Journal of Theoretical and Applied Information Technology
10th February 2013. Vol. 48 No.1

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

THE MINING METHOD BASED ON PAGE VISITING SEQUENCE OF THE PATH CLUSTERING

[1] CHI MA, [2]JIAN KONG, [3] CHUN NA ZHANG

[1, 3] College of Software, University of Science and Technology Liaoning, Anshan 114051, China

[2] Yantai Engineering&Technology college, Yantai 264006,China

E-mail: [1]aschima@126.com, [2]kwith@163.com [3]zcn1979@yahoo.com.cn

ABSTRACT

The clustering algorithm based on users visiting path, which intends to discovery the user's visiting behavior to achieve the architectural optimization of websites. The visiting sequence and visiting times reflect users' interest measures, those users who have similar interests are clustered in the same master and the implementation of the clustering process is achieved by the identification log. This essay refers to the KSearch algorithm in order to obtain the behavior model of the clustered users, and resets the relationship between interestingness and similarity. The essay also introduces the longest fitting path to define cluster center, the experiment results show that the algorithm is effective.

Keywords: *Path Cluster, User Visiting Transaction, KSearch Algorithm, Cluster Center*

1. INTRODUCTION

With the rapid development of the Web, both the architecture of a Web site and its service schema are quite different from the past, and there are deep changes in user access patterns and tendencies freedom as well. It is desirable for Web designers to optimize the site as much as possible to attract users. They also try to discover some implicit information, such as useful patterns or potential knowledge of some Web activities, correlation, visiting habits, and sequential relationship of users from log files. Thus, behavioral patterns of users can be identified through these effective analyses.

In general sense, the procedure of a clustering is to group limited data objects in a limited spatial domain, and its ultimate goal is to divide these data objects into different clusters with low inter-cluster similarity while high intra-cluster similarity. Take a Web site as an example, the relationships between the Web pages can be depicted as networks, in which pages are denoted as nodes and linked to each other. The similarity of sub-pages derived by the same node is relatively high, which means the interest between users and these sub-pages are similar. There is a positive correlation between sequential characteristics and interest measure of users visiting Web sites. For example, a user $us1$ visits Web page $UR1, UR2, UR3, ..., URn$ in a sequence, then the relation between the users' visiting sequence and interest measure can be illustrated as follows:

$$I(UR1) > I(UR2) > I(UR3), ..., > I(URn) \quad (1)$$

Formula (1) shows that there exists some relation between the users' visiting sequence and interest measure. The interest of the earlier visited page is relatively high. Thus, it is desirable to deal with the relations in the log files using a clustering method, which groups the pages into similar clusters. Then, mangers of the Web site could make use of the pattern to adjust the site structure to meet user personalization requirements. So far, the usual steps of analyzing the log files of a Web site is to preprocess the information of the site at first, then process the users' visiting behaviors, finally cluster the paths. Reference [3] identifies users' visiting transactions using fuzzy clustering, and proposes a new path clustering method, possibilistic. Based on fuzzing theory, possibilistic takes the user visiting sequence and time into consideration. However, it lacks analysis of visiting frequent measure, which is contrary to the focus of this paper. The method presented by reference [4] is a partition clustering method for pages. It studies the user visiting paths according to its classification, and utilizes the user visiting sequences when clustering. The method is simple and efficient, but it only considers the visiting sequence while neglecting the other attributes of pages, such as user visiting frequency, time on page. Reference [5] computes the

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

correlation of Web pages based on similarity matrix, and clusters pages using PageGather where the visiting frequencies of users compose the tuples in the matrix. This approach still does not take the page visiting sequence as the metric of clustering. Reference [6] process the user visiting transactions with fuzzing clustering methods, and still does not take the user visiting sequence into consideration. While reference [7] develops a novel method for preprocessing pages, which is the premise for data arranging. The common shortcoming of the above mentioned methods is lack of Web page visiting sequence, i.e. the order relationship between different nodes of the same type, and this is the focus of this paper.

2. PATH CLUSTERING

Path clustering is to find certain characteristics in user access behaviors, which is reflected in the domain formed by the path similar of user clicks on pages. The visiting procedure of a user to a Web site can be described by an interest trajectory which is constituted by a set of link addresses of visiting paths. When users exhibit similar paths, these users can be grouped into a class which can be further defined as a visiting behavior pattern. Clustering then can be conducted on the pattern's domain to obtain useful information.

Users visiting information are stored in the log files, and the format should follow the W3C standards. The structure of a Web site can be described as a directed graph, for example $D=(V,M)$, in which V denotes the set of pages, M denotes the set of links between pages. Thus, the user visiting behavior pattern is obtained.

Definition1(Web site log file): Let $F=\{(f_u.ip, f_u.us, f_u.url, f_u.time) \mid 1\le u\le n, f\in F\}$, in which F denotes the log information of the Web site, f_u denotes the actual user u visiting information including the IP address of the user $f_u.ip$, the identifier of the user $f_u.us$, the visiting page address of the user $f_u.url$, the specific page visiting time of the user $f_u.time$.

The focus of the next step is the log files clean which excludes the objects not in the research interest, such as information about videos, images, sound, and other multimedia files, retain the information about the text and HTML format. After the clean, it is of need to identify user transaction information.

Definition 2 (User visiting transaction): Let $\{f_u.url \mid 1\le u\le n\} =< ip_u, us_u, \{(f_1.url, f_1.time)\cdots (f_m.url, f_m.time)\}>$ be a visiting transaction.in which, m denotes the last visiting page of a user. The behavior of the user in the Web site is denoted as $T=f_i.time-f_j.time$, in which T denotes the time window of the user, $1\le i, j\le m$.

In order to traverse each visiting transaction, this paper attempts to find a general approach, and the steps are as follows. The pseudo-code of the traversal algorithm can be written as Table1.

Table 1: The Pseudo-Code Of The Algorithm To Traverse Each Visiting Transaction

```
ObjData fwn_SDataSet(ObjUrl ou_usurl , ObjLog ol_uslog,String s_userid,int i_timeval)
{
   fn_LogPreProc(ol_uslog);
   Int i_vncount=fn_SearchUrl(ou_usurl, s_userid);
   ObjData ob_ds=fn_DataProc(i_timeval, ou_usurl, ol_uslog, i_vncount);
   return ob_ds;
}
```

1) Preprocess the log files of the site;

2) Group the users in the log files, that is arranging the relevant content of every IP address into a set;

3) Determine the visiting transactions, identifying the visiting contents of every user according to the time window T , in which a part of a visiting content consists a data set, which is called user visiting transaction;

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4) Determine visiting time of a user according to the log file, the record of a user visiting the same address is considered only once in the visiting transactions, while the other visiting is only recorded with times but not put into the data set;

5) Sort the data set according to the time sequence of the user visiting transactions, thus they form a sequential transaction set.

Here, every new formed user transaction represents a non-repeated visiting path. The data set stores the users' visiting paths within a particular time, and these paths can be clustered using some method.

3. ALGORITHM ANALYSIS

In this paper, we present a path clustering method, called KSearch. KSearch clusters characteristics according to generated data sets, while takes the users' interest measure into consideration. In terms of the sequence of users visiting pages, after several clustering iterations, the actual situation of the user visiting sequence can be obtained. Fully given the premise of visiting order, KSearch removes the extra visiting records of the same address. Compared with other methods, KSearch is more efficient and reliable. The basic ideas of the proposed method are as follows:

1) Given data set F , and parameter k for clustering partitions;

2) Design the cluster center, and initializing the cluster center. Partition clustering the n transactions in the data set in terms of k . Notice that it is of need to minimize the total similarity between every user transaction and the cluster center;

3) Re-compute the cluster center according to the cluster center model, and transform the clustering result into a matrix. Then compute the cluster center and relevant matrix, and get the convergence evaluation function Q ;

4) Compute the convergence evaluation function Q within limited iterations, the return value of the algorithm should be convergence until meeting the requirement, or else return to step 3).

Definition 3(interest measure):The attention of users visiting a Web site can be denoted as $\text{Interest}(f_u url)$, the sequence of users visiting address can be denoted as $(f_1 url, f_2 url, \cdots, f_n url)$, different interests determine the visiting sequence: $\text{Interest}(f_1 url) > \text{Interest}(f_2 url) > \cdots > \text{Interest}(f_n url)$, in which $\text{Interest}(f_k url)\ 1 < k < n$.

Let $u_1, u_2, \cdots, u_n$ be a group of users, the visiting can be denoted as the following series:

$$u_1 - url_{11}, url_{12}, url_{15}, url_{14}, url_{13}$$
$$u_2 - url_{13}, url_{12}, url_{11}, url_{15}, url_{14}$$
$$u_3 - url_{11}, url_{13}, url_{15}, url_{12}, url_{14}$$
$$u_4 - url_{11}, url_{12}, url_{15}, url_{13}, url_{14}$$
$$u_5 - url_{11}, url_{12}, url_{15}, url_{13}, url_{14}$$

It can be seen that the visiting sequence is suitable for not only individuals but also for group characteristics. By analyzing the series we can get the group user visiting path $url_{11}, url_{12}, url_{15}, url_{13}, url_{14}$, and the corresponding interest measure can be illustrated as: $\text{Interest}(f_U url_{11}) > \text{Interest}(f_U url_{12}) > \text{Interest}(f_U url_{15}) > \text{Interest}(f_U url_{13}) > \text{Interest}(f_U url_{14})$, where U stands for group, and the visiting sequences of individuals determine the interest measure of the whole group.

Definition 4 (Similarity): There is some causality between similarity and interest measure, which means in the same link, if the interest measure of two users to the same Web page are consistent, then the two users are similar; or conversely. Let the user visiting transaction of two users are S_i and S_j respectively, then the similarity can be denoted as: $sim(S_i, S_j) = \sum_{p=1}^{n} (|\text{Interest}(f_{s_i} url_p) - \text{Interest}(f_{s_j} url_p)|)$ in which, the absolute value of the difference between the interest measure is closer to zero, the more similar they are.

Definition 5(Clustering center): Based on the scatter chart in regression analysis, we assume that the user interest measure as the independent variable, and the selected link address as the dependent variable. Thus the relevant data points of visiting users of different types should generate convergence, and finally fitting. Based on the Definition 3, we give the following definition: the longest user visiting fitting path, i.e., starting from the initial address, the longest path of most similar

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

visiting address of all set members. Users' access paths are shown in figure1.

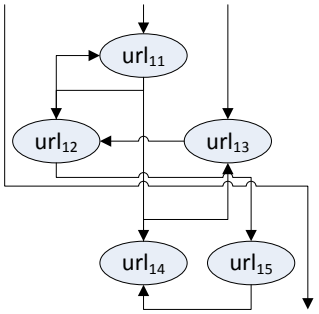

Figure :User Access Path

Thus, we can get cluster result $C=\{c_i \mid 1<i<n\}$, where c_i includes several user visiting transactions. The similarity of transactions from the same cluster is relatively high. The group center, i.e. the accumulative path of interest measure in the cluster, can be illustrated as follows:

$$cen(c_i)=\{f_j url_1,\ f_j url_2,\cdots,\ f_j url_k\} \quad (2)$$

where the longest path is denoted as:

$$\max(\sum \text{Interest}(f_j url_i))$$

With reference to the series of Definition 3, we can get the longest fitting path in which we assume that the length of a user visiting path is fixed. And in order to facilitate the processing, we convert it into a vector form which can be denoted as follows:

$$W=(w_1,w_2,\cdots,w_n) \quad (3)$$

Then transform the user visiting transaction vectors into a matrix:

$$V_{(m\times n)}=\begin{pmatrix} W_{11} & & \\ W_{21} & & \\ \vdots & & \\ W_{m1} & & W_{mn} \end{pmatrix} \quad (4)$$

where m denotes all the user transactions from Web site log files, n denotes the length of every user transaction, i.e. the number of visiting addresses. Here we assume the length is a fixed length.

The proposed KSearch algorithm is not a hierarchical clustering method in a traditional sense, whereas it takes the similarity as a reference condition to classify the paths. The core idea of KSearch is to cluster the transaction set into several clusters, $cen(c_i)$ denotes the pages formed according to the ith interest measures. Notice that this path might not exist in real. The problem is about equation (2) and (4), and defines the clustering fitting function as follows:

$$Q(V,W,C)=\min(\sum_{i=1}^{m}(\sum_{j=1}^{n}v_{i,j}sim(W_j,cen(c_i))) \quad (5)$$

where $v_{i,j}$ is an item of the matrix, the maximum m and n could equal or not. When solving the fitting function, we assume that as follow: because W is a sub-set of V, the operations here is divided into two steps, which treats the other two parameters except W in $Q(V,W,C)$ as constant to solve.

Step1: Let C be a constant, solve $Q(V,W,C)$:

$$v_{i,j}=\begin{cases}1 & sim(W_j,c_i)\le sim(W_j,c_k)\\ 0 & i\neq k\end{cases}$$

Where $1\le k\le m$.

Step2: Let W be a constant, solve $Q(V,W,C)$:

1) Give an initial constant C_0, solve $Q(V,W,C)$, then get the initial W_0;

2) Let $k=0$, utilize C_0 and W_0 to deduce C_1, and the like, thus get $C_2,\cdots,C_m$, then $W_1,W_2\cdots,W_n$. In deducing process, if $Q(V,W,C_i)=Q(V,W,C_{i+1})$, then output the corresponding W, the algorithm terminates; or else, perform step three;

3) Utilize C_0 and W_0 to deduce W_1, then get $W_2,\cdots,W_n$ as well, then $C_1,C_2\cdots,C_m$. At this point, if $Q(V,W_j,C)=Q(V,W_{j+1},C)$, then output C, the algorithm terminates; or else, perform step 2).

According to this computation, within limited iterations, the algorithm converges and gets a local optimal solution.

ISSN: 1992-8645 www.jatit.org E-ISSN: 1817-3195

4. EXPERIMENTS AND EVALUSATION

The experiment data are derived from the server log files of an e-commerce travel site in the period of 2012. After cleaning the full-year data, we get some visiting records. The whole Web site contains 325 pages, and the volume of user visiting page is 89M. Analyzing by the algorithm, we get 4225 user visiting transactions, and the average visiting path of group users are 6.3. We partition the data into five sets according to time interval, and validate our algorithm on these data. Comparison of algorithm applied before and after is shown in Table 2 as follows:

Table 2: Comparison Of Before And After The Algorithm Is Applied

Group	1	2	3	4	5
Number of records	47258	47110	42156	45697	53252
Number of user transactions	315	326	268	336	371
Number of clusters(applied before)	51	54	47	49	58
Average of cluster centers (applied before)	7	7	7	8	11
Number of clusters(applied after)	66	67	61	68	77
Average of cluster centers(applied after)	10	10	9	11	15

According to the above analysis, after applying the algorithm, the number of clusters rose 31%, and the number of cluster centers rose 37%, which means the effect of clustering is significantly improved and proves to be sensible.

5. CONCLUSIONS

Based on visiting sequence of users to a Web site, this paper proposes KSearch algorithm, a partition clustering method, to cluster user visiting paths. This method emphasizes the users interest measures to understand the real needs of users, thus site managers can tailor the site structure to cater to the customers. At the same time, this method takes the relations between pages into account, and provides some personal serve in terms of user similarity. Experimental results show that the proposed algorithm is effective.

Acknowledgements: Thanks for the software technology laboratory at USTL.

REFRENCES:

[1] Vassiliki A. Koutsonikola, Athena I. Vakali. “A fuzzy bi-clustering approach to correlate web users and pages”. *International Journal of Knowledge and Web Intelligence*, Vol.1, No2, 2009, pp. 3-23.

[2] Y.F. Liu. “A text information extraction algorithm based on tag xpath clustering” .*Computer applications and software*.Vol.21, No.11, 2010, pp.199-202.

[3] S. H. Yang, H. L. Lin, Y. B. Han. “Automatic data extraction from template generated Web pages”. *Journal of Software*, Vol.19, No.2, 2008, pp.209-223.

[4] Q Lu ,Z.N. Zhang. “Research on path clustering in web sites”. *Computer Applications and Software*, Vol.25, No.8, pp.205-226.

[5] J. C. Song, J. Y. Shen. “An Intelligent Recommendation System Based on Web Navigation Path Clustering”. *Information and Control*, Vol.36, No.1, 2007, pp.119-124.

[6] H. Yu, H. Luo, S. S. Chu. “Web Users Access Paths Clustering Based on Possibilistic and Fuzzy Sets Theory”. Proceedings of ADMA, Vol.1, 2010, pp.12-23.

[7] Jae-Min Lee,Byung-Yeon Hwang. “Two-Phase path retrieval method for similar XML document retrieval”.*R.Khosla et al.(Eds.): KES*,2005, 3681:967-971.

[8] J. J. Wu, J. J. Chen, S. Z. Zhao. “Research of Path Clustering Based on the Access Interest of Users”. *Computer engineering and applications*.Vol.41, No.36, 2005, pp.170-182.

[9] J. M. Chen, H. Lu, Y. Q. Song. “A Possibility Fuzzy Clustering Algorithm Based on the Uncertainty Membership”. *Journal of Computer Research and Development*, Vol.45, No.9, 2008, pp.1486-1492.

[10]Y. J. Yu, H. Z. Lin, C. Chen. “Personalized web recommendation based on path clustering”.*ICCOMP'06 Proceedings of the 10th WSEAS international conference on Computers*, 2006, 148-153