Market User Analyzer using OKH Algorithm

G. P. Mohole
Department of computer engineering,
Marathwada Institute of Technology, A'bad (MH)

S. A. Kinariwala
Department of computer engineering,
Marathwada Institute of Technology, A'bad (MH)

ABSTRACT
For the analysis of business, a lot of research attention in the field of computational statistics and data mining has been made. Due to recent technological advances in the field of data clustering, the researchers face ever-increasing challenges in extracting relevant information from enormous volumes of available data. The paper focus on large data sets obtained from online web visiting and categorizing this into clusters according some similarity it helpful tool for the top level management to take optimized and beneficial decisions of business expansion. Clustering is the assignment of a set of observations into subsets. Cluster analysis is widely used in market research when working with multivariate data from surveys. Market researcher partition the general population of consumers into market segments and understand the relationships between customers. To achieve robustness and efficiency in data clustering combine Partitional and hierarchical (Optimized K-means algorithms) satisfiable clustering results.

General Terms
Use concept Data Mining using Partitional and Hierarchical algorithm

Keywords
Clusters, Partitional Algorithm, K-Means, Optimized K-means, Hierarchical algorithm, Single Link algorithm.

1. INTRODUCTION

1.1 Existing System

The premise of data mining is that it is necessary to establish the enterprise-level customer-information data warehouse which can assemble the customer data and can provide a correct, complete and single customer data circumstance for instituting better customer service strategies.

It becomes highly challenged to digital firms due to the exponential growth of customer data in today's electronic business environment. Hence market segmentation proves to be one of the possible solutions. The goals of market segmentation include retention of customer, allocation of advertising resource, and an increase of profit margins. The outcome of market segmentation plays important role in market development, advertising, product pricing, attracting new customers and other marketing strategies.[3],[4]

A good analytical tool should be able to compare the characteristics of different segments (group) and identify important attributes of each segment (group) to create opportunity for targeting customer which can be accomplished by getting the complete context of the customer. Clustering is just as the normal saying that "things of one kind come together", that is, to classify a group of individuals into several groups. Its aim is to make the distance between the individuals of the same group as short as possible, and the distance between the individuals of different groups as far as possible.

Clustering analysis can find out the data distribution mode and proper inter relationship between data properties from the macroscopical view, search for the useful relation between the object data in certain data set and divide the records in database into a series of meaningful subclass.

By combining the features of partitional and hierarchical clustering methods, the input data set into several small sub clusters in the first phase and then continuously merges the sub clusters based on cohesion in a hierarchical manner in the second phase using optimized K-means algorithm. As shown in figure below, the data set is partitioned into 15 sub clusters and these sub clusters are next grouped into two clusters.

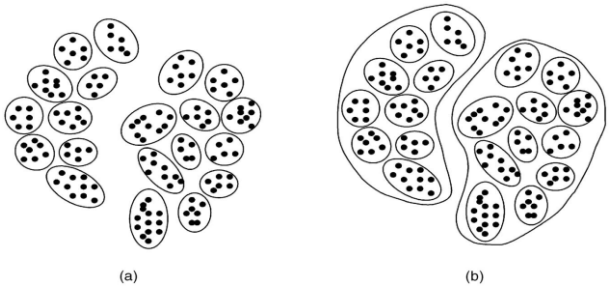

Fig.1.1.Partitional and Hierarchical clustering algorithm. (a) Several small clusters.(b)Merge small subclusters into clusters.

As shown by our performance studies, this algorithm is very robust and possesses excellent tolerance to outliers in various workloads. More importantly, it is able to cluster the data sets of arbitrary shapes very efficiently and provides better clustering results than prior methods. K-means method is one of the most commonly used approaches for classification and is an exclusive clustering algorithm, where if a certain data point belongs to a definite cluster then it could not be included

in another cluster. K-means is considered a fast method because it is not based on computing the distances between all pairs of data points. The commonly used distance in K-means method is Euclidean distance. In this paper, we made an optimization of the standard K-means method, and the new method make the segmentation much faster, while still get exactly the same result as the traditional K-means algorithm. Thus, the application helps determine the user's browsing details and monitor customer population and generates a summarized representation of the same.

1.2 Necessity

The focus area of this system is market research and analysis. It is a web-based application and aims at determining target markets and consumer density and identifying potential customers. We have used the concept of Cluster analysis for the same. This application will help determine the user's browsing details and monitor customer population.Web User analysis is a simple template that provides a graphical, time-phased overview of process in terms of conceptual design, mission, analysis, and definition phases.

Fig. 1.2 Market User Analysis template diagram

It becomes highly challenged to digital firms due to the exponential growth of customer data in today's electronic business environment. Hence market segmentation proves to be one of the possible solutions. The goals of market segmentation include retention of customer, allocation of advertising resource, and an increase of profit margins. The outcome of market segmentation plays important role in market development, advertising, product pricing, attracting new customers and other marketing strategies. A good analytical tool should be able to compare the characteristics of different segments (group) and identify important attributes of each segment (group) to create opportunity for targeting customer which can be accomplished by getting the complete context of the customer.

With this system interested to target the customer of my own web portal in major cities of Maharashtra. To find that from which location of the world the visitors of the website to trace the future market at that place. Maximum used services by this visitors also referral of the website and whole control of the system at administrator. The similar kind of facility available with Google Analytic Tools but for every domain we have to pay the different fees also the whole database with Google. Report formation not as per our requirement which is the major task in every system.

1.3 Objectives

The main objective of this paper is market research from mass of real time data which work faster, better and robust. Along with market research the project will cover the following aspects**:**

i. Market research.
ii. Analyzing visitor traffic country wise and product wise.
iii. Customer tracking
iv. Referrer researches
v. Product Positioning
vi. Business expansion
vii. Predicting future markets
viii. Retrieve user's browsing details
ix. Reporting

In modern society, customer becomes the most important asset of enterprise. Efficient customer relationship management is the necessary method to improve enterprise competition advantage. By use of data mining technology, rules and patterns can be extract from a mass of data, but how to explain and apply these rules and patterns is the key point enterprises care about and also the determinant factor to assist the enterprise decision-making. [6]

1.4. Theme

To fulfil the objectives the cluster analysis concept decided in this project. We know that cluster analysis or clustering is the assignment of a set of observations into subsets (called *clusters*) so that observations in the same cluster are similar in some sense. As also Cluster analysis is widely used in market research when working with multivariate data from surveys and test panels. Market researchers use cluster analysis to partition the general population of consumers into market segments and to better understand the relationships between different groups of consumers/potential customers.

For Web User Analyzer, try to do an experiment on real time dataset: Web User Analyzer (Data after the click on our Web Portal) . [7] When the visitors click on our website we try to collect the database by tracking his IP on the basis of the Geo-Graphical Location (Longitude And Latitude of the City) the cluster formation take place. Also try to find which application selected by the visitor and of which city to collect the more detail data about the visitor.

On initial level we define the center of sub clusters on the basis of the city user visited. As more clicks happen the cluster center is constant a sub cluster form on the basis of

which application of that city visited by the user *eg.* Job, Sale, Education of Nashik City using optimized K-means algorithm. This all sub clusters are combined and prepared the Merge Clusters by applying the hirarchical clustering algorithm which find the details about particular city which service request by the visitor more. In project major interest is to form the clusters by using optimized K-means algorithm. So the performance of the system improved and by hierarchical algorithm the system becomes very much robust and efficient.

2. RELATED ALGORITHMS

2.1 Standard K- Means Algorithm

The k-means algorithm is one popular partitional algorithm for data clustering. The *k*-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster. [8]

The algorithm steps are

1. Choose the number of clusters, *k*.
2. Randomly generate *k* clusters and determine the cluster centers, or directly generate *k* random points as cluster centers.
3. Assign each point to the nearest cluster center, where "nearest" is defined with respect to one of the distance measures discussed above.
4. Recompute the new cluster centers.
5. Repeat the two previous steps until some convergence criterion is met (usually that the assignment hasn't changed).

The main advantages of this algorithm are its simplicity and speed which allows it to run on large datasets. Its disadvantage is that it does not yield the same result with each run, since the resulting clusters depend on the initial random assignments. From the algorithmic framework, we can see that the algorithm need to adjust the sample classification continuously, and calculate the new cluster centers constantly. Therefore, the time consumption is fairly considerable when the dataset is large. It is necessary to improve the efficiency of the algorithm's application. If a data point is far away a center, it is not necessary to calculate the exact distance between the point and the center in order to know that the point should not be assigned to this center. So most distance calculations in standard K-means are redundant.

2.2 Optimized K-means Algorithm

In this paper, we use triangle inequality to reduce these redundant calculations. In this way we improved the efficiency of the algorithm to a large extent. As can be seen from the generally acknowledged truth, the sum of two sides is greater than the third side in a triangle. Euclidean distance meets the triangle inequality, which we can [1] extend to the multi-dimensional Euclidean space. We can take three vectors in Euclidean space randomly: x, a, b, then:

$$d(x,a) + d(a,b) \geq d(x,b)$$

$$d(a,b) - d(x,a) \leq d(x,b)$$

$d(Ci,Cj)$, is the distance between two cluster centers.

If $2d(x,a) \leq d(Ci,Cj)$ then:

$$2d(x,Cj) - d(x,Cj) \leq d(Cj,Ck) - d(x,Cj) \quad (1)$$

according to equation (1) then : $\boldsymbol{d(x,Cj) \leq d(x,Ck)}$

First, select initial cluster centers, and set the lower bound y(x,f)=0 for each data point and cluster center. Second, assign each data point to its nearest initial cluster, we will use the results obtained previously to avoid unnecessary distance calculations in this process. Each time d(x,f) is computed, set y(x,f)=d(x,f).

2.3 Hierarchical Algorithm

As its name implies, a hierarchical clustering algorithm [2] establishes a hierarchical structure as the clustering result.Consider the example in Fig. 2 one possible hierarchical Structure. With the hierarchical structure, we can obtain different clustering results for different similarity requirements.

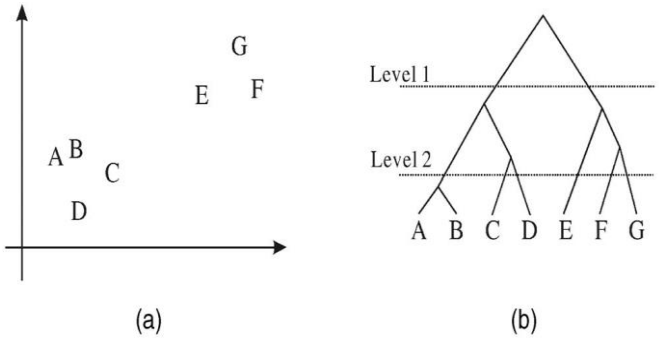

Fig. 2.1 Illustrative hierarchical clustering results for a data set of seven points. (a)The input data set. (b) A possible hierarchical tree.

As shown in Fig. 2b, if the similarity requirement is set at level 1, the input data set is partitioned into two clusters, i.e., {A,B,C,D} and {E,F,G}. However, if the similarity requirement is set at level 2, then the input data set is partitioned into six clusters, i.e., {A,B}, {C},{D}, {E}, {F} and {G}.

Most existing hierarchical clustering algorithms are variations of the single-link and complete-link algorithms. By good quality of clustering results, hierarchical algorithms are widely used, especially in document Clustering and classification.[5] The outline of a general hierarchical clustering algorithm is given below:

Hierarchical Clustering Algorithm

1. Initially, each data point forms a cluster by itself.

2. The algorithm repetitively merges the two close Clusters.

3. Output the hierarchical structure constructed.

A single-link clustering algorithm differs from a complete

link clustering algorithm in the intercluster distance measure, i.e., Step 2. The single-link algorithm uses the distance between the two closest points of the two clusters as the intercluster distance [2]

1. Initially, each data point forms a cluster by itself.

2. The algorithm repetitively merges the two closest

clusters.

3. OKH ALGORITHM

Now, we describe the proposed algorithm based on Optimized K-Means and Hierarchical Clustering as follows:

Algorithm OKH

//Input : The input data set, the size of the data set, n, the number. of sub clusters, m, and the desired number of clusters, k.

//Output : The hierarchical structure of the k clusters.

1. Apply Optimized k-means on the input data set to obtain m sub clusters.

2. Apply the single-link clustering algorithm on the m sub clusters produced in Step 1 with based on the similarity measure of attributes and stop when k clusters are obtained.

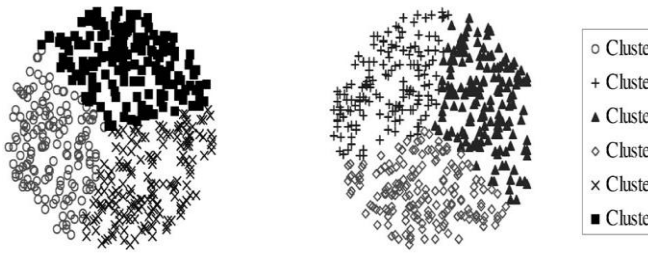

Fig. 3.1. Example of OKH Algorithm.

Algorithm OKH is a clustering algorithm. In the first step, it adopts the Hierarchical algorithm to divide the input data set into m subclusters. At the beginning of step second step it obtains these m subclusters produced in the first phase and produced k clusters. It is known that the optimized k-means algorithm is good for obtaining clusters of isotropic shape, while the single-link algorithm is able to find clusters of any shape. With prior knowledge, we can make the clustering algorithm adapt to various inputs by adjusting the parameter m.

4. REVIEW OF ALGORITHMS

In order to test the effectiveness of the improved algorithm, we try to do an experiment on real time dataset: Web User Analyzer (Data after the click on our Web Portal). [7] When the visitors click on our website we try to collect the database by tracking his IP on the basis of the Geo-Graphical Location (Longitude and Latitude of the City). Also try to find which application selected by the visitor and of which city. On initial level we define the center of sub clusters on the basis of the city user visited. As more clicks happen the cluster center is constant a sub cluster form on the basis of which application of that city visited by the user eg. Job, Sale, Education of Nashik City using optimized K-means algorithm. This all sub clusters are combined and prepared the Merge Clusters by applying the hierarchical clustering algorithm which find the details about particular city which service request by the visitor more. As per the search we find that by using optimized K-means algorithm the performance of the system improved and by hierarchical the system become very much robust and efficient.

5. APPLICATION

If the system develops to trace that which application of the website requested by the Visitors the Business Intelligence take place. Where with the good performace real time clusters are forms for every click and it generate the reports that represented in different kind of charts which reflect that the owner should concentrate on which application of website. After implementing the said algorithm with the seconds the reformulation of the Cluster take place

6. CONCLUSION

In modern society, customer becomes the most importantasset of enterprise. Efficient customer relationship management is the necessary method to improve enterprise competition advantage. By use of data mining technology, rules and patterns can be extract from a mass of data, but how to explain and apply these rules and patterns is the key point enterprises care about and also the determinant factor to assist the enterprise decision-making. [6]As we known, the standard K-Means algorithm is used in many fields. While, the efficiency is unsatisfactory when faced with large-scale dataset. In this paper, we improved the standard K-Means algorithm using triangle inequality and combining the Hierarchical Algorithm [5] for robustness in the data. We run the optimized algorithm on Visitors Data and Hierarchical algorithm on their applications requested. The proposed method definitely more efficient than the standard K-Means algorithm and by combining hierarchical algorithm the efficiency also improved. The proposed system is a solution to the problem of market research and analysis. It helps in determining target markets and consumer density and identifying potential customersall address text. For two addresses, use two centered tabs, and so on. For three authors, you may have to improvise.

7. REFERENCES

[1] Xiaoping Qin, Shijue Zheng , Tingting He□Ming Zou, Ying Huang, *"Optimizated K-means algorithm and application in CRM system"* 2010 International Symposium on Computer, Communication, Control and Automation

[2] Cheng-Ru Lin and Ming-Syan Chen, IEEE," *Combining Partitional and Hierarchical Algorithms for Robust and Efficient Data Clustering with Cohesion Self-Merging*" IEEE Transactions On Knowledge And Data Engg. Vol-17,No-2 Feb 2005 1041-4347/05

[3] Haibo Wang, Da Huo, Jun Huang ,Lixia Yan, Wei Sun, Xianglu Li " *An Approach for Improving K- Means Algorithm on Market Segmentation"* 2010 International Conference on System Science and Engineering.ICSSE2010 978-1-4244-6474-6110

[4] Wanghualin, "*Data Mining and Its Applications in CRM***",** 978-0-7695-4043-6/10 2010 IEEE DOI 10.1109/ICCRD.2010.184

[5] George Karypis ",*Hierarchical Clustering And DataSets"* Springer 2005 Data Mining and Knowledge Discovery 141-168 2005

[6] Zhe zhang,"Data Mining and Its Application in Customer Relationship Management", Shanghai: Fudan University Press, Aug.2007

[7] A.G. Buchner and M. Mulvenna, "Discovery Internet Marketing Intelligence through Online Analytical Web Usage Mining," Proc.

[8] J. Han and M. Kamber, Data Mining: Concepts and Techniques., 2006.