Knowledge Discovery Technology Based on Access Information Mining on Knowledge Warehouse

Yongdang Chen, Yang Wang, Xiao Xiao
College of Mechanical & Electrical Engineering
Xi’an Polytechnic University
Xi’an, China

Meihong Shi
College of Computer Science
Xi’an Polytechnic University
Xi’an, China

***Abstract*—This study was conducted to explore a new kind of mining method based on web access information and its algorithm. The paper analyze general web mining and knowledge discovery method, and summarize the limitation of common data mining method based on weblog access information. On this basis, it put forward a new approach of web access information mining based on self-built access information database, and expatiated on the description method of knowledge warehouse navigation page sets and user access action, and the mining algorithm based on access database. This new method is more simple, flexible and efficient.**

Keywords-web access information; knowledge management; data mining; knowledge discovery

I. INTRODUCTION

With the rapid development of internet application, a mass of web access information was accumulated on World Wide Web (WWW). Through effectively data mining on this information, knowledge about user access action is available. This knowledge can serve service providers of web sites to help improve the design of Web site, improve the performance of web services and increase more personalized service. Mining on web access information has become an important study field in the world[1-3]. Currently, the research mainly focused on system improvement, user modeling, discovery and navigation model, improvement of the efficiency of the web site, real time personalized recommendation and business intelligence discovery[4-6]. Its mining object is log file records on the server, including server log data[7,8]. However, in practice, because of the client cache, the client agent, repeat calls to navigation pages and so on, weblog data records may be inaccurate or incomplete. Thus, the effect of web access information mining is affected. For these problems, the paper analyze general web mining and knowledge discovery method, and summarize the limitation of common data mining method based on weblog access information. It put forward a new approach of web access information mining based on self-built access information database, and expatiated on the description method of knowledge warehouse navigation page sets and user access action, and the mining algorithm based on access database.

II. KNOWLEDGE DISCOVERY AND WEB MINING

KDD（Knowledge Discovery based on Database）is to extract people interested knowledge from the database and data warehouse. This knowledge is implicit, previously unknown, potentially useful, easily understood information. KDD is a gradually developed branch of computer science in recent years and a new attempt in artificial intelligence field. KDD has been successfully used in industrial, agricultural, military, financial and commercial aspects, and become one of the current focus of computer science.

Currently, KDD mainly research task description, knowledge evaluation and knowledge representation of knowledge discovery. The effective knowledge discovery algorithm is the key. Specifically, that is mining knowledge such as association rules, data clustering, classification rules, sequential pattern, similar model, chaotic pattern and so on with the above method and its integrate technologies in all kinds of real-world database (relations, interpretation, temporal, spatial, distributed, object-oriented).

KDD process is usually divided into the following steps, as shown in Figure 1.

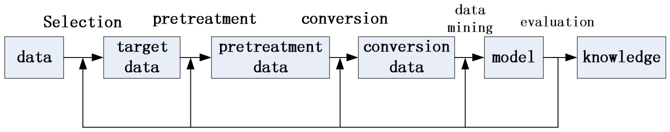

Figure 1. KDD process

From Figure 1, obviously, KDD is a process linked up by many steps and iterative process of human-computer interaction. KDD application reflected web mining in internet or intranet. Logically, the web can be seen as a directed graph G = (N,E) in the physical network. Among, node set N corresponds to all documents on the web, and directed edge set E corresponds to the hyperlinks between nodes. Further division on the node set, N={Nl,Nnl} (as shown in Figure 2). All non-leaf nodes Nnl is HTML document which contain the tag to specify the document's properties and internal structure, or embed a hyperlink to indicate structure relationships between documents with the exception of text. The leaf node Nl can be a HTML document, and also be a document in other formats such as PDF, text files, as well as graphics, audio and

978-1-61284-181-6/11/$26.00 ©2011 IEEE

other media files. Each node N has a URL, which contains the information of the Web site and the directory path.

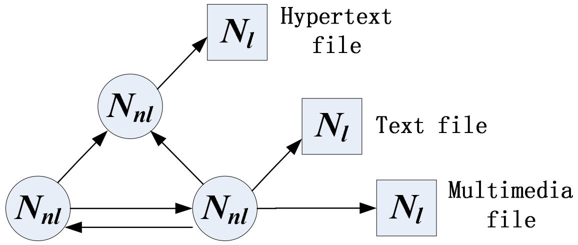

Figure 2. Web logic structure (directed graph G)

Diversity of information on the web determines the diversity of web mining tasks, which can be divided into content mining, access information mining and structure mining and so on. With web mining, need knowledge can be extracted from the web page. General users access action, frequency and content is analyzed. Universal knowledge about user group access action is gained, and be used to improve the design of web services, such as fast and efficient channel to access the highly relevant content, efficient and accurate personalized service for customers.

III. KNOWLEDGE WAREHOUSE ACCESS INFORMATION MINING

Generally, the primary data objects of access information mining are access path and query information which is query keyword entered by the user. The main data source of web mining is user access log file (Weblog) on web server. However, in knowledge management system based on knowledge warehouse, the limitations of mining based on Weblog as follows:

- Client caching makes Weblog data inaccurate.
- Client agent (Proxy) may be screen the detail of the users who access the web.
- Because of the call for knowledge in the knowledge warehouse, it often return to the same navigation page and makes Weblog data records incomplete.

Therefore, this paper put forward a new approach of web access information mining based on self-built access information database.

A. The description of knowledge warehouse navigation page set

Topology map of knowledge warehouse navigation page set is a directed graph. User browsing mode is a subgraph within a period of time. The similar user access subgraph form a similar user group. The path that user often access make up frequent path.

Define 1. Knowledge warehouse navigation page set is a directed graph: $G=(N,N_p,E,E_p)$, among, N is a node set, that contains the full URL address in the knowledge warehouse navigation page set. $N_p=\{node\in N,\{(USERID,hits)\}^n\},n\geq 1$, is the node property set that records USERID and the number of the user access node. E is a directed edge set. $E_p=\{e\in E,\{\{NumberofPath\}^p\}^m,p\geq 1,m\geq 1$, is a property set of the directed edge, that records the number of the path of the edge connecting.

B. The description of user access action

Define 2. User access affair to knowledge warehouse TW:

$TW=(TASK,UID,URL,TIME,TIMELENGTH)$, among, TASK、UID、URL、TIME、TIMELENGTH respectively is the user task group, user ID, access URL, access time and the length of time. Next, treatment on this information can reflect user access action within a period of time.

Define 3. A user access action *tw*:

$tw=(uid_{tw},\{(S_k^{tw}\cdot uid,S_k^{tw}\cdot URL,S_k^{tw}\cdot time,S_k^{tw}\cdot timelength)\}^m)$. Among, S^{tw} represents a access sequence of the user. *k* represents serial number. *m* represents the number of access sequence, $1\leq k\leq m$.

Define 4. Within a fixed period of time T , the number of a user access a URL:

$Hit(u,URL)=\left\|\{uid_{tw}=u,URL\in\{S^{tw}\cdot URL\},tw\in TW\}\right\|$. In order to filter out the transition page scanned while looking for knowledge, it requirements $Hit(u,URL)\geq 2$.

Define 5. The relative long of a user access a URL:

$$TimeLength(u,URL)=\frac{\sum_{t=1}^{m}S_j^{tw_t}.timelength}{\max(\sum_{t=1}^{m}S_j^{tw_t}.timelength)}.$$

Among, $tw_t\in TW,uid_{tw_t}=u,S_j^{tw_t}.URL=URL$, the value is a relative value, indicating a time preference characteristics of a user access a URL. Taking into account the user usually using other tools to work while using the knowledge management system, spending time in a page has been exaggerated, and more than 20 minutes to access a page is counted 20 minutes.

Define 6. The access new degree of a user access a URL:

$$New(u,URL)=\frac{\dfrac{\sum_{i=1}^{Hit(u,URL)}(Time_i(u,URL)-Time(Start))}{Hit(u,URL)}}{T}$$ .Among,

$Time_i(u,URL)=S_j^{tw_i}.time,\quad tw_i\in TW,uid_{tw_i}=u,S_j^{tw_i}.URL=URL$. $Time_i(u,URL)$ represents the moment that the user u *i*th access the URL in a fixed time period time T. $Time(Start)$ is the moment of start recording. The value is a relative value, and represents the freshness degree characteristics of a user access a URL. Obviously, if a user has recently frequently accessed, the value is greater, otherwise the value is smaller.

Define 7. The degree that a user *u* emphasis on a URL:

$Concern(u,URL) = \frac{\sum_{i=1}^{Hit(u,URL)} (\frac{\|S_j^{tw_i}.URL\| - k + 1}{\|S_j^{tw_i}.URL\|})}{Hit(u,URL)}$. Among, $tw_t \in TW, uid_{tw_t} = u, S_j^{tw_t}.URL = URL, j = 1 \cdots m, 1 \le k \le m$. The emphasis degree value is a relative value, express importance property of a user access a URL. Obviously, if the user first access the URL every time, then the degree of user emphasis on the URL must be a larger. Suppose that a user access m URL. In these URL, if he first access URL, then his attention to the URL is highest, the value is mark $m/m=1$. If he second access URL_j, his attention to the URL_j is higher, the value is mark $(m-1)/m$, …, The value of last URL to be accessed is $1/m$.

Define 8. The degree of a user interest in a URL:

$$Interest(u,URL) = Hit(u,URL) \times (1 + TimeLength(u,URL) + New(u,URL) + Concern(u,URL)).$$

According to this information, the interest degree matrix of user access can be defined.

Define 9. The interest degree matrix of user access:

$$M_{mxn} = \begin{bmatrix} I_{11} & I_{12} & \cdots & I_{1j} & \cdots & I_{1n} \\ I_{21} & I_{22} & \cdots & I_{2j} & \cdots & I_{2n} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ I_{i1} & I_{i2} & \cdots & I_{ij} & \cdots & I_{in} \\ \cdots & \cdots & \cdots & \cdots & \cdots & \cdots \\ I_{m1} & I_{m2} & \cdots & I_{mj} & \cdots & I_{mn} \end{bmatrix}.$$

Among, all URL addresses can be obtained from the node set N in the directed graph G. USERID and the interest degree of user access each node can be directly or indirectly obtained from the node attribute set. I_{ij} is the interest degree of user j access URL i in a certain time period. Each row vector $\overrightarrow{M}(i,*)$ contains the information of all user access the URL i. Each column vector $\overrightarrow{M}(*,j)$ contains the information of the user j access all URL.

From the above definition, row vector represents the topology map of knowledge warehouse navigation page set, also voiced a common user access pattern. Column vector reflect the user type and show the subgraph of user access. Therefore, figuring out the similarity degree between the row vectors and between the column vectors, the relevant navigation page sets and similar user group can be known. Calculation of similarity degree between vectors is based on the Euclidean distance (EUCLID).

$$EU_d(X,Y) = \left[\sum_{i=1}^{n} (X_i - Y_i)^2 \right]^{1/2}$$

The Euclidean distance between vectors can be calculated. Euclidean distance is smaller, the similarity degree between the vector is higher.

C. The mining algorithm based on access database

User browsing action to knowledge warehouse can be defined as described earlier. Topological structure of navigation page set for knowledge warehouse is known. Even though the browse mode of different users in different time periods is different, overall knowledge need for the larger size task is stable in a large period of time. Thus, similar user group and relational navigation pages and frequent access paths can be discovered by analysis for user browsing action to knowledge warehouse in a time period.

1) Nnavigation page clustering algorithm

As shown above, row vector $\overrightarrow{M}(i,*)$ of the interest degree matrix $M_{m\times n}$ indicates the interest degree information of all users access to URL i. If some knowledge content is accessed by the user performing similar tasks, the knowledge content is also relevant and can be aggregated to same class as the basis for knowledge supply. While clustering, Hierarchical Clustering Method which is the most common system clustering method at home and abroad can be used. It is able to generate a hierarchical nested clusters and high accuracy.

For the row vector $\overrightarrow{M}(i,*)$ of the interest degree matrix $M_{m\times n}$, the process of system clustering method is as follows:

a) Each row vector $\overrightarrow{M}(i,*)$ be seen as a cluster $c_i = \overrightarrow{M}(i,*)$ with a single member. These clusters constitute a cluster $C = \{c_1, ..., c_i, ..., c_m\}$.

b) Calculate similarity degree $Sim(c_i, c_p)$ between each pair of cluster (c_i, c_p) in the C, among $1 \le i \le m, 1 \le p \le m, i \ne p$.

c) Selecting the pair of cluster with the greatest similarity $\arg\max sim(c_i, c_p)$, and merge c_i and c_p to a new cluster $c_k = c_i \cup c_p$, to form a new cluster $C = \{c_1, ..., c_k, ..., c_{m-1}\}$ of $M_{m\times n}$.

d) Repeat the above steps, until remaining a cluster in C.

e) From the above process, a clustering tree is constructed. It contains the cluster level information, as well as the similarity degree in all clusters and between clusters.

f) Decide the number and the class of the cluster.

2) mining algorithm for similar users

As mentioned earlier, the column vector $\overrightarrow{M}(*,j)$ of the interest degree matrix $M_{m\times n}$ represents the subgraph of user personal access. The user with similar access subgraph can be form similar user group. The algorithm can be divided into the following three steps:

a) Using Euclidean distance, calculate the similarity degree between current user and all other users.

b) Select n users best similar to the current user as close neighbors from big to small. Here, if the value of n is too large, then users who have high similarity will be introduced excessive noise. If the value of n is too small, so for those who do not have a high similarity will be mislead.

Calculate the averages value of the neighbors interest degree in each URL, for recommended.

3) The mining algorithm for frequent access path

The links exist between relevant knowledge warehouse navigation pages, and it is made constraints by the topology graph of knowledge warehouse navigation page set. In other words, while clustering for the navigation page, at least part of the navigation page form a cluster because of the topology structure of knowledge warehouse navigation pages set. Therefore, analyzing the results of clustering of navigation page, the frequent access paths can be got. For frequent access paths mining, the process of analysis for each navigation page cluster *CW* as follows:

According to the feature set E of the directed edges in graph G, all URL in the CW are checked and determined whether they are on a path. If all URL is not on a path, they cannot form a frequent access path. Then, the next CW will be analyzed. If there are n paths in the CW, it also need to be checked whether they can be merged. Frequent path *fp* is calculated by define 10.

Define 10. Path frequency: The access frequency of path $p = \{URL_1, URL_2, ..., URL_i, ..., URL_n\}$ is defined as the ratio of the total hits number for each shortest path node and the total hits number for all nodes in knowledge warehouse navigation page set.

$$fp = \sum_{i=2}^{n} hits(URL_i) / \sum_{j=1}^{\|Sweb\|} hits(URL_j) .$$

For any one path, if its access frequency exceeds the threshold value λ, it is considered as a frequent access path.

Because a navigation page usually contain multiple hyperlinks pointing to different direction, the selection of knowledge warehouse access path is a problem in the face. Discovery of frequent access path is help to provide the candidate appropriate hyperlinks for users. Frequent access path provides a good basis for the improvement of structure design of knowledge warehouse.

IV. CONCLUSIONS

Web information mining is widely used in financial services, retail, government management, manufacturing, medical services and other fields. Its application prospect is mainly in three areas: *1)* E-commerce. With web mining technologies, the hidden model information can be automatically found from the data in the server and client log. According to the system access mode and user behavior pattern, the predictive analysis can be made. *2)* Website design. Through mining for user access logging information and grasping the interest of users, it is help website to provide information push service and customized service for personal. *3)* Search engines. Through web content mining, clustering, classification, category browsing and retrieval for network information can be achieved. through analysis for user questioning history records, effectively questioning way is extended and the effect of user retrieval is improved.

ACKNOWLEDGMENT

This work is supported by the Scientific Research Fund for Doctor of Xi'an Polytechnic University(No.BS0711), and Shaanxi Nature Science Foundation Research Project(No.SJ08-ZT14).

REFERENCES

[1] ZHANG Yajun, "One new web log data mining model and it's application in digital library," Information Science, vol.1, 2011, pp. 27–33

[2] Yu-Sheng Lo, Yu-Ting, Chien-Tsai, "Using web usage mining techonology to evaluate the effectiveness of internet health promotion activities," Proceedings of 2010 Third International Conference on Education Technology and Training , pp. 440–443, November 2010

[3] YU Xiao-bing, GUO Shun-sheng, HUANG Xiao-rong, "Intelligent e-commerce based on Web usage mining and its application," Computer Integrated Manufacturing Systems, vol.2, 2010, pp.53–58

[4] ZHONG Qian-lin, WANG Huan-min, DU Ya-jiang, "Integration techniques of web mining based on web services," Taiyuan Science & Technology, vol.8, 2009, pp.29-35

[5] Yang Yifan, Zhu Ming, Li Huahu, "Study and improvement on linkage similarity-based web mining algorithm," Computer Applications and Software, vol.1, 2011, pp.64-69

[6] BAI Juan, BU Hui, "Intelligent recommendation for e-commerce based on web mining," Microcomputer Information, vol.21, 2010, pp.71-77

[7] LIU Shi-jie, "Web mining model based on application layer log," Proceedings of 2009 International Forum on Information Technology and Applications, pp. 504–507, May 2009

[8] FU Xiang, JIN Ou, "Data preprocessing method for web usage mining," Computer Systems & Applications, vol.8, 2010, pp.58-63