Web User Analysis Using Hierarchical and Optimized K-mean Algorithm for Online Market Analysis.

Ms.Poonam L. Rakibe(ME-II, Computer)
Prof.P.N.Kalvadekar(Guide)
Department of Computer Engineering,
S.R.E.S, Kopargaon
University of Pune
Email: poonamrakibe@rediffmail.com

Abstract

Data mining plays really very important part in the analy-sis of business.Computational statistics and information re-trieval has attracted attention of many researchers.As there is enormous data there are various challenges for the researchers to extract particular data as per users requirement. Considering large volumes of data which is produced due to various users visiting online online web portal clustering of that data needs to be done and categorizing of that data on basis of some similar attributes of the users.When standard K-means partitional algorithm is not that efficient because when it is subjected to large data set it consumes consider-able time. So use of Optimized K-Means and Hierarchical algorithm is to be done which will yield better results for on-line shopping web portal.Reports are generated from output of clustering algorihm .Results can helpful for business owners or website owner in market research for deciding marketing strategies,customer retention strategies ,production and oper-ations taking place in the business. .

Key terms

Data mining„partitional algorithm,hierarchical algorithm,K-means ,optimized K-Means algo-rithm,Market research

1. Introduction

1.1 Existing System

Data Mining Can be Used in Various business application for different purposes such as decision support system,customer retention strategies,selective marketing,businesss management,user profile analysis to name a few [1],[2],[3],[10].Data mining is the process of discovering the knowledge. In todays electronic informa-tion era it becomes highly challenged to digital firms to manage customer data to retrieve useful information as per their requirement from that data,so market segmentation can be used.Market segmaentation also include customer retention strategies,allocation of resources for advertising,to check profit margins.so outcome of segmenatation plays big role in deciding price of the products,iattracting new customers and identifying potential customers.Clustering analysis is able to find out data distribution and proper inter relationship between data items.clustering is defined as "grouping of similar data" .Clustering divide records in the database or data ob-jects in the dataset into series of meaningful subclasses or group.Data mining is basically a useful process in which the enterprise retrieve useful knowledge from a load of information which is incomplete and random that has been generated from various business tasks such aaas production,marketing,customer services of the enterprise.

A good analytical tool should be able to compare between different characteristics or attributes of different groups and indentify different important characteristics of each segments to decide different business strategies.Clustering analysis can find out the distribution of diff erent data entities as well can find out proper inter relationships between the data objects so that it can divide the data set into series of meaningful subclasses.

The motivational state of the art:

Limitations of K-means algorithm:

The constraint on K-means algorithm is that if a data point is included in one cluster then it cannot be included in another cluster .K-means algorithm has significant disadvantages[9][4].Time taken by K-means algorithm when large amount of data is applied to it is considerable.the result or the the output I.e. clusters formed using K-means algorithm are not always same though same data is applied to the K-means algorithm.it means output is different for each run because result depends upon random initial assignment of initial k clusters . The application helps to determine users visiting details ,which products he/she has viewed and which products he/she has purchased and keeps track oof customer population visiting the online shopping portal.To avoid limitations of traditional K-means methos such as output dependency .So there is increas-ing need of improved clustering algorithm which will yield same result but will take less amount of time to processs large amount of data in online shopping portal.so the optimization of K-means algorithm is to be done.

It becomes highly challenged to digital firms due to the increasing growth of customers data in todays electronic business world to analyse data.A good analytical tool should be able to analyses the data by comparing the characteristics of different data entries so as to take very important business decisions.

This paper describes the existing system in section 1 and describes motivation of the new are (why there is need of new improved algorithm) .Section 2describes work related to the proposed system which is nothing but liter-ature survey,associated challenges,drawbacks,efficiency of the system and experimental setup and software requirement specification.Section 3 describes Program-mers design input to the system and outcome of the system.Section 4 is about results and validation of outcomes and section 5 describes conclusion.

2. Related Work

2.1 Litearature survey:

Traditional K- means algorithm is traditional partitional algorithm and can be used as clustering algorithm.and it is most popular algorithm because of its simplic-ity. Data Mining Can be Used in Various business application for different purposes such as decision support system,customer retention strategies,selective marketing,businesss management,user profile analysis to name a few [1],[2],[3],[10].

The premise of data mining is that it is necessary to establish a customer-information data warehouse which consist of customer data and will guide the busi-ness managers for designing better customer strategies as wee as better customer retention strategies.

Data in digital business firm is increasing exponentially in todays electronic n fast world.hence,to manage such data,market segmentation is solution.customer re-tention strategies [5],allocation of advertising resources and increase in profit margins are included in market segmentation[2],[5].

For that a comparisons of different characteristics of different segments(group) is to be done so as to identify important attributes of each group.Thia can be done using a good analytical tool.This tool will gives opportuninty for targetting customers from different locations,purchasing different products,having numerous different attributes and so on.Cluster analysis can yield the same result as mentioned above.clustering is nothing but ”grouping similar kind of data together”.Using clustreing and the analytical tool proper interrelationships between data point s can be found out.

Existing system uses traditional standard clustering algorithm such as K-means algorithm[6].It is most commonly used traditional method for categorizing or clustering data and it is exclusive clustering algorithm.K-means uses concept of unsupervised learning.In K-means algorithm each centroid of the cluster is assigned by each data point in the data set to which ever centroid is nearest.The centre or centroid is computed by averaging all points in the cluster.Centroid has co-ordinates in the space and that co-ordinates are computed by taking arithmetic mean for each dimensions separately for all data points in the specific cluster[6].

K-means algorithm is simple and having high pro-cessing speed even though it is applied to large data sets but has significant disadvantages as discussed in [8][4].

K-means algorithm is simple and having high processing speed when applied to large amount of data.k-means calculates centroid of the clusters by averaring the data points in the data set.

Limitations of traditional K-means methods such as output of k-means algorithm depends upon depends upon order in which input are given and K-means tends to result in local mininmum and limited applicability to to only the data set consisting of isotropic clusters

.When K-means algorithm is used thre is need to adjust the sample space continuously .so,we should calculate the centres of the clusters again and again.If data set applied to to the K-means algorithm is large then more time is consumed by algorithm to produce the clusters. So more fast algorithm is required and there is need to improve efficiency of K-means algorithm.In K-means method while assigning each data object in the data set D to the cluster centres which are initially assigned some calculation regarding distance of the data object from the centre of cluster which was randomly assigned initially is to be made but if the distance of the data object is far from the centre then there is no need to measure the exact distance between that data point and centre of the cluster in order to know that that point should not belong to the respective cluster. This was case for one data point as data set is large then such dis-tance calculations are unnecessary in K-means method .and these redundant calculations should be avoided in order to increase processing time of the K-means method.

2.2 Related algorithm

We are using optimized version of K-means algorithm.Hierarchical algorithm,OKH algorithm .all the algorithm are as follows: Optimized K-means Algorithm:

We use triangle inequality theorem to reduce redundant calculation in K-means algorithm.The triangle inequality theorem says that the sum of two sides of the triangle is always greater than the remaining side of the triangle.this theorem is applied to Euclidean space[9] and can be extended to multidimensional Euclidean space.
Take three random vectors in Euclidean space
:p,q,r Then
Then
d(p,q) + d(q,r) »= d(p,r)
d(q,r) - d(p,q) «= d(p,r)
d(Ai,Aj) is the distance between the two cluster centres namely i,j.
if
2d(p,q)«= d(Ai,Aj) then
2d(p,Aj) - d(p,Aj) «= d(Aj,Ak) - d(p,Aj)...... (1)
Then according to above equation (1)
d(p,Aj) «= d(p,Ak)

The optimized K-means algorithm is as follows:
1)select centres for clusters and set lower bound y(p,f)=0 for each data object as wellas for centre of the cluster. 2)Now start assigning each data object to centre of clus-ter by computing distance between cluster centre and data object[3].we can use previously obtained data to avoid redundant calculations.each time d(p,f) is calcu-lated and set its value to y(p,f).
Hierarchical Algorithm:

As the the name of the algorithm suggest the output of hierarchical algorithm[14] is nothing but hierarchical structure formed using data entities.Single link or complete link hierarchical algorithm can be used.Consider the example below.With this a we can get different hierarchical structure with different similarity requirement .From fig.2.1 if the similarity requirement is set at

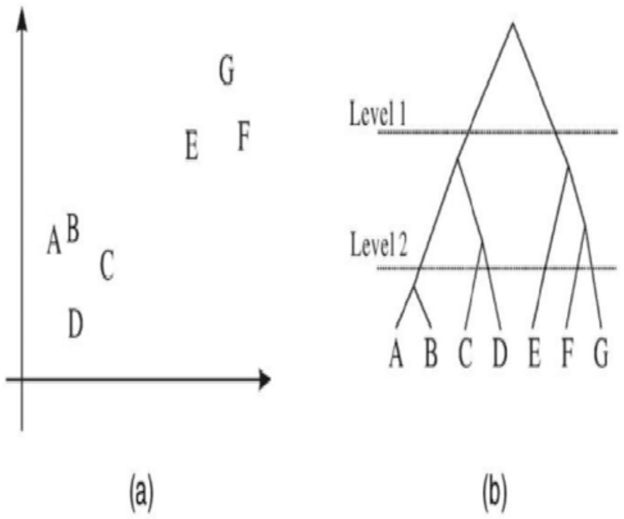

Figure 1: Level 1Illustrative hierarchical clustering results for a data set of seven points.(a)The input(b)A possible Hierarchical tree

level 1 , the input data set is partitioned into different clusters. As given in the above figure A,B,C,D,E,F,G are data objects given.when similarity requirements are given at level 1,the input data objects are divided into 2 clusters i.e(A,B,C,D) and the (E,F,G).when similarity requirement is set at level 2 then the data objects are divided into six clusters i.e.(A),(B),(C),(D),(E),(F) and(G).Most hierarchical clustering algorithm are varia-tions of the single-link and complete linkage hierarchical algorithm.By good quality of clustering output.

The outline of a general hierarchical algorithm steps : Singe link or complete link hierarchical algorithm can be used.

1)each data point form cluster .

2)and algorithm merges to closest cluster

repetitively. 3)Hierarchical structure is outputted.

OKH Algorithm:

This algorithm is nothing but combination of optimized k-means algorithm and hierarchical aclustering algorithm.

Algorithm OKH:

Input:data set D containing data objects, size of data set s,number of sub clusters m,desired number of cluster n

Output: hierarchical structure of n clusters.

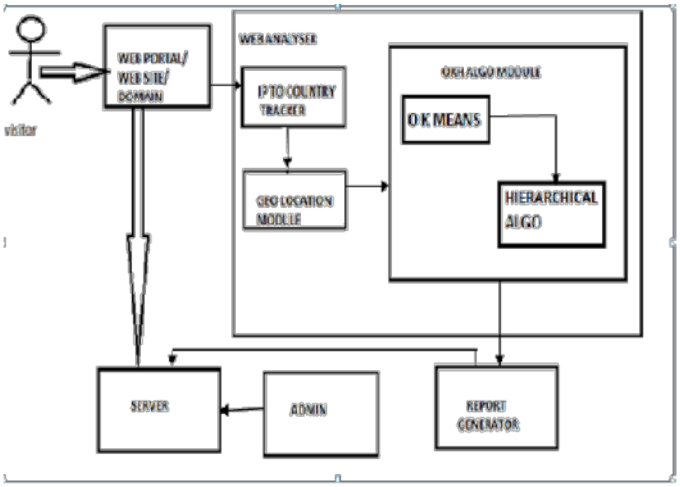

Figure 2: Archictecture of web user analyser

1)apply O K-means on input data set to get m subcluster.
2)Apply single link algorithmon m subcluster to get n cluster based on some similarity.

Experimental setup including SRS :-

The proposed system is an Analytical tool developed for online shopping web portal. we the system builders are third party who are going to develop a system that will help a particular domain to use its users data in order to expand their business. developers are located at the server site of the domain.The users Of the domain can be anywhere in the world. We are going to collect data of users of particular domain ,then we are going to apply clustering algorithm on it and are generating reports according to purachase,enquiries,browsing de-tails of users and then are sending reports to particular domain so that domain owner can use these reports for deciding business expansion and customer retention strategies. Instead of manual surveying the system is providing domain with ready reports for deciding business strategies,as the existing system uses K means algorithm for intended purpose

The experimetal setup of ovarall system is as shown in figure 1.
Instead of manual surveying the system is providing domain with ready reports for deciding business strategies,as the existing system uses K means algorithm

for intended purpose and its performance degrades when subjected to large datasets.proposed system is efficient, fast and performs well when subjected to large datasets as it is using new improvised optimized k means hierarchical algorithms.

Normal Requirements:

They are also called as basic requirements. they are explicitly expressed by customers and may be simply placed into the SRS.Following are some of the normal/basic requirements.

NREQ1:

Description : System can be used by more than one user at the same time.

Other factors affecting the requirement:access control/priviledges given by admin to different users.

NREQ2:

Description : System Is Supposed To Connected To Internet Through Wi-Fi Or Dial Up Or Broad Band Connection. Other factors affecting the requirement: Processing Speed

NREQ3:

Description: System Is Supposed To Work For Users Who Have Moderate Amount Of Computer Experience(E.G.Excel,Analytical Tools Etc.)

Other factors affecting the requirement: Versions Of Softwares(Excel 2003/2007)

NREQ4:

Description : Backup Is To Be Maintained At Regular Time Intervals So As To Avoid Loss Of Data During Crashes.

Expected Requirements

These are implicit but expected requirements . they are regarded by customers as obvious and rarely are placed in the first version of the SRS, however if they are not fulfilled, the customers are dissatisfied.Following are Expected Requirements: EREQ5:

Description: Owner of the domain as well as all authorized persons are allowed to use the output(Reports) of the system.

Other factors affecting the requirement: access control/priviledges given by admin to different users.

EREQ6:

Description : System can be run on any platform

EREQ7 :

Description : Should Be Able To Handle Defined Amount Of Data Approx. Number Of Records Required

Initially Plus Anticipated Growth Eg 1,000 Customers Increasing At 100 Per Year; 10,000 Jobs Increasing At 1,000/Year; Etc.

EREQ8:

Description : System Should Be Maintainable. Maintainability/Future Expansion. Who Will Maintain The System In Future (Eg In-House), Any Plans For Future Developments, Etc.

Other factors affecting the requirement: Backup Facility During System Crash/Failure.

EREQ9:

Description : Source Code Should Be Provided So As To Maintain System In Future Or For Future Expansion. EREQ10:

Description : System Should Be Reliable.It Should Have Downtime As Minimum As Possible..

Other factors affecting the requirement: Input Data Subjected To The System For Processing.

2.1.3 Excited Requirements

These are implicit and unexpected requirements. the customers may be not aware of them, but if they are fulfilled, the customers even may be excited.

XREQ11:

Description : System can be used by more than one domain concurrently.

Other factors affecting the requirement: Contents Of Domain And Server Location.

XREQ12:

Description : This System Need High Level Of Security

(Eg Different Group With Different Access Rights)Is Required.

XREQ13:

Description : External Communications Should Be Involved. Eg Automated Faxing, Import Csv Data , Export Data To Particular System, Etc.

XREQ14:

Description : The Deliverables Should Contain Help Files,Manuals,Support Files,Source Code, Training To Use The System Efficiently Etc. Other factors affecting the requirement: Extra Cost Required.

XREQ15:

Description : Existing data should be converted to the format of the current reports from proposed system. Other factors affecting the requirement: Extra Amount Of Work Needed And Extra Time Needed.

Software Requirements: -

Operating System: windows
2x Database: sql server 2005
Language of Implementation: ASP .net,C
sharp H/w Requirements:
Hard disk:80gb

Ram:512 Mb

Processor:2.1GHz

3. Programmer's design

3.1 Input to the proposed system:
Input to the the proposed system is nothing but the the various user visiting the website.so inputs are nothing but user navigation on web portal or web logs maintained by the server of the website . 3.2 Outcome of the system:
Ouput of the system is nothing but report generated by the report generator from clusters formed by applying OKH algorithm.as well as consisting details of the customer along with items he has viewed as well as items he has purchased .The proposed system consists of following components.each model does its work mentioned below.

A)Module for Admin :Whenever user visits a particular web his details (eg.ip,login detais or account details)are sent to server of the website .Admin module is used to give access control privileges to the different users of the system.. Admin performs functions like creation, deletion. updation of account, restricted access control to other intruders or users.
B)Server Of Domain:Server of domain is nothing but the server of the web portal/wesite or domain for which the system is to be implemented.The server serves the user request.
C)IP to Country tracker: : Whenever user visits the domain .all the details of the user are tracked by the server(e.g. IP address,zip code,country code etc.) When a visitor requests a web page that contains tracking code, the visitor's browser will read this tracking code and call a JavaScript file installed on the web server. Once called, the JavaScript file will open a data collection session and set (or reset) a one-year, first-party cookie in the visitor's browser. This cookie will be stored in the visitor's browser until it expires after one year, or until the user deletes it.
D)Module for Geo Location : Apply this characteristics

on Google maps API for getting its geographical location (Exact longitude and lattitude) And the output of that API is geographical representation of the user location Geo-Location Module helps to locate visitors geographical location using inputs like longitude and latitude of the user from user details for locating him.

E)Module for OKH algo: On that data values OKH is applied i.e. at first a hierarchical algo is applied then optimized K means algorithm is applied to produce clusters of these data values.

3.1. Mathematical Model

Web User Analysis Using Hierarchical And Optimized K-Means Algorithm For Online Market Analysis is of P Class because:

1. Problem can be solved in polynomial time.
2. GA always produces strong feature set.
3. Optimized K means hierarchical algorithm in this case gives us optimum solution.

Let S be the set of visitors, ip addresses, clusters, clustering reports S = I,R,T,M where I represents the set of visitors who visits domain,which are input to web analyser , R represents the details of visitors which are input to the clustering algorithm and T are clusters which are are generated as output of analyser as clustering output and M be the set of reports for respective clusters.

I = (I1, I2, I3,...) R=

(R1, R2, R3,...)

T=(T1,T2,...)

M = (M1,M2,...)

Input is mapped to output which is shown in the following vein diagram:

3.2. Dynamic Programming and Serialization

In proposed approach various dynamic programming as-pects are used.detail is as follows:

The whole sysyem divided into four parts according to its functionality.first part retrives user data from navigation of user on web portal, all the data are applied to

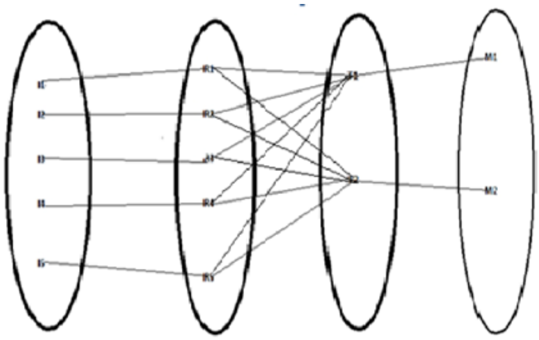

Figure 3: Venn Diagram

The first phase is again divided into subphases .To find location of the user longitude and latitude of the user is retrieved by the system then according to values ofg latitude and longitude city of the user will be decided. By applying those deatails to geo-location API.hence it makes use of optimal substructure approach of dynamic programming. With this technique problem is reduced and memorization is achieved.

In OKH phase Branch and bound method can be used for construction of clusters and for outlier detection.

3.3. Data independence and Data Flow architecture

the OKH algorithm to get te desired result.

A DFD allows you to identify the transformations that take place on data as it moves from input to output in the system.

6

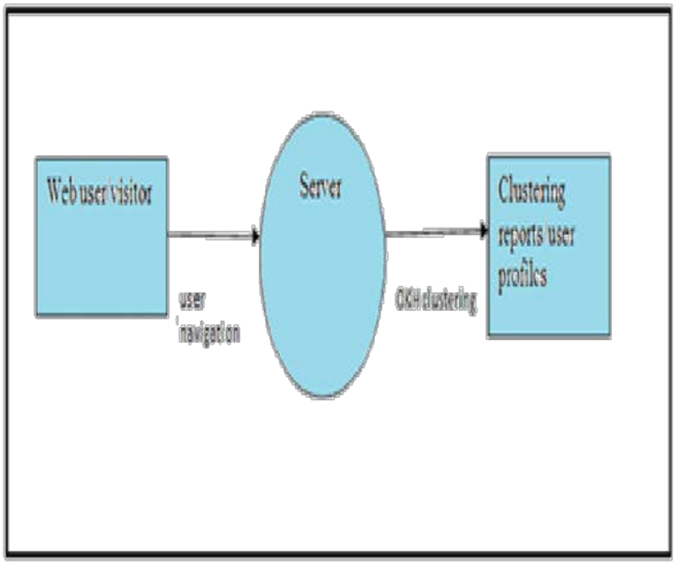

Figure 4: Data Flow Diagram

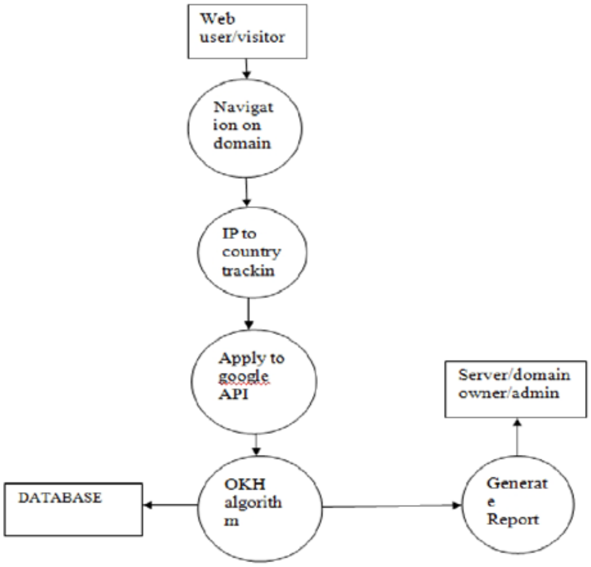

Figure 5: Level 1 Data Flow Diagram

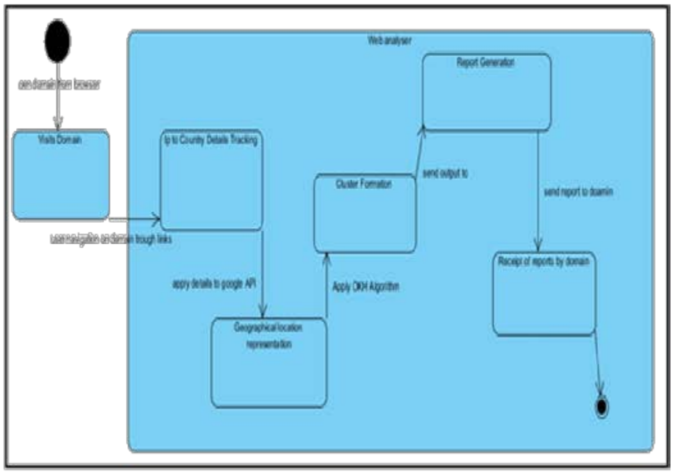
Web analyser
open domain from browser
Visits Domain
user navigation on domain trough links
Ip to County Details Tracking
apply details to google API
Geographical location representation
Apply OKH Algorithm
Cluster Formation
send output to
Report Generation
send report to doamin
Receipt of reports by domain

Figure 6: State Diagram of web user analyser

3.4. Turing Machine

Current State	Description	Output	Next state
S1	Visiting domain	Web logs are tracked by domain	S2
S2	IP to country tracking	Details of user from ip addresss to country are tracked	S3
S3	Geo-location	Geographical Details of user are represented on maps	S4
S4	Cluster forming by apply-ing OKH algorithm	Cluster formed	S5
S5	Report generator	Report gen-erated as per user require-ment	S6
S6	Receiption of report by owner of web portal	Admin can view details reports	[End State]

4. Results and Discussion

The performance of the proposed approach can be ana-lyzed using five fold or Tenfold cross validation technique. For measuring performance of the proposed approach de-tails of input to system is as follows

ID	List of Input
1	Browsing from the server located in Maharashtra
2	Browsing From Machine with IP 117.228.80.124
3	Browsing from machine with IP 10.147.1.0, 203.88.23.33

And the corresponding output is as follows:

IID	Output
I1	117.228.80.124 INDIA PUNE MAHARASHTRA IN 18.5196 73.8553 +05:30
I2	14.97.245.36, INDIA, MUMBAI, MAHARASHTRA, IN , 19.0144, 72.8471, +.5:30
I3	14.140.40.14 INDIA ,PUNE, MAHARASHTRA ,IN, 18.519 ,73.8553,+05:30

The final output can be validated by comparing quality of the outputs of the K-means algorithm as well comparing time taken by algorithms to process same amount of data.

5. Conclusion

In this electronic information era,digital enterprises should make more or full use of information available from various information resources and instead of using product centric management model should use customer centric management model.so,business owner should focus on building customer information data warehouses which will support business managers in deciding marketing strategies,customer retention strate-gies.operations as well as production strategies.

If the proposed system traces which application of the website requested by visitors then business intelligence takes place.With this real time clusters can be formed for every click of customer and it will generate reports in terms of different charts regarding the real time data..These different charts will reflect that the owner of the website should concentrate on which application (product)of the website.

This application is designed by making use of improved (optimized)standard K-means algorithm .this optimized algorithm makes use of triangle inequality theorem and incorporating hierarchical clustering algorihm for robust data.this approach applies optimized algorithm on visitors data and applies hieararchical algorithm onto the products they requested.the proposed system is surely efficient than that of standard K-means Method.This system helps to identify target markets as well as customer density and potential customers of website or particular product,hence helps in target marketing.

After implementing the said algorithm time is saved for reformulation of the clusters.

References

[1] A.G. Buchner and M. Mulvenna, "Discovery Internet Marketing Intelligence through Online Analytical Web Usage Mining," Proc. ACM SIGMODâĂŹ98,vol.27,no.4,PP.54-61,dec 1988

[2] M.S.Chen,J,Han and P.S. Yu,"Data mining âĂ An Overview from database perspective ",IEEE trans Knowledge and data engineeringvol5,n0.1,PP.866-883,Dec.1996

[3] U.M.Fayyad,G.piatetsky-shapiro,p.Smyth and R.Uthurasamy"dvances in Knowledge Discovery and data mining"Cambridge,Mass:MIT ,1996.

[4] Mrs.G.P.Dharne ,Mrs.S.A. Kinariwala ,Mrs.A.S.vaidya,MS.P.V.Pandit"A web user nalyser by hierarchical and optimized K-means algotrithm",vol.1,issue7,dec.2011

[5] Google Latitude Website http://www.google.com/latitude/intro.html

[6] Wanghualin, âĂIJData Mining and Its Applications in CRMâĂİ, 978-0-7695-4043-6/10 2010 IEEE DOI 10.1109/ICCRD.2010.184

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

[8] Xiaoping Qin, Shijue Zheng , Tingting HeMing Zou, Ying Huang, "Optimizated K-means algorithm and application in CRM system"2010 International Sym-posium on Computer, Communication, Control and Automation

[9] G.P.Mohole,S.A. Kinariwala "Market user analyser Using OKH algorithm"International journal of Com-puter application(IJCA)2012.

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

[11] Xiaoping Qin, Shijue Zheng , Tingting HeMing Zou, Ying Huang, "Optimizated K-means algorithm and application in CRM system" 2010 International Sym-posium on Computer, Communication, Control and Automation

[12] Zhe zhang,"Data Mining and Its Application in Cus-tomer Relationship Management", Shanghai: Fudan University Press, Aug.2007

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

[14] K.S.Beyer,J.Goldstein,R.ramakrishnan and U.ShaftâĂİWhen is nearest neighbor meaning ful?âĂİProc.International conference Database theory(ICDTâĂŹ 99)pp.217-235,1999