatic CH rotation and re-scheduling, DHAC avoids re clustering and achieves uniform energy dissipation through the whole network. Simulation results in the NS-2 platform demonstrate the longer network lifetime of the DHAC than the better-known clustering protocols, LEACH and LEACH-C.

III. CONCLUSION

Clustering has also been a topic of interest in many different disciplines for a long time. Many clustering methods have been successfully used in other application areas. This paper is analysis the concepts of hierarchical clustering algorithm. Hierarchical clustering is a method of cluster analysis which seeks to build a hierarchy of clusters. Hierarchical clustering plays an important role in the performance of WSNs, and research associated with routing is always a focus. One major requirement in WSNs is energy efficiency. In this paper also discussed some research papers of hierarchical clustering methods to WSNs. This research results indicate that the approach has potential to provide an efficient and flexible way to manage issues of WSNs using hierarchical clustering.

REFERENCES

[1]. Pavel Berkhin (2000), Survey of Clustering Data Mining techniques ,Accrue Software, Inc..
[2]. Jiawei Han and Micheline Kamber (2006), Data Mining: Concepts and Techniques, The MorganKaufmann/Elsevier India.
[3]. K. Florek, J. Lukaszewicz, J. Perkal, H. Steinhaus, and S. Zubrzycki. Sur la liaison et la division des points d'un ensemble fini. Colloqium Math., 2:282–285, 1951.

[4]. L. L. McQuitty. Elementary Linkage Analysis for Isolating Orthogonal and Oblique Types and Typal Relevancies. Educational and Psychological Measurement, 17:207–209, 1957.
[5]. P. H. A. Sneath and R. R. Sokal. Numerical taxonomy: the principles and practice of numerical classification. W. H. Freeman, 1973.
[6]. Y. Song, S. Jin and J. Shen, A unique property of single-link distance and its application in data clustering, Data & Knowledge Engineering, 70 (2011), 984-1003.
[7]. D. Krznaric and C. Levcopoulos, Optimal algorithms for complete linkage clustering in dimensions, Theoretical Computer Science, 286 (2002), 139-149.
[8]. Agglomerative Clustering in Web Usage Mining: A Survey Karuna Katariya M. Tech Scholar R. K. University Gujarat, India Rajanikanth Aluvalu School of Engineering R. K. University Gujarat, India International Journal of Computer Applications (0975 – 8887) Volume 89 – No 8, March 2014
[9]. A Study of Network Intrusion Detection by Applying Clustering Techniques Subaira.A.S1, Anitha.P2 PG Scholar, Department of CSE, Dr.N.G.P.Institute of Technology, Coimbatore, India1 Assistant Professor, Department of CSE, Dr.N.G.P.Institute of Technology, Coimbatore, India2
[10]. Comparison Of Hierarchical Agglomerative Algorithms For Clustering Medical Documents Fathi H. Saad1, Omer I. E. Mohamed2, and Rafa E. Al-Qutaish2 1 National Health Services (NHS), London, UK f_miligi@yahoo.com 2 Al Ain University of Science and Technology, Abu Dhabi, UAE omar.mohamed@aau.ac.ae, rafa.alqutaish@aau.ac.ae
[11]. Agglomerative Connectivity Constrained Clustering for Image Segmentation Jia Li∗ Jia Li is an Associate Professor in the Department of Statistics at the Pennsylvania State University. Email: jiali@stat.psu.edu
[12]. Agglomerative clustering of a search engine query log Doug Beeferman, Lycos Inc. 4002 Totten Pond Road Waltham, MA 02451 dbeeferman@lycosinc.com Adam Berger , School of Computer Science Carnegie Mellon University 5000 Forbes Avenue Pittsburgh, PA 15213 aberger@cs.cmu.edu
[13]. Hierarchical Design Based Intrusion Detection System For Wireless Ad Hoc Sensor Network Mohammad Saiful Islam Mamun Department of Computer Science, Stamford University Bangladesh, 51, Siddeshwari, Dhaka. E-mail : msimamun@kth.se A.F.M. Sultanul Kabir Department of Computer Science and Engineering, American International University Bangladesh, Dhaka. afmk@kth.se
[14]. Data Compression Algorithm based on Hierarchical Cluster Model for Sensor Networks Wang Lei, Wang Tongsen, Yang Ronghua Department of Electronic Information & Electrical Engineering Fujian University of Technology, Fuzhou, P.R.China, 350108 International Journal of Advanced Science and Technology Vol. 2, January, 2009
[15]. Hierarchical Clustering Of Business Process Models Jae-Yoon Jung_, Joonsoo Bae__,1 and Ling Liu___Department of Industrial Engineering Kyung Hee University Yongin-si, Gyeonggi-do, 446-701, Republic of Korea jyjung@khu.ac.kr Received July 2008; revised December 2008
[16]. Applying Hierarchical Agglomerative Clustering to Wireless Sensor Networks Chung-Horng Lung, Chenjuan Zhou, Yuekang Yang Department of Systems and Computer Engineering Carleton University Ottawa, Ontario, Canada K1S 5B6 {chlung, cjzhou, yyang}@sce.carleton.ca
[17]. Generalized Agglomerative Clustering with Application to Information Systems Sadaaki Miyamoto Volume 5285 of the series Lecture Notes in Computer Science pp 158-166 Latifur Khan · Mamoun Awad · Bhavani Thuraisingham
[18]. A new intrusion detection system using support vector machines and hierarchical clustering Received: 13 January 2005 / Revised: 10 June 2005 / Accepted: 21 July 2005/published online august 2006 c Springer-Verlag 2006
[19]. Hierarchical Clustering Algorithms for Document Datasets YING ZHAO yzhao@cs.umn.edu GEORGE KARYPIS karypis@cs.umn.edu University of Minnesota, Department of Computer Science and Engineering and Digital Technology Center and Army HPC Research Center, Minneapolis, MN 55455 Editor: Usama Fayyad Submitted June 21, 2003; Revised July 23, 2004
[20]. Hierarchical Clustering of WWW Image Search Results Using Visual, Textual and Link Information Deng Cai1* Xiaofei He2* Zhiwei Li* Wei-Ying Ma* and Ji-Rong Wen* Using Hierarchical cluster analysis in nursing research Jason W.beckstead Western journal of nursing research 2002, 24(3),307-319
[21]. Hierarchical clustering algorithm for fast image retrieval Santhana Krishnamachari Mohamed Abdel-Mottaleb Philips Research 345 Scarborough Road Briarcliff Manor, NY 10510 {sgk,msa}@philabs.research.philips.com Part of the IS&T/SPIE Conference on Storage and Retrieval for Image and Video Databases VII San Jose, California, January 1999. pp427-435.
[22]. Using hierarchical agglomerative clustering in wireless sensor networks: An energy-efficient and flexible approach Chung-Horng Lung *, Chenjuan Zhou Department of Systems and Computer Engineering, Carleton University, Ottawa, Ontario, Canada K1S 5B6

AUTHORS PROFILE

S.Aravindhan received his M.Phil degree from Thiruvalluvar university,Vellore in the year 2012.He has received his MCA degree from Anna university,Chennai in the year 2011.He is pursuing his Ph.D degree at Periyar University, Salem, Tamilnadu, India. His areas of interest include Data Mining, Cloud Computing and Computer Networks.

Dr.D.Maruthanayagam received his Ph.D Degree from Manonmanium Sundaranar University, Tirunelveli in the year 2014. He has received his M.Phil, Degree from Bharathidasan University, Trichy in the year 2005. He has received his M.C.A Degree from Madras University, Chennai in the year 2000. He is working as Assistant Professor, Department of Computer Science, Sri Vijay Vidyalaya College of Arts & Science, Dharmapuri, Tamilnadu, India. He has 14 years of experience in academic field. He has published 1 book, 15 International Journal papers and 23 papers in National and International Conferences. His areas of interest include Grid Computing, Cloud Computing and Mobile Computing.