Volume: 05 Issue: 01 Pages:1842-1846 (2013) ISSN : 0975-0290

Modified Agglomerative Clustering for Web Users Navigation Behavior

T. Mekala[1] P. Nandhini[2]
[1]*Assistant professor*
CSE Dept, M.Kumarasamy College of Engineering- Karur, India
[2]*Program Analyst*
Cognizant Technology Solutions-Tamilnadu, India
[1]mekalathangavel@gmail.com
[2]nandhinime@yahoo.co.in

ABSTRACT

-Clustering is the task of finding natural partitioning within a data set such that data items within the same group are more similar than those within different groups. At present all available clustering techniques for web usage mining are based upon usage patterns derived from user's page preferences. To improve the efficiency of usage pattern, the similarities in user's access behavior based on time locality of user's navigational acts are used. The tuning algorithm defines cluster that show similar visiting behavior at the same time period and users page preferences by varying the priority given to the time visit.

Keywords- Web mining, Web users clustering, Navigation, Access time.

Date of Submission : May 31, 2013 Date of Acceptance : July 15, 2013

I. INTRODUCTION

Interest in the analysis of user behavior on the Web has been increasing rapidly. This increase stems from the realization that added value for Web site visitors are not gained merely through larger quantities of data on a site, but through easier access to the required information at the right time and in the most suitable form.

Web Mining is the extraction of interesting and potentially useful patterns and implicit information fro m artifacts or activity related to the World Wide Web. Web usage mining provides the support for the web site design, providing personalization server and other business making decision, etc. In order to better serve for the users, web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users' visiting characteristics, and then extracts the users' using pattern[l].

According to the differences of the mining objects, there are roughly three knowledge discovery domains that pertain to web mining: Web Content Mining, Web Structure Mining, and Web Usage Mining. Web content mining is the process of extracting knowledge from the content of documents or their descriptions. Web document text mining, resource discovery based on concepts indexing or agent; based technology may also fall in this category. Web structure mining is the process of inferring knowledge fro m the World Wide Web organization and links between references and referents in the Web. Finally, web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns
in web access logs.

A. Web Usage Mining

Web usage mining, also known as Web Log Mining, is the process of extracting interesting patterns in web access logs. Web servers record and accumulate data about user interactions whenever requests for resources are received. Analyzing the web access logs of different web sites can help understand the user behavior and the web structure, thereby improving the website design. Log record has lots of useful information such as URL, IP address, time and so on. Analyzing and discovering log could help us to find more potential users of the web site and trace service quality of the site. The large majority of methods that have been used for pattern discovery from Web data are clustering methods. Clustering has been used for grouping users with common browsing behavior.

Web users clustering [2] is to use web access log files to partition a set of users into clusters such that the users within a cluster are more similar to each other than users from different clusters. The discovered clusters can then help in on -the-fly transformation of the web site content. In particular, web page scan be automatically lin ked by artificial hyperlinks. The idea is to try to match an active user's access pattern with one or mo re of the clusters discovered from the web log files [3]. Pages in the matched clusters that have not been explored by the user may serve as navigational hints for the user to follow the aspects based on a weight factor. This paper we highlight the fact that grouping Web users based on their navigational

Volume: 05 Issue: 01 Pages:1842-1846 (2013) ISSN : 0975-0290

II. EXISTING SYSTEM

More specifically, we define three different user visiting structures in order to capture all aspects of interrelations in page and time visiting. A vector is used to represent the frequency of a user's visits to particular pages (with no information about the time of visits) while a second one records the frequency of the user's visits at particular timeframes (with no information about which page was visited). Moreover, the lack of the complementary information in each of these vectors, motivated us to define a table which will incorporate the overall information (seen as a set of vectors). In particular, this table represents the frequency of visits to particular pages incorporating the exact knowledge about the timeframes of these visits. These structures are summarized next:

A. Time Visiting Vectors

A time visiting vector TV (i, :), where i = 1 . . . n, represents a user's accessing behavior with respect to time (timeframes). It is also a multivariate vector consisting of t measurements:

$$TV (i, :) = (TV (i, 1). . . TV (i, t)) \tag{1}$$

Where the TV (i, l) element, l = 1, . . . , t, indicates the number of times the user i visits the whole site (all the p pages) during the l timeframe. All the TV(i, :) vectors are organized in the two dimensional n x t users' time visiting table TV.

For example, in Fig. 1(b), which depicts the table TV, the fact that TV (2, 2) = 9 means that the user identified as 2 has made 9 visits to the whole site during the timeframe 2.

Table 1 Time visiting table

	Timeframes	
Users	2	4
	8	9
	6	10

B. Clustering Phases

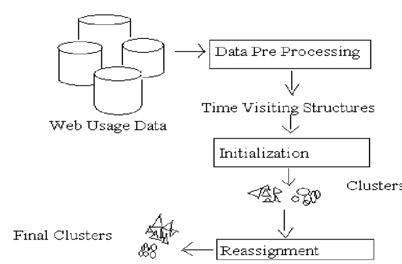
This time related clustering algorithm is an unsupervised hard partitioned method and it is used to minimize the objective function E_t(6).

Normally in this clustering phase there are two steps i.e., initialization and reassignment behavior should be faced as a twofold problem that will: (i) deal with the different users' page preferences and (ii) identify the time dependencies involved in the usage navigational patterns. Thus, the problem that has to be addressed should combine the above two criteria namely the users' page preferences i.e. the page aspect and the time their visits were logged i.e. the time aspect. Since the proposed approach aims at advancing the earlier ones (which considered only the page preferences).

Thus, we adopt two algorithmic approaches that differ in their initialization step and tune the two aspects based on a weight factor. The first tuning approach initiates with the page preferences and then proceeds to the time aspect while the second one follows the reverse logic.

1) Initialization: K-means partitioned clustering algorithm is used to produce the k clusters. K-means algorithm is: given n points to be clustered, a distance measure d to capture their dissimilarity and the number of clusters k to be created, the algorithm initially selects k random points as clusters' centers and assigns the rest of the n − k points to the closest cluster center (according to d). Then, within each of these k clusters the cluster representative (also known as centroid or mean) is computed and the process continues iteratively with these representatives as the new clusters' centers, until convergence (8).

Figure 1 Clustering Process



In this framework, given the n users and the number of k clusters are to be created. In time related algorithm the clustering considers the time aspect via time visiting structure (i.e. TV table) and uses dt as users' dissimilarity distance measure.

2) Reassignment: The reassignment step aims at producing a CL∗ clustering which enhances the initial CL to meet the two criteria of the time-aware problems. Given the initial CL clustering, it aims at finding a CL∗ that minimizes the objective function E_t. More specifically, the reassignment step begins with the set of k clusters produced by the CL and involves a number of iterations.

During each iteration, compute for each user ui the fluctuation of the value of the underlying objective function (i.e. E_t) caused by moving user ui to one of the rest k – 1 clusters. The reassignment phase follows K-means idea for its convergence(7), ending either after a number of iterations or when the objective function improvement between two consecutive iterations is less than a minimum amount of improvement specified.

III. PROPOSED SYSTEM

Our defined problems are of NP-hard nature since they are a generalization of the well -known clustering problem [4] and thus we can only aim for approximate solutions. Based on the previous section, we define two algorithms to solve the TUNING TIME-AWARE CLUSTERING problem (tuning algorithms). These algorithms adopt local search heuristics which are similar in spirit with the well-known K-means algorithm [5], which is used for our initial clustering setup. Although agglomerative does not provide approximation guarantees, it has been proved very effective in many practical problems.

A. Agglomerative Clustering

Hierarchical clustering algorithms are either top-down or bottom-up. Bottom-up algorithms treat each document as a singleton cluster at the outset and then successively merge (or agglomerate) pairs of clusters until all clusters have been merged into a single cluster that contains all documents. Bottom-up hierarchical clustering [9] is therefore called hierarchical agglomerative clustering or HAC. Top-down clustering requires a method for splitting a cluster. It proceeds by splitting clusters recursively until individual documents are reached. The agglomerative clustering includes three linkage criteria namely single, complete, average linkage. The hierarchy within the final cluster has the following properties

- Clusters generated in early stages are nested in those generated in later stages
- Clusters with different sizes in the tree can be valuable for discovery.

The most commonly used linkage method is single linkage.

1) Single Linkage: In cluster analysis, single linkage, nearest neighbour or shortest distance is a method of calculating distances between clusters in hierarchical clustering. In single linkage [10], the distance between two clusters is computed as the distance between the two closest elements in the two clusters. The distance measure used here is Euclidean distance.

Mathematically, the linkage function – the distance $D(X,Y)$ between clusters X and Y – is described by the expression

$$D(X,Y) = \min_{x \in X, y \in Y} d(x,y) \quad (2)$$

where X and Y are any two sets of elements considered as clusters, and d(x,y) denotes the distance between the two elements x and y.

Algorithm: Agglomerative Clustering

SIMPLE HAC($d_1,\ldots d_N$)
for n←1 to N
do for i←1 to N
do C[n][i]←SIM(d_n,d_i)
I[n] ←1 (keeps track of active clusters)
A←[] (assembles clustering as a sequence of merges)
for k←1 to N-1
do(I,m) ←arg
max $\{(i,m) : i \neq m \wedge I[i]=1 \wedge I[m]=1\}$ C[i][m]
A.APPEND(i,m) (storemerge)
for j←1 to N
do C[i][j] ←SIM(I,m,j)
C[i][j] ←SIM(I,m,j)
I[m] ←0 (deactivate cluster)
Return A

IV. EXPERIMENTAL RESULT

To obtain the efficient clustering results the data set used is web log of Amazon web access data.

The data set includes 98 numbers of instances, 3 attributes and 5 to 10 number of URLs. The three attributes taken are session ID, Time and Page preferences by users.

The result obtained is

Figure 2 Squared Euclidean Distance

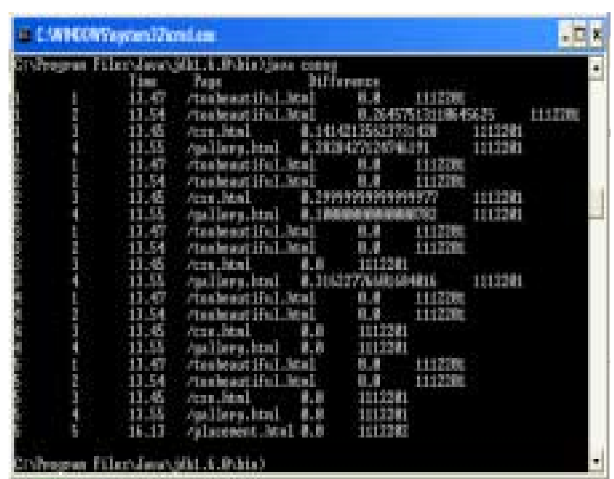


Figure 3 Agglomerative Clustering

Volume: 05 Issue: 01 Pages:1842-1846 (2013) ISSN : 0975-0290

The Figure 3 describes that Agglomerative clustering seems to be more effective to cluster the web usage data then k means algorithm

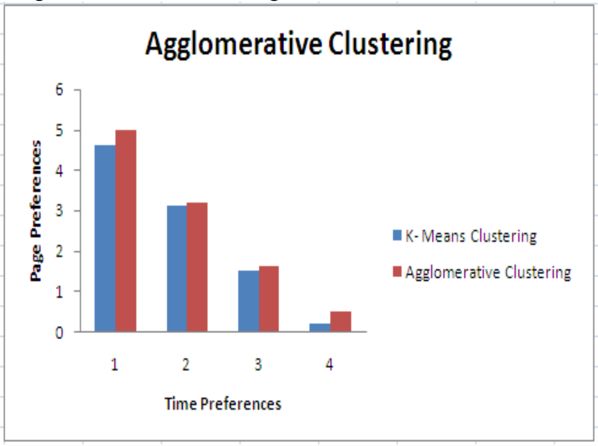


Figure 4 Comparative Result

V. CONCLSUION

This paper introduces and evaluates time aware clustering approaches: the so-called TUNING time aware clustering. The web access sessions are too complex to be converted into simple numerical features. In fact, the URLs in a website always have a hierarchical or tree-like structural directory. The future work includes comparison of same algorithm with various distance measures.

REFERENCES

[1] Qingtian Han, Xiaoyan Gao, Wenguo Wu, Study on Web Mining Algorithm Based on Usage Mining", Co mputer- A ided Industrial Design and Conceptual Design, 2008.CAID/CD 2008. 9th International Conference on 22-25 Nov.

[2] Yan, T.W., Jacobsen, M. Garcia Molina, H, Dayal, U.: From "User Access Patterns to Dynamic Hypertext Linking", proc. of Fifth International WWW Conference (1996)

[3] Rajni Pamnani, Pramila Chawan "Web Usage Mining: A Research Area in Web Mining" Department of computer technology, VJTI University, Mumbai

[4] M. Garey and D. Johnson, Computers and Intractability: "A guide to the Theory of NP-Completeness." New York, NY, USA: Freeman, 1979

[5] T. Hastie, R. Tibshirani, and J. Friedman, "The Elements of Statistical Learning: Data Mining, Inference, and Prediction." Springer, 2001

[6] A.Bianco, G.Mardente, M. Mellia, M. Munafo, and L. Muscariello, "Web user session characterization via clustering techniques," in GLOBECOM'05, IEEE, Dec. 2005, p. 6

[7] J. Srivastava, R. Cooley, M. Deshpande, and P. Tan, "Web usage mining:Discovery and applications of usage patterns from web data," SIGKDD Explorations, vol. 1, no. 2, pp. 12–23, 2000

[8] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets," in Proc. of the SIAM Int. Conference on Data Mining, Newport Beach, CA, USA, Apr. 2005, pp. 358–3

[9] W. Yang, Z. Wang, and M. You, "An improved collaborative filtering method for recommendations' generation," in *Int. Conference on SMC'04, IEEE*, Netherlands, Oct. 2004, pp. 4135 – 4139

[10] Y. Zhao and G. Karypis, "Topic-driven clustering for document datasets," in *Proc. of the SIAM Int. Conference on Data Mining*, Newport Beach, CA, USA, Apr. 2005, pp. 358–369

Authors Biography

T.Mekala completed her ME CSE in Bannari Amman institute of Technology, Erode in 2012 and completed her BE CSE in PSNA College of Engineering and Technology, Dindigul in 2010. Her interested areas are cryptography (crypto- Biometric), Network Security and Data Mining.

P.Nandhini completed her ME CSE in Bannari Amman institute of Technology, Erode in 2012 and completed her BE CSE in M.Kumarasamy College of engineering , Karur in 2009. Her interested areas are Data Mining, Network security.