ISSN:0975-9646

Analysis and Individualization of Clients concern based on the Internet Browsing History

V.Ravindranadhan, P.Sarath, M.Gopi Mahesh, M.Rajasekhar K.V.D.Kiran
Vaddeswaram, K.L.C.E, Andhra Pradesh, India

***Abstract*--This paper evaluates the interests by mining the internet browsing history. To count the visiting information of the interests, visiting time and regularity of groups. The idea of this paper is that the information of web page is accurate by using an algorithm called grouping related refined segments (GRRS) and combining the rules of association and time series analysis for individualization of one's search.**

Keywords—mining, webpage, interest, associationrules and time series analysis.

I. INTRODUCTION

In today's epoch of information explosion, Internet and World Wide Web are growing exponentially. It is progressively more difficult for users to locate what they need most in such an ocean of information. "Data is affluent, knowledge is meager". In this case, one of the problems that people are concerned about is how to obtain constructive information accurately, fast and resourcefully. This is not only a particular user's problem in fact it is a universal problem. Web Page Browsing is one of the imperative ways for people to obtain information. Every user visits hundreds of web sites with an enthusiasm, all this browsing is done based on a particular interest of the users. Analyzing the internet browsing history helps to the development of personalization technology, using which the job of the user will become more productive and expeditious.

The preliminary task of this paper is to gather some information by drilling down and analyzing some web browsing record, the thorough analysis of the gathered data, results in some regularity conclusion. Secondly, the paper proposes an enhanced technique called GRRS based on HAC (Hierarchical agglomerative clustering) and k - means algorithm. By this technique, it classifies people by their visiting concern. This algorithm reimburses for k - means having to determine the classification number (K) in advance, because in most cases we don't know the exact classification number. It also overcomes the complexity of choosing the division and merge points in HAC, because once chosen wrong, it cannot be patched up, even get into vicious circle and get a terrible clustering results. Finally, the algorithm proposes the stability of user's main interests, that is, user's interests are stable by the increasing number of visiting.

II. REFINEMENT AND EXTRACTION FROM DATA LOG

According to a log of KLuniversity network center server in Vijayawada, it records the university campus network users' nearly three months of visiting information. A complete record forms for:

ID	LOG IN	CLIENT IP	WEB IP	WEB SITE	GROUP

Preprocess the data and remove some incomplete or insignificant part which is a part of data cleaning and refinement, for which many proven techniques are available. In this paper, visiting time, user IP and interests' groups are to be used to examine and process some necessary group, such as splitting some large groups and combine some small and similar groups. Weight processing is essential to some groups, which has large access amount such as Google. It has to avoid incomplete groups which are prominent but cover some less attention groups, such as law, cookery, religion and so on. Although their access amount is very little, it can be a sign of visitors' interests obviously, so we can multiply it by a coefficient as weights, to avoid them from other important groups covered.

III. TIME SERIES ANLYSIS FOR INDIVIDUALIZATION OF SEARCH

Extract required data from the log containing browsed history, such as number of users visited, web pages accessed and their groups, and access time etc over the period. We came across two valid points:

(1). we dig out three parts of log which contain a day, a week and a month likewise, to count the access amount of each group. Results are shown in Fig. 1 to 3.

(2). we also mine three parts of log which contain an hour, a day and a month likewise, to count the access amount of all groups. Results are shown in Fig. 4 to 6.

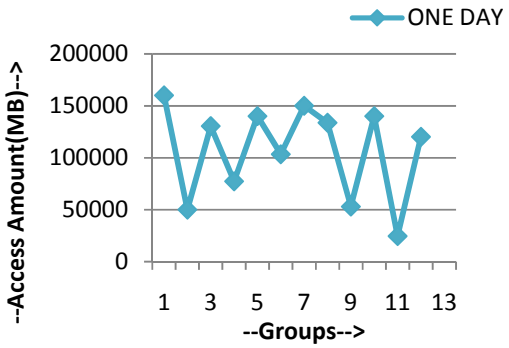

Fig 1. *The access amount of each group in one day*

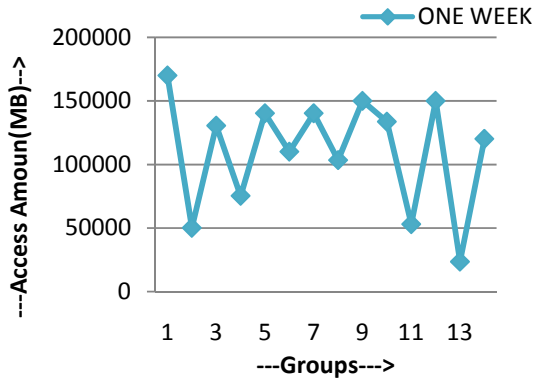

Fig 2. *The access amount of each group in one week*

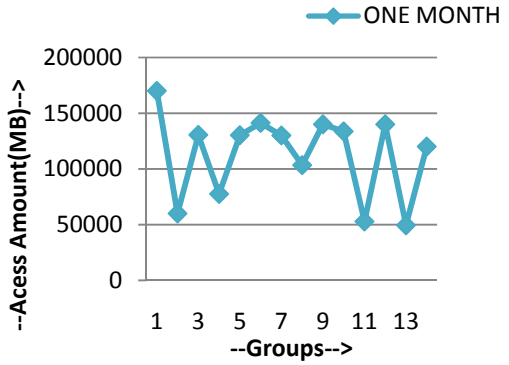

Fig 3. *The access amount of each group in one month*

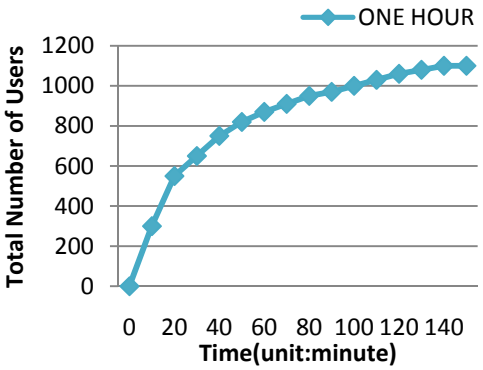

Fig 4. *The total number of users in one hour*

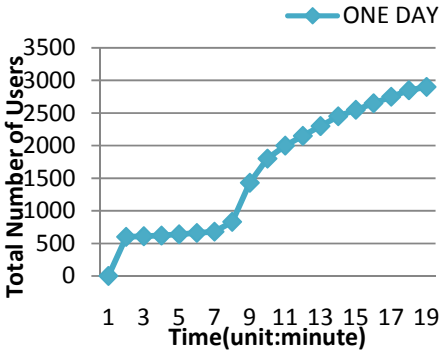

Fig 5. *The total number of users in one day*

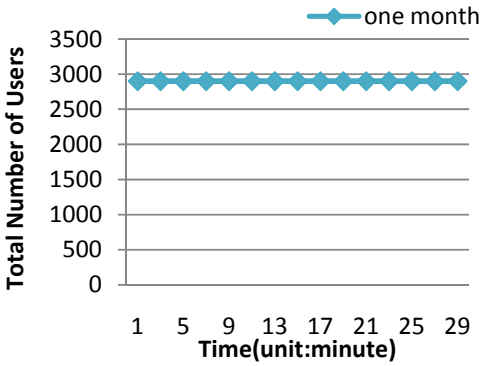

Fig 6. *The total number of users in one month*

We can get some conclusions from above figures.

(1). The relationship of the access amount of each group and the visiting time. As the visiting time increased,the access amount of each group is reaching linear increase.

(2). The relationship of the proportion of the access amount of each group and the visiting time. As the visiting time increased, the proportion of the access amount of each group is reaching stable value. When the visiting time increases to a certain extent, the change of the proportion is less, almost a stable value. Because the general people visiting internet every day is still stable, the total interests are tend to stability. In this data, search engine has been occupying most visited group. It has a very good greement with the habit of using the search engines.

(3). The relationship of the access amount of all groups and the visiting time. Generally, the total number of users is fixed, so with the growth of time, the growth in the users slows down, and finally become stable and remains unchanged,similar to the logarithmic growth.

IV. USING THE GRRS(GROUPING RELATED REFINED SEGMENTS) ALGORITHM

A.*Problem Definition:*

Every client visit various webPages which he is concerned in, such as news, military, education and so on. But every portion has the different proportions. We take the set of all clustering objects to set $X=\{x_1, x_2 \ldots x_n\}$. In this set, each object has a limited number of indicators to evaluate it, and each indicator represents a certain characteristic of x_i. Therefore, x_i can be explained by the vector $P(x_i)=(x_i,x_{i2},\ldots, x_{im})$, and x_{ij} correspond to the j feature of object x_i.Cluster is to analyze the likeness of n objects corresponding $P(x_i)$ in the objects set X, to divide into a number of subsets X_1, X_2,..., X_m, which are not intersectant, and m is the number of sort. It need satisfy the below conditions.

$$X_1 \cup X_2 \cup \ldots\ldots \cup X_m = X_{1\ldots\ldots} \cap X_i \cap X_j \neq \emptyset$$

$$1 \leq i \neq j \leq m \qquad (1)$$

The Affiliation function

$$\varphi_{ij} = \begin{cases} 1 \\ 0 \end{cases} \quad 1 \le i \le n \text{ and } 1 \le j \le m. \quad (2)$$

If $x_i \in X_j$, then $\varphi_{ij} = 1$, else $\varphi_{ij} = 0$

and the Affiliation Function must satisfy the

$$\text{Condition, } \begin{cases} 0 < \sum_{i=1}^{n} \varphi_{ij} < 1, \forall j \\ \sum_{j=1}^{m} \varphi_{ij} = 1, \forall i \end{cases}$$

That is to say each object only belongs to one group. And each group is the non-empty really subset of X.

Besides, there are several techniques to get the likeness. This paper uses the Cosine amplitude method, and the specific algorithm formula.

$$r_{ij} = \frac{\left|\sum_{k=1}^{m} x_{ik} x_{jk}\right|}{\sqrt{\left(\sum_{k=1}^{m} x_{ik}^2\right)\left(\sum_{k=1}^{m} x_{jk}^2\right)}}$$

Where i , j =1, 2 ,…, n

To resolve the above problem, we can use the conventional methods of HAC (hierarchical agglomerative clustering) and k-means to cluster the users by their interests.

B. *The GRRS (Grouping Related Refined Segments) Algorithm and its performance*

The theory of HAC is a kind of bottom up approach. First, you have to set a infimum limit, and then take the each entity x_i of X={x_1, x_2 ...xn} to a sort. Secondly, combine these sorts by some likeness, and the rest can be done in the similar way, and terminate it until clustering met with Infimum conditions. In this way, we can get the outcome of clustering.

The theory of k-means is different from HAC, and it is divided the objects only one gradation into k sort. According to a way to get k clustering canters, calculating the similarity of each object x_i and clustering center, classified as the sort which has the maximum similarity. Then these objects in k groups are set to new k clustering centers, clustering them again and so on. Stop it until all clustering centers reaches the stable value.

HAC algorithm is simple, but the clustering process of identifying the merge points is very difficult. It could cause vicious circle and get conflicting results and decreased quality of the sort, if chosen improperly. Compared to HAC algorithm, k-means can deal with the cluster of entities in which we have to determine the k value in advance, which is the one of the greatest difficulty.

Therefore, this paper uses an improved cluster algorithm (GRRS), which combines the advantages of above two algorithms, to make up for the deficiency. By the proving of F-measure, that is a method of evaluating the effectiveness of cluster, this algorithm its complexity has increased, but the result is more accurate. It avoids the k - means algorithm into the local optimal and the defect of predefined classification number. A concrete realization procedure is as follows.

Step 1: Take each line x_i of objects set X={x_1 ,x_2 ,...x_n } to each sort alone. These n sorts constitute a cluster X={X_1 ,X_2 ,...,X_n }.

Step 2: Calculate the similarity

$$r_{ij} = \frac{\left|\sum_{k=1}^{m} x_{ik} x_{jk}\right|}{\sqrt{\left(\sum_{k=1}^{m} x_{ik}^2\right)\left(\sum_{k=1}^{m} x_{jk}^2\right)}}$$

Where i , j =1, 2 ,…, n

Step 3: Set a infimum limit, and select the maximum of similarity. If max $\ge \mu$, then combine x_i and x_j, to get $x_i = x_i \cup x_j$. So they constitute X={x_1, x2...x_{n-i} } Step 2 and Step 3 are repeated. If max μ, stop the algorithm, so you can get k clusters X={x_1, x2 ...xk}.

Step 4: Set k and the central point of average value as the sort number of k-means and cluster centers. Stop the algorithm until the cluster centers are stable, so we can get the new result of cluster as the final result.

According to the above the improved algorithm, we can control the infimum limit μ, to get different degree of clustering. The value of μ depends on the concrete issue. The different μ has a great influence on the clustering results. The infimum limit bigger, cluster is more refined, and the number of cluster is bigger. So it requests the similarity of the inner of clusters to be bigger. Otherwise, the infimum limit smaller, cluster is rougher, and the number of cluster is smaller. So it requests the similarity of the inner of clusters to be smaller.

V. CONCLUSION

In this manuscript, the internet browsing history is examined. Firstly, we have done some statistical analysis of users and visiting categories. Secondly, this manuscript puts forward and proposed an improved algorithm called as grouping related refined segments. The algorithm uses the paper to personalized recommendation and combining the mining of association rules and the time series analysis. In this way, we can take advantage of internet browsing history better, to produce practical application value.

ACKNOWLEDGMENT

We are greatly delighted to place my most profound appreciation to Er.K.Satyanarayana Chancellor, K.L.University, Dr.K.Rajasekhar Rao, Principal and Prof.S.Venkateswarlu, Head of the department, under their guidance and encouragement and kindness in giving us the opportunity to carry out the paper. Their pleasure

nature, directions, concerns towards us and their readiness to share ideas helped us to rejuvenate our efforts towards our goal. We also thank the anonymous references of this paper for their valuable comments.

REFERENCES

[1] Zhang Ning, Cao Yi, Study of the Users' Interests Based On the Internet Browsing History, 2011

[2] Zhang Weijiao; Liu Chunhuang; Li Fangyu. Method of Quality Evaluation for Clustering [J]. Computer Engineering, 2005

[3] Mitra S. An evolutionary rough partitive clustering [J]. Patterns Recognition Letters, 2004

[4] De S, Krishna P. Clustering web transactions using rough approximation [J].fuzzy Sets and System, 2004

[5] Wang Shi; Gao Wen; LI Jin Tao; and Xie Hui. Path Clustering: Discovering The Knowledge in the web site. Journal of Computer Research and Development, 2002

[6] Perkowitz M. Etzioni O. Towards adaptive Web Sites: Conceptual framework and case study. Artificial Intelligence, 2000

[7] Jose Borges, Mark Levene, Data Mining of User Navigation Patterns, In Proceeding of WEBKDD'99,San Diego, CA, USA, August 15-18,1999