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

ABSTRACT
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 a input. Agglomerative have many drawbacks such as initial error propagation, dimensionality, complexity and data set size issues. In this paper we have introduced solution for data set size problem that helpful for information retrieve from large web data, web log data files are as a input for agglomerative clustering algorithms and output is efficient clustering that will be used further for information extraction in web usage mining.

General Terms
Clustering, Agglomerative Clustering

Keywords
Web Usage Mining, Clustering, Agglomerative Clustering

1. INTRODUCTION
Now a day's explosive growth of data collection, so data is stored in data warehouses and that data is accessed by intranet or internet. Data mining is a process of extract or retrieves use full information from the large set of data. Today number of user is increasing so use of web is growing exponentially; World Wide Web is a wide source of information. Web mining is a data mining technique that is used to automatically extract information from web [1].web mining is divided in three type such as content mining(content of pages), structure mining(structure of pages), usage mining(access or use of pages).

Web usage mining is to discover browsing pattern from user's behaviors [4]. Web usage mining is helps to deal with certain web scaling problem such as user trend analysis through surfing, traffic flow analysis, distributed control and handling, web traffic management and many more [2].

Clustering is an unsupervised learning technique. Using clustering technique on extracted information from web data, to separate all the information is known as clustering. Similar characteristics data is placed in one group and remaining data is placed in another group.

There for in Cluster certain data point is similar with same cluster data points and "dissimilar" with other clusters data points. Such methods are used in Data Mining, Pattern recognition, Image analysis, Bioinformatics, Machine Learning. Voice mining, Image processing, Text mining, Web cluster engines, Whether report analysis [6]. Requirement of clustering is a Scalability, Ability to deal with different types of attributed, Discovery of clusters with arbitrary shape; Minimal requirements for domain knowledge of determine input parameters, Ability to deal with noisy data, Insensitivity to the order of input records, High dimensionality [14], Constraint-based clustering Interpretability and usability. Complexity is high in clustering and inability of clustering.

2. RELATED WORK
2.1 Type of Clustering
A collection of similar data objects so clustering [11] have two types of similarities.

- Intraclass similarity- Objects are similar to objects in same cluster. Intraclass similarity [15] is based on the near data objects in a cluster. In a similarity measured in clustering is an important role in doing good clustering so intraclass similarity gives the large value for good clustering. it is measured by using this equation:

$$ICS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \frac{1}{|C_i|^2} \sum_{d_{1d'\epsilon C_i}} Similarity(d, d')$$

- Interclass dissimilarity- Objects are dissimilar to objects in other clusters. Interclass similarity [15] is give less value for good clustering. So similarity between different cluster object is less. This formula is used of measured similarity of cluster objects in this method.

$$ECS(\pi(D)) = \sum_{i=1}^{|\pi(D)|} \sum_{j=i+1}^{|\pi(D)|} \frac{1}{|C_i|\,|C_j|} \sum_{d\epsilon C_{i,d'\epsilon C_i}} Similarity(d, d')$$

2.2 Various Types of Clustering Method
Clustering is a process of classify data into number of cluster. Clustering classifies data based on different types methods that method follows are:

- Hierarchical clustering method
- Partitioning clustering method
- Density based clustering method
- Grid based clustering method
- Model based clustering method

2.2.1 Partitioning clustering

In partitioning method have N object database and that database is partitioned [18] in k groups using partitioning method. In all objects contain in one cluster and at least one object contain in each group. This method is suited for small to medium sized data set to finding spherical-shaped clusters. It is used for complex data set and cluster very large data set. The representative partitioning [18] clustering algorithms are K-MEDOID, K-MEANS.

K-means clustering algorithm is work in that manner first select the randomly one node from the number of objects. K is a center of cluster. Similar [18] data form a cluster, similarity find based on the distance between object and center.

2.2.2 Model based clustering

Model based clustering method construct the model for every clusters and find a data which is fit to that model and this method is automatically give the number of clusters. This method is robust. The representative model-based clustering algorithm is EM. Model based method is often based on probability distribution of data. Individual distribution is called component distribution .in this method probability distribution is done by the mixture density model.

EM method acquire statistic from traditional mixture model and depends on that statistic it perform clustering in model based clustering method.

2.2.3 Density-based clustering

In density based method divide data in cluster based on the density of objects. So distance between cluster objects is less and growing the number of objects [21] though the density based clustering algorithm. So density of cluster is growing. And it have same advantages such as reduced effect of noise (outliers) and discover clusters of arbitrary shape, input data scan only once, needs density[21] parameters to be initialized.

Here Density-based clustering algorithms the data space contain dense regions of objects is consider as a cluster and clusters are separated by regions of low density. Density based algorithms depends on each object with a density value defined by the number of its neighbor objects within a given radius. Density [21] of objects is greater than a specified threshold is defined as a dense object and initially is formed a cluster itself. Two clusters are merged if the y shares a common neighbor that is also dense. The DBSCAN, OPTICS, HOP, and DENCLUE algorithms are representative of density-based [21] clustering algorithms concept.

2.2.4 Grid based clustering

In grid based clustering method is divided data space in number of cell that forms grid structure. Grid structure is depends on the number of cell rather than number of object. Perform clustering in a grid so complexity is reduced in grid [13] based clustering method. Statistic attribute are gathered form grid cell. Performance is depends on the size of the grid that is less than the number of objects contain in cluster. The representative grid based clustering [13] algorithms are STING, WAVE CLUSTER, and CLINQUE.

2.2.5 Hierarchical clustering

Hierarchical clustering builds a cluster hierarchy such as a tree of clusters. Hierarchical clustering processes is a sequence of divide or merge clusters. In which each cluster have chilled and structure that is more informative than the unstructured set of clusters returned by flat clustering. No need of give number of cluster initially. Good result visualization in this method and complexity is high. Hierarchical clustering [12] is used in information retrieval. In which distance of objects is measured and merge or divide cluster objects based on the distance. Unstructured data is divided or merged effectively in hierarchical clustering.

2.2.5.1 Divisive (Top-Down) Approach

Here starting with a one cluster of all objects and recursively splitting each cluster until the termination criterion is reached [12]. The most useful part of hierarchical clustering is that it can be applicable to any type of attributes, so easy to apply for any task related to web usage mining with respect to web data.

2.2.5.2 Agglomerative Technique (Bottom-Up)

Hierarchical and partition is a clustering method, in the partitioning method required the number of clusters as a input while hierarchical clustering method are no need to number of cluster as a input, so unknown data set given as a input. Hierarchical clustering contains two methods top-down and bottom-up. Agglomerative clustering [7] is a bottom-up method. That method is simple and very flexible method.

Agglomerative clustering algorithm works by grouping the data one by one on the basis of the nearest distance measure of all the pair wise distance between the data point. Again distance between the data point is recalculated but which distance to consider when the groups has been formed [7]. Single linkage, complete linkage, average linkage and centroid distance between two points, grouping the data until one cluster is remaining.

Agglomerative clustering starting with one point clusters and recursively merging two or more most similar clusters [12] to a single cluster (parent) until the termination criterion is reached. (E.g. k – clusters have been built).

It has advantages that No apriori information about the number of clusters required and Easy to implement and gives best result in some cases. In which Algorithm can never undo what was done previously. And sometimes it is difficult to identify the correct number of clusters by the dendogram. Major drawbacks are that initial error propagation, dimensionality, complexity and large data set size.

ALGORITHM

Agglomerative clustering algorithm steps:

1) Let be the set of data points.
2) Then find distance between the clusters.
3) Merge pair of clusters that have smallest distance.
4) Update distance matrix.
5) If all the data points are in one cluster then stop, else repeat from step 2) [7].

2.2.5.2.1 Various technique of agglomerative clustering

1) single linkage

Single linkage clustering is one of the methods of agglomerative hierarchical clustering. In single linkage clustering link between two clusters is made by single object pair, it is also known as a nearest neighbor clustering. Distance between two clusters is based on points of clusters that is small or nearest. Mathematically, the linkage function is $D(X, Y)$ – the distance between clusters [10] X and Y – is described by the equation

$$D(X,Y) = \min_{x \epsilon X, y \epsilon Y} d(x,y)$$

- Where X and Y are any two sets of objects considered as clusters
- $D(x, y)$ function denotes the distance between the two elements x and y.

2) Complete linkage

In complete-linkage clustering, the link between two clusters contains all element pairs, it is also known as a farthest neighbor clustering. Distance between two clusters is depends on objects of clusters that is maximum or farthest. Mathematically, the complete linkage function—the distance D(x, y) between clusters X and Y— is described by the following expression [10]:

$$D(X,Y) = \max_{x \epsilon X, y \epsilon Y} d(x,y)$$

Where

- d(x,y) is the distance between elements $x \epsilon X$ and $y \epsilon Y$;
- x and y are two sets of elements (clusters)

3) Average linkage

In average linkage clustering the link between two clusters contains one point of cluster to all points of other cluster [10]. In which average distance between pairs of clusters data points is denote the distance of two clusters.

$$Sim(C_i, C_j) = \frac{1}{|C_i \cup C_j|(|C_i \cup C_j - 1)} \sum_{\vec{x} \in (C_i \cup C_j)} \sum_{\vec{y} \in (C_i \cup C_j), \vec{y} \neq \vec{x}} Sim(\vec{x}, \vec{y})$$

4) Centroid linkage

In centroid linkage clustering the link between two clusters contains one point center of cluster to center points of other cluster. The distance between clusters is defined as the (squared) Euclidean Distance between cluster centroids [10].

2.2.5.2.2 Issues in agglomerative clustering

In general agglomerative clustering have issues such as dimensionality, initial error propagation, complexity and large data set size issues.

Large data set size

Today's data is growing so storage of data is expanded. And data set sized is increased day by day. There for clustering of data is needed because of all data or information is not important for doing any operation. And all data is classified in attributes. But in agglomerative clustering algorithm has a issues of not classified large data set [16]. In which number of data or information of attribute is large but not important that number of attribute is more. So it is independent to the how many attribute contain in a data base.

Agglomerative clustering algorithm is performing on small or medium size data set. This method has limitation that is not work on large data set size, so large data set size issue is a major issue in agglomerative clustering algorithm. Large data set sized[16] problem can be depends on a three reason such as 1) Data set can have large number of element 2) In data set contain each element can have number of features. 3) Many clusters can be contain in data set for discover all the data.

High Dimensionality

Agglomerative clustering algorithms are designed for two or three dimension data. So high dimension data is a major challenge for clustering because of dimensionality [17] is increased. Small number of dimension is relevant to the exits clusters. And many dimensions are irrelevant. There for noise in data is increased. When dimensionality [17] is increased data is become parse because data points are located on different dimension subspaces.

Dimensionality is depend on the number of attribute contain in data set instead of large number of features [17] of each attribute.

Complexity

Complexity is an amount of time or space required by an algorithm for given input size. Agglomerative clustering algorithm complexity [19] is O (n^3) because the similarities N x N metrics scan every time and find similarity N-1 iterations and give the N number of clusters as an input. So complexity [19] of agglomerative clustering algorithm is high.

And in which storage space is needed for similarity metrics so complexity is high in that method.

Initial error propagation

In an agglomerative clustering method error is contain in an initial step of the clustering and that is propagate [20] last step of the process of clustering. In agglomerative clustering algorithm merge number of cluster in one cluster at that time one cluster contain error and that error is not solved that step and that cluster [20] is merged with other cluster have data is a without error or true data, so data of that two cluster may be clash, So error can be occurred in that data is more.

In agglomerative clustering algorithm all that types of cluster is merge and error propagation can be more so result of that method is a not accurate at the final step of algorithm [20].So initial error propagation is a problem of agglomerative clustering algorithm.

Table 1: Measurement result of different methods

Algorithm	**Time**	**Accuracy**
Single Linkage	0.28	0.98
Complete Linkage	3.68	4.56
Average Linkage	1.78	8.62

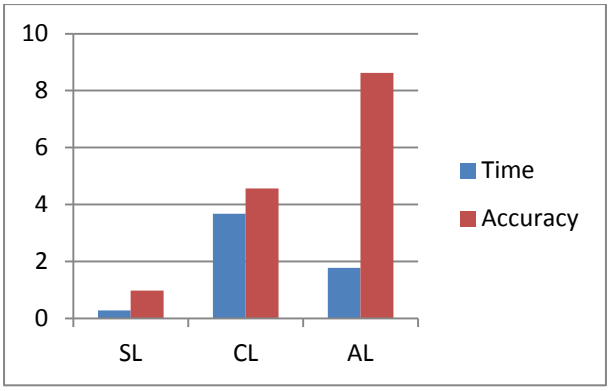

Fig 1: Comparison graph of different methods

3. CONCLUSION

We have come across many issues such as initial error propagation, complexity, dimensionality and large data set size during the review of various clustering algorithms mainly agglomerative approach. The study also revealed that Dimensionality and Initial error propagation issues are solved [7]. The issues related to data set size and complexity need to be resolved to perform usage trend analysis, path analysis and clustering based on data.

Agglomerative clustering algorithm can be improved using linkage methods and an initial error propagation issue can be solved using single linkage method. Single linkage, complete linkage, average linkages are various agglomerative clustering distance calculation methods. By improving these methods, these methods can be useful for agglomerative clustering algorithm.

4. REFERENCES

[1] Anjali B. Raut, G.R. Bamnote," Web Document Clustering Using Fuzzy Equivalence Relations", Volume 2, Issue 1, February 2011.

[2] Sonali muddalwar, Shashank Kawan, "Applying artificial neural networks in web usage mining", international journal of computer science and management research, vol 1 issue 4 [NOV-12].

[3] Sumaiya banu, kayitha, swetha, sathiya raj "Amended agglomerative clustering for web users navigational behavior" Indian journal of engineering, Vol 3, issue 8, June 2013.

[4] Jaykumar Jagani," A survey on web usage mining with neural network and proposed solutions on various issues", ICRD-ETS-2013.

[5] Jaydeep Srivastava, "Web Mining: Accomplishments and future directions", http: //www. cs. unm .edu/ faculty/ srivastava. html.

[6] A. K. jain, M. N. murty, p. j. Flynn, "data clustering: A review", volume 31, issue 3, September 1999.

[7] phivos Mylonas, Manolis Wallace, and Stefanos Kollias," Using k-nearest Neighbor and Feature Selection as an Improvement to Hierarchical Clustering", issue-2004.

[8] http:// sourcemaking. com / uml/ modeling –it - systems/structural- view/generalization specialization- and inheritance

[9] http: / /www. chegg.com /homeworkhelp /definitions /optimization-29

[10] http: // bus. Utk .edu /stat /stat579 /Hierarchical % 20 Clustering%20Methods.pdf

[11] L. V. Bijuraj," Clustering and its Applications" Conference on New Horizons in IT-NCNHIT 2013

[12] Aastha joshi, Rajneet Kaur," A Review: Comparative Study of Various Clustering Techniques in Data Mining", Volume 3, Issue 3, March 2013

[13] Wei-keng Liao, Ying Liu,Alok Choudhary," A Grid-based Clustering Algorithm using Adaptive Mesh Refinement", Appears in the 7th Workshop on Mining Scientific and Engineering Datasets 2004

[14] Osmar R. Zaiane, Andrew Foss, Chi-Hoon Lee, and Weinan Wang," On Data Clustering Analysis: Scalability, Constraints and Validation"

[15] Selim Mimaroglu and A. Murat Yagc," A Binary Method for Fast Computation of Inter and Intra Cluster Similarities for Combining Multiple Clusterings"

[16] M.Vijayalakshmi, M.Renuka Devi, "A Survey o f Different Issue of Different clustering Algorithms Used in Large Datasets" , Volume 2, Issue 3, March 2012

[17] Andrew McCallum, Kamal Nigam, Lyle H. Unga,r" Efficient Clustering of High-Dimensional Data Sets with Application to Reference Matching" ,issue 07/2000.

[18] Jiawei Han, Micheline Kamber "Data Mining: Concepts and Techniques"

[19] Bruce Walter, Kavita Bala, Milind Kulkarni, Keshav Pingali," Fast Agglomerative Clustering for Rendering"

[20] Manolis Wallace, Stefanos Kollias," Robust, generalized, quick efficient agglomerative clustering ",issue 2004.\

[21] Hans-Peter Kriegel, Martin Pfeifle," Density-Based Clustering of Uncertain Data"