Selecting type of clusters that are most appropriate for capturing overlapping interests of different types of users in personalization tasks using Web Usage Mining

Sanjay B. Thakare, Sangram Z. Gawali

Information Technology, Bharati Vidyapeeth University
College of Engineering, Pune-43, Maharashtra, India
mail2sbt@gmail.com
gsangram@gmail.com

***Abstract*— Clustering is one of the most important phase of the web personalization, so selecting proper types of cluster to justify the user interest for creating user profile is again very important. In this paper, we address the problem, which types of cluster are most appropriate for personalization task and how to identify and make use of such cluster in personalization. If user interest is overlapped amongst the cluster then which one or others are most appropriate for users likes. We have to find the parameters that are required to select the clusters that potentially capture overlapping interests of different types of users.**

***Keywords*— Clustering, Personalization, Web Usage Mining**

I. INTRODUCTION

With the explosive growth of information sources and information available on the World Wide Web, it has become increasingly necessary for data miners to develop efficient and effective automated tools to find the desired pattern. Web Mining is the extraction of interesting and potentially useful patterns and implicit information from source or activity related to the World Wide Web. There are roughly three knowledge discovery domains that related to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web Usage Mining also known as Web Log Mining is the process of extracting interesting patterns in web access logs. Web Usage Mining analyses the usage patterns of web sites in order to get an insight of site to understand the user's interests and requirements. This information is especially valuable for E-Business sites and other applications in order to achieve improved customer satisfaction.

Cluster analysis classifies a set of data into two or more mutually exclusive unknown groups based on combination of interval variables. The purpose of cluster analysis is to discover groups where member of the groups share common properties. A clustering is a process which create set of clusters which share the common properties. Important distinction between hierarchical and partitional sets of clusters will discussed here. Partitional Clustering is a division data objects into non-overlapping subsets (clusters) such that each data object is in exactly one subset. Hierarchical clustering is a set of nested clusters organized as a hierarchical tree.

II. PROBLEM DEFINITION

Let T be the set of web access transactions T={ t1, t2, t3,……..,tn} and U be the m distinct users U={u1, u2, u3,……um}. Let C be the set of k distinct cluster of the transaction which will help us to form m distinct user profile but it not happen in that way. The user interest is overlapped on many clusters. Consider the a user ui, for which set of the transactions are Ti={t1 ,t2,.…,ti} are classified over three cluster c1,c2 and c3 as shown in figure below.

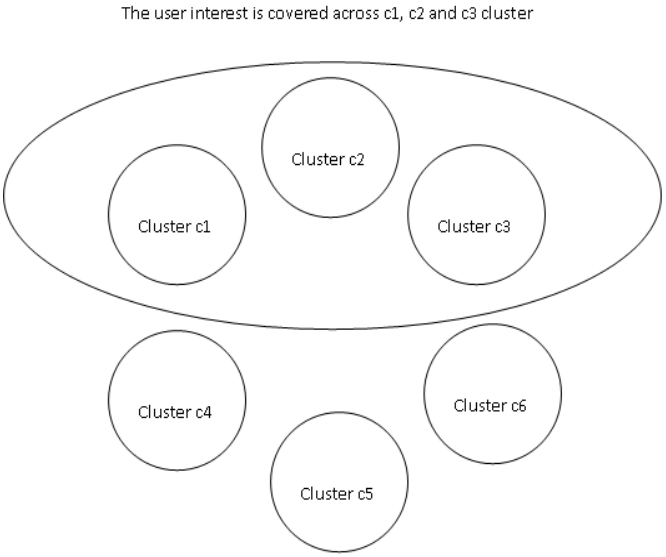

Fig. 1. User interest is covered across c1,c2 and c3 cluster

As the above figure shows that the user interest or like is scattered over three clusters. Now we have to select the proper cluster that perfectly match to the user likes. The problem is that how to select the cluster and how many cluster that represent the user likes.

III. PURPOSED SOLUTION

A. How to select the cluster

Now we assume that the cluster have centroid that represent the cluster. Let T be the user transaction and C be the centroid of the cluster.

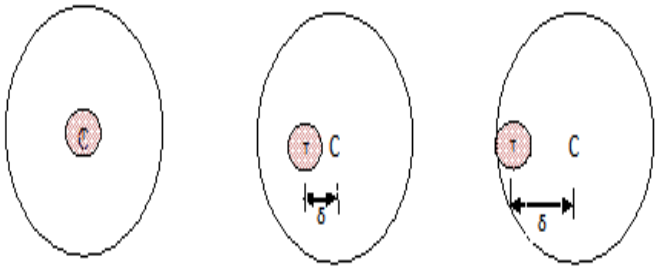

Fig. 2. a) Very close to cluster b) at moderate distance from centroid c) at edge from centroid

δ (deviation) is calculated in the following manner:

1] Distance Measurements between Data Points

The Euclidean distance function measures the 'as-the-crow-flies' distance. The formula for this distance between a point C (***X1, X2***) and a point ***T*** (***Y1, Y2***) is:

$$\delta = \sqrt{\left(\sum_{i=0}^{n} (xi - yi)^2\right)}$$

a. Deriving the Euclidean distance between two data points involves computing the square root of the sum of the squares of the differences between corresponding values.
b. The following figure illustrates the difference between Manhattan distance and Euclidean distance:

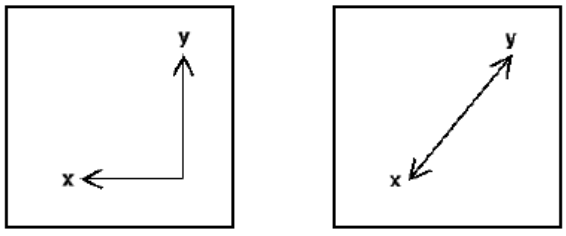

Fig. 3. a) Manhattan distance b) Euclidean distance

This parameter specifies how the distance between data points in the clustering input is measured.

a. Euclidean: Use the standard Euclidean (as-the-crow-flies) distance.
b. Euclidean Squared: Use the Euclidean squared distance in cases where you would use regular Euclidean distance in Jarvis-Patrick or K-Means clustering.
c. Manhattan: Use the Manhattan (city-block) distance.
d. Pearson Correlation: Use the Pearson Correlation coefficient to cluster together genes or samples with similar behavior; genes or samples with opposite behavior are assigned to different clusters.
e. Pearson Squared: Use the squared Pearson Correlation coefficient to cluster together genes with similar or opposite behaviors (i.e. genes that are highly correlated and those that are highly anti-correlated are clustered together).
f. Chebychev: Use Chebychev distance to cluster together genes that do not show dramatic expression differences in any samples; genes with a large expression difference in at least one sample are assigned to different clusters.

2] Distance and Weight Combination

Some time it happens that the least perceived pages are close to centroid; in that case we have to concentrate on the weights to avoid this problem. The weight is the average time taken by the user to see that particular page. Let's consider V is average time taken by the user to see the page and α be the threshold value.

$$V = \begin{cases} 1 & V \geq \alpha \\ 0 & V < \alpha \end{cases}$$

$$\delta = V * \sqrt{\left(\sum_{i=0}^{n} (xi - yi)^2\right)}$$

B. How may cluster required to select

One needs to select the cluster that match the users interest for creating users profile. The user interest is overlapped over more than four cluster then we have to put such user in the general category. The minimum number of the cluster required to create the user profiles should not exceed than three. One can combine the selected cluster with common start, common end or have matching substring in the access stream.

IV. CONCLUSION

One of the most significant challenges for web usage mining research for personalization are personalization solutions must effectively leverage user profiling, adaptive web publishing and privacy considerations. We believe that efforts toward personalization should focus on smaller goals initially until technology, policy and user acceptance of various techniques and methods reach a level where more

sophisticated personalization features are possible and delivered to the user. Ultimately the user should happy with the effort taken by the web usage mining researcher and this would be the best reward for their work.

The knowledge on the internet can deliver on its promise to become a personalized information portal that will enrich its users in new and unforeseen ways. This system will be effective and must be nurtured by a well-designed and paced process that can build slowly and scale on each new innovation.

REFERENCES

[1] Jianxin Wang, Reducing the Overlap among Hierarchical Clusters with a GA-based Approach, 2009.
[2] D.Vasumathi, A.Govardhan, K.Suresh, Effective Web Personalization Using Clustering, 2009.
[3] R. Forsati, M. R. Meybodi, A. Ghari Neiat, Web Page Personalization based on Weighted Association Rules, 2009
[4] Minxiao Lei, Lisa Fan, A Web Personalization System Based on Users'Interested Domains, 2008.
[5] B.Bahmani Firouzi, T. Niknam, and M. Nayeripour, A New Evolutionary Algorithm for Cluster Analysis, 2008.
[6] Norman A. Graf, Clustering Algorithm Studies, 2000.
[7] Bamshad Mobashes, Robert Cooley, Jaideep Srivastava, Automatic Personalization Based on Web Usage Mining , 1999.