ISSN No. 0976-5697

Volume 2, No. 5, Sept-Oct 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

User Access Pattern Classification Scheme for Optimizing Web Directories

P.Rubini*
Asst.Professor
Department of Information Technology,
Gnanamani College of Technology,
Namakkal.
ruby_nila@yahoo.co.in

K.Sudhakar
Asst.Professor
Department of Computer Science & Engineering,
Sengunthar College of Engineering,
Namakkal.
ksudhakar.cs@gmail.com

A.Anbarasan
Asst.Professor
Department of Information Technology,
Gnanamani College of Technology,
Namakkal.
new.anbu27@gmail.com

Abstract: Web is the massive information source in the world. Information retrieval is the complex task in the web environment. Search engines handle the information retrieval process in two ways. They are query-based information retrieval and directory based information retrieval. Google provides the query based information retrieval model. All the information is fetched with respect to the user query value. Yahoo provides directory and query based information schemes. All information is arranged in the hierarchical domain order. Web directory is a hierarchical tree structure of domain and sub domain information. The web directories are classified into two types. Artificial web directories are constructed with reference to the web document contents. Real web directories are constructed with the usage data. Personalization can be applied on the real web directory environment. Objective Community Directory Miner (OCDM), Objective Probabilistic Directory Miner (OPDM) and Objective Clustering and Probabilistic Directory Miner (OCPDM) methods are applied in the existing web directory personalization schemes. The proposed system is designed to perform web directory optimization using the classification techniques. The probabilistic latent semantic analysis algorithm is used for the classification process. The fuzzy logic technique is used to enhance the PLSA scheme for weight optimization. The web directory optimization model uses ISP based user access logs.

Keywords: Personalization, Probabilistic Latent Semantic Analysis Algorithm, Fuzzy logic, Web directory optimization.

I. INTRODUCTION

At its current state, the Web has not achieved its goal of providing easy access to online information. As its size is increasing, the abundance of available information on the Web causes the frustrating phenomenon of "information overload" to Web users. Organization of the Web content into thematic hierarchies is an attempt to alleviate the problem. These hierarchies are known as Web Directories and correspond to listings of topics, which are organized and overseen by humans [14]. A Web directory, such as Yahoo (www.yahoo.com) and the Open Directory Project (ODP) (dmoz.org), allows users to find Web sites related to the topic they are interested in, by starting with broad categories and gradually narrowing down, choosing the category most related to their interests.

A. *Problem Statement:*

The information for the topic that a user is seeking might reside very deep inside the directory. Hence, the size and the complexity of the Web directory itself are canceling out the gains that were expected with respect to the information overload problem, i.e., it is often difficult to navigate to the information of interest to a particular user.

We claim that we can overcome the deficiencies of Web directories and Web personalization by combining their strengths, providing a new tool to fight information overload. In particular, we focus on the construction of usable Web directories that model the interests of groups of users, known as user communities. The construction of user community models, i.e., usage patterns representing the browsing preferences of the community members, with the aid of Web Usage Mining has primarily been studied in the context of specific Web sites [4]. In our work, we have extended this approach to a much larger portion of the Web through the analysis of usage data collected by the proxy servers of an Internet Service Provider (ISP).

More specifically, we present a knowledge discovery framework for constructing community-specific Web directories. Community Web Directories exemplify a new objective of Web personalization, beyond Web page recommendations [5], [6], or adaptive Web sites [7]. The members of a community can use the community directory as a starting point for navigating the Web, based on the topics that they are interested in, without the requirement of accessing vast Web directories. Thus, personalization can be of particular benefit to large generic directories such as ODP, or Yahoo!. Personalized versions of these directories can also be employed by various services on the Web, such as Web portals, in order to offer their subscribers a personalized view of the Web. Moreover, Web search engines to provide personalized results to queries can exploit community Web directories. The construction of community directories with usage mining raises a number of interesting research issues, which are addressed in this paper. One of the challenges is the analysis of large data sets in order to identify community behavior. Moreover and apart from the heavy traffic expected at a central node, such

as an ISP proxy server, a peculiarity of the data is that they do not correspond to hits within the boundaries of a site, but record outgoing traffic to the whole of the Web. This fact leads to increased dimensionality and semantic incoherence of the data, i.e., the Web pages that have been accessed.

II. RELATED WORK

Web usage mining has been used extensively for Web personalization. A number of personalized services employ machine learning methods, particularly clustering techniques, to analyze Web usage data and extract useful knowledge for the recommendation of links to follow within a site, or for the customization of Web sites to the preferences of the users. A thorough analysis of these methods, together with their pros and cons in the context of Web Personalization, is presented in [1] and [2].

PLSA has been used in the context of Collaborative Filtering and Web Usage Mining. In the first case, PLSA was used to construct a model-based framework that describes user ratings. Latent factors were employed to model unobservable motives, which were then used to identify similar users and items, in order to predict subsequent user ratings. PLSA was used to identify and characterize user interests inside certain Web sites. The latent factors segmented user sessions to support a personalized recommendation process. A similar approach was followed in [3], where each user session was "mapped" onto a sequence of latent factors, named tasks that correspond to a more abstract view of user behavior. The resulting "task-sequences" were used for statistical analysis of user behavior, such as finding the most frequent tasks, or for generating recommendations. In [8], PLSA was exploited to build user profiles, where each profile consists of the distribution of Web pages over the set of the latent factors. Subsequently, user profiles supported personalized Web search. More recent work [11] used PLSA to cluster users in legitimate and malicious ("shilling") groups.

On the other hand, a number of studies exploit Web directories to achieve a form of personalization. Users build their profiles by specifying a set of categories from the ODP hierarchy. Automatic profile construction is proposed in [12]. The user profiles linked to categories of the directory are used typically for personalized Web search, while the directory itself is not personalized. The personalization of Web directories is mainly represented by services such as Yahoo! and Excite (www.excite.com), which support the manual selection of interesting categories by the user. An initial approach to automate this process was the Montage system, which was used to create personalized portals, consisting primarily of links to the Web pages that a particular user has visited, while also organizing the links into thematic categories according to the ODP directory. A related technique for mobile portal personalization was presented, where the portal structure was adapted to the preferences of users. A Web directory was used as a "reference" ontology and the Web pages navigated by a user were mapped onto this ontology using document classification techniques, thus resulting in a personalized ontology. Finally, in recent work [13], the similarity between users, based on navigation data within the ODP, was used to create clusters of ODP categories. These clusters were further exploited to recommend shortcuts within the Web directory.

Our work differs from the above cited approaches in several aspects. First, instead of using the Web directory for personalization, it personalizes the directory itself. Compared to existing approaches to directory personalization, it focuses on aggregate or collaborative user models such as user communities, rather than content selection for single user. Furthermore, unlike most existing approaches, it does not require a small set of predefined thematic categories, which could complicate the construction of rich hierarchical models. Finally, the work presented in [13], which is closest to ours, is limited to the recommendation of short navigation paths in the ODP hierarchy, rather than the personalization of the whole Web directory structure. Moreover, that method makes the assumption that usage data are collected from the navigation of users within the Web directory. Thus, its applicability to independent services such as a Web portal is questionable.

In this paper, we propose a knowledge discovery framework for building Web directories according to the preferences of user communities. Community Web directories are more appropriate than personal user models for personalization across Web sites, since they aggregate statistics for many users under a predefined thematic taxonomy, thus making it possible to handle a large amount of data, residing in a sparse dimensional space. To our knowledge, this is the first attempt to construct aggregate user models, i.e., communities, using navigational data from the whole Web. Compared to our earlier work on this topic, in this paper, we address the problem of "local overload." We achieve this by combining thematic with usage information to model the user communities. On this basis, we present new versions of the approaches introduced in [9] and a new method that combines crisp clustering with probabilistic models.

III. DISCOVERY OF COMMUNITY WEB DIRECTORIES

The construction of community Web directories is a fully automated process, resulting in operational personalization knowledge, in the form of user models. User communities are formed using data collected from Web proxies as users browse the Web. The goal is to identify interesting behavioral patterns in the collected usage data and construct community Web directories based on those patterns. The process of getting from the data to the community Web directories is summarized below:

Usage Data Preparation comprises the collection and cleaning of the usage data, as well as the identification of user sessions. Web Directory Initialization provides the characterization of the Web pages included in the usage data, according to the categories of a Web directory. We compare two different approaches for the characterization of the Web pages. The first approach organizes Web pages into an artificial Web directory using hierarchical document clustering. The second approach classifies them onto an existing Web directory, like ODP.

Community Web Directory Discovery is the main process of discovering the user models from data, using machine learning techniques and exploiting these models to build the community Web directories. The first two stages result in the construction of the required structures for the discovery of community Web directories. These stages are presented in Sections 3.1 and 3.2. An initial discussion of

the third stage is presented in Section 3.3, while more details are provided in the sections that follow.

A. Usage Data Preparation:

The usage data that form the basis for the construction of community Web directories are collected in the access log files of ISP cache proxy servers. These data record the navigation of the subscribers through the Web, and hence, they are usually diverse and voluminous. The outgoing traffic is much higher than the usual incoming traffic of a Web site and the Web pages more diverse semantically. The task of usage data preparation, detailed in [9], is to assemble these data into a consistent, integrated, and concise view. The next stage is the identification of individual user sessions. The fact that we are focusing on the discovery of behavioral patterns in the data, rather than individual users, allowed us to overcome the lack of user registration data or other means of user identification, such as cookies, and led us to exploit a simple kind of user session. A user session is defined as a sequence of log entries, i.e., accesses to Web pages by the same IP address, where the time interval between two subsequent entries does not exceed a certain time threshold. More formally:

B. Web Directory Initialization:

The next stage toward the construction of community Web directories is the association of the users' browsing data with the Web directory.

a. Artificial Web Directory:

An artificial Web directory is constructed from the usage data themselves. In particular, by exploiting the vector space representation of the Web pages, a taxonomy is built using a hierarchical agglomerative approach for document clustering. The resulting hierarchy is a binary tree, representing clusters of Web pages that form thematic categories. This hierarchy corresponds to the initial Web directory, which provides directly a mapping between the Web pages and the categories that the pages are assigned to. Details of how the artificial Web directory is constructed can be found in [9].

b. "Real" Web Directory:

The structure of the artificial directory is similar to that of existing Web directories since each node can be considered a category of the directory containing a set of semantically similar Web pages. Nevertheless, there are also notable differences between a "real" Web directory and the artificially constructed one. One such difference is that the initial directory (without personalization) is already strongly focused on the usage data, rather than being a general resource. As a result, the potential benefits of personalization are rather limited. A more practical difference concerns the artificiality of thematic categories (document clusters). The artificial directory is a binary tree, where each node clusters exactly two subnodes. In addition, the number of nodes has been chosen using only statistical properties of Web page contents. Finally, the artificial Web directory might not "cover" the semantics of new sessions due to "overfitting" of the document clustering approach on the initial data. These observations motivated us to study the personalization of a "real" Web directory and in particular the ODP. The main difficulty in this effort was the association of usage data, i.e., the Web pages, to categories of the Web directory, given the small proportion of Web pages that are explicitly assigned (manually) to categories of the directory. We approached this problem by an automated page classification method developed in [10].

The hierarchical classification of Web pages requires us to redefine the notion of user session in order to work with the categories in the Web directory rather than the Web pages themselves. In our approach, pages are mapped onto thematic categories of the hierarchy, and therefore, a user session is translated into a sequence of categories. We define the user sessions, which result from this mapping as thematic user sessions since they do not contain the Web pages themselves, but rather their thematic representation.

C. Community Web Directory Discovery:

Having determined the mapping and the associations between Web pages, user sessions, and Web directory categories, we employ unsupervised learning to discover patterns of interest in the thematic user sessions. In our recent work, we employed two methods for the discovery of community Web directories. We presented an extension of the cluster mining algorithm, named Community Directory Miner (CDM), while in [9], we presented an approach based on the discovery of latent semantics using PLSA. These algorithms are used to extract a subset of the categories of the initial Web directory that correspond to the community models, i.e., usage patterns that occur in data and represent the browsing preferences of community members. Each community model _ is subsequently exploited to construct the community Web directory. The general process of discovering community Web directories can be seen as a construction of the subgraph G' of the Web directory G which corresponds to the community Web directory. More formally:

IV. OBJECTIVE CATEGORY INFORMATIVENESS

To alleviate the "local" information overload problem discussed above, we introduce an additional criterion in the discovery of user communities. This criterion incorporates a measure of a priori informativeness of categories, which is taken into account when pruning leaf nodes from the Web directory. The inclusion of leaves that satisfy this criterion selectively reduces the generality of the directories, making them reflect more "fine-grained" user interests and resulting in a better distribution of the information that is indexed. The proposed measure of objective informativeness is based on the entropy of the category. More formally, let $c_i \in C$ a category of the Web directory. This category corresponds to a node of the Web directory and can be represented by a Boolean random variable C_i. This variable is true for Web pages of the category and false for the remaining Web pages in the directory. Thus, the probability distribution of the variable $p(C_i)$ is estimated across the number of Web pages in the directory as:

$$p(C_i = true) = \frac{\#\ pages\ \ in\ \ the\ \ category}{total\ \ \ number\ \ \ of\ \ \ pages}$$

The a priori amount of information in this category before examining the users' browsing behavior can be estimated by its entropy $H(C_i)$, i.e.,

$$H(C_i) = \sum_{k \in \{true,\, false\}} p(C_i = k) \log p(C_i = k) \quad \text{-(2)}$$

V. COMMUNITY WEB DIRECTORY DISCOVERY ALGORITHMS

In this section, we describe the pattern discovery methodology that we propose for the construction of the community Web directories. This methodology aims at the selection of categories of the Web directory that satisfy the criteria mentioned in the previous sections, i.e., community as well as objective informativeness. The selected categories are used to construct the subgraph of the community Web directory. The input to the pattern discovery algorithms is the user sessions, which have been mapped to thematic user sessions. We note that there is a many-to-one mapping of the Web pages to the leaf categories of the Web directory. In other words, more than one Web page within a user session can be mapped to the same leaf category. Thus, the number of distinct entries in a thematic user session u_i is generally smaller than its simple counterpart v_i that contains Web pages. The removal of duplicates leads to the thematic session set which is defined as follows:

A. Objective Community Directory Miner (OCDM):

The first machine learning method that we employed for community discovery is the CDM algorithm. The enhanced version of CDM, incorporating the OCIA criterion, is named as Objective-Community Directory Miner (OCDM). Similar to CDM, OCDM is based on the cluster mining algorithm which has been employed earlier [4] for site-specific community discovery. Cluster mining discovers patterns of common behavior by looking for all maximal fully connected subgraphs (cliques) of a graph that represents the users' characteristic features, i.e., thematic categories in our case. The algorithm starts by constructing the graph, the vertices of which correspond to the categories, while the edges to category co-occurrence in thematic session sets.

Vertices and edges are associated with weights, which are computed as the category occurrence and co-occurrence frequencies, respectively. The connectivity of the graph is usually high. For this reason, we make use of a connectivity threshold that reduces the edges of the graph. This threshold is related to the frequency of co-occurrence of the thematic categories in the data. Once the connectivity of the graph has been reduced, the weighted graph is turned to an unweighted one. Finally, all maximal cliques of the unweighted graph are generated, each one corresponding to a community model. One important advantage of this approach is that each user may be assigned to many communities, unlike most crisp user clustering methods. Moreover, the clusters generated by OCDM group together characteristic features of the user. Each clique discovered by OCDM is thus already a community model, i.e., a set of interesting categories.

B. Objective Probabilistic Directory Miner (OPDM):

A powerful statistical methodology for identifying latent factors in data is PLSA. Similar to the approach that we followed for the OCDM algorithm, we recall the relation $\Re = (\overline{U}, C)$, with the pair $(\overline{u}_k, c_j) \in \Re$ representing the access c_j of category c_j or one of its leaf subcategories during the thematic session set $\overline{u}_i$. The PLSA model is based on the assumption that there exists a set $Z = \{z_1, z_2, \ldots, z_k\}$ of latent factors such that each instance $(\overline{u}_k, c_j) \in \Re$, i.e., each observation of a certain category inside a thematic session tree, is related to a latent factor $z_k \in Z$. To further formalize our assumption, we define the following probabilities: $p(\overline{U}_i)$, the a priori probability of the thematic session set $\overline{u}_k$ i.e., the number of times the session appears in the usage data; $p(Z_k | \overline{U}_i)$, the conditional probability of the latent factor z_k motivating the observation of session $\overline{u}_k$ and $p(C_j | Z_k)$, the conditional probability of category c_j being accessed, given the latent factor z_k. We can describe a probabilistic model for generating the categories "observed" in sessions as follows: select a thematic session set with probability $p(\overline{U}_i)$, select a latent factor z_k with probability $p(Z_k | \overline{U}_i)$, and select a category c_j with probability $p(C_j | Z_k)$. The results of the above process allow us to estimate the probability of observing a particular (session-category) pair $(\overline{u}_k, c_j)$, using joint probabilities and Bayes's theorem as follows:

$$p(\overline{U}_i C_j) = \sum_k p(Z_k)\, p(\overline{U}_i | Z_k)\, p(C_j | Z_k). \quad (3)$$

C. Objective Clustering and Probabilistic Directory Miner (OCPDM):

In addition to the enhanced methods presented above, we also introduce here a new hybrid method for the discovery of community models. This method combines a clustering algorithm with PLSA. We apply the popular k-means clustering algorithm on the relation $\Re = (\overline{U}, C)$ for the creation of the initial communities. This approach differs from CDM clustering, as it produces nonoverlapping clusters, i.e., each category belongs to only one cluster. However, as we have explained above for PLSA, the explicit modeling of latent factors is considered advantageous. The new algorithm Objective Clustering and Probabilistic Directory Miner (OCPDM) invokes OPDM for each of the K clusters. In particular, the categories of the cluster on which each latent factor has the maximum impact are selected using the LFAP threshold.

Finally, the OCIA criterion is also applied. This process leads to a more "specialized" community Web directory than the one that would be produced by k-means. The community Web directory includes only the most "dominant" categories of the community, i.e., the categories that satisfy not only the explicit associations, but also the hidden knowledge that exists in the data. The categories that do not satisfy these associations are dropped from the community model.

D. Community Web Directory Refinement:

The result of the aforementioned pattern discovery methods is a hierarchy that corresponds to the community Web directory, i.e., to a prototypical model for each community, which is representative of the participating users. The construction of the directory is based on the

selection of the categories by each algorithm and their mapping onto the original Web directory. However, the construction of useful community Web directories needs to go beyond the selection of categories by the pattern discovery algorithms.

VI. PROPOSED METHODOLOGY

The proposed system is designed to perform web directory optimization using the classification techniques. The probabilistic latent semantic analysis algorithm is used for the classification process. The fuzzy logic technique is used to enhance the PLSA scheme for weight optimization. The web directory optimization model uses ISP based user access logs. The real web directories are constructed using the usage data values. Access sequence weight is estimated with the user data and frequency values. The access sequence based weight estimation model is enhanced with fuzzy weight assignment models. The probabilistic latent semantic analysis (PLSA) algorithm is used to classify the directory entries. The system is divided into four modules. They are Usage Data Analysis, Access Patterns, Weight Estimation and Directory Construction.

The usage data analysis module is designed to perform session conversion and preprocessing tasks. The access pattern extraction module is designed to group up all the access information using the session information. The weight estimation module is used to assign frequency based weight values. The directory construction module is designed to build the web directory based on the classification scheme.

A. Usage Data Analysis:

The user access log maintains the page access history of a particular web site or search engine. The search engine log maintains the access information for all sites that are visited through the search engine information. The web site based log maintains the information about the page access entry for the current web site. This system uses the search engine logs to process access information for all sites.

The usage log analysis module is divided into three sub modules. They are log preprocessing, session conversion and session analysis modules. The log preprocess module is designed to remove noisy and redundant data values. The noisy data refers the incomplete data entries in the web access logs. Same page requests are updated multiple times in case of refresh process. The redundant entries are also removed from the log files. The log entries are maintained with unique session identification values. The session values are grouped into single entry for each session. The session analysis is representing the session summary information.

B. Access Patterns:

The session information shows the user visiting sequence for a user in a session. The session entries are mapped into single graph format. The session entries are mapped with relationship sequence. The community values are identified with respect to the session mapping process. The community information is arranged into tree structure manner. The access pattern reflects the user interest on the web sites. The positive and negative factors are also considered in the pattern identification process.

C. Weight Estimation:

The weight estimation module is designed to assign the weight values for the access sequence structure. The weight values are assigned with respect to the link values. The weight values are used in the latent semantic and probabilistic latent semantic analysis operations. The weight values are modified due to dynamic entry behaviors of the web logs. The fuzzy logic based weight assignment model is used in the system. The weight values are converted into fuzzy weight model. The weight models are statistical model. The fuzzy weights are used in the classification process.

D. Directory Construction:

The directory construction module is designed to build the web directory based on the access information. The access sequence information are extracted from the web user logs. The log values are assigned with fuzzy weights for each session and patterns. The probabilistic latent semantic analysis algorithm is used with fuzzy weights. Directory entries are maintained in a tree-structured model. The distance between the patterns I used to separate the directory values.

VII. CONCLUSION

This paper advocates the concept of a community Web directory, as a Web directory that specializes to the needs and interests of particular user communities. Furthermore, it presents the complete methodology for the construction of such directories with the aid of machine learning methods. User community models take the form of thematic hierarchies and are constructed by employing clustering and probabilistic learning approaches. Search engine log based web directory construction scheme is proposed in the system. Fuzzy weight based directory optimization model is used to assign weights for the web pages. The system assists the user for fast information retrieval. Automated web directory construction and visualization scheme enables web personalization.

VIII. REFERENCES

[1]. P.I. Hofgesang, “Online Mining of Web Usage Data: An Overview,” Web Mining Applications in E-Commerce and E-Services, pp. 1-24, Springer, 2009.

[2]. G. Castellano, A.M. Fanelli, and M.A. Torsello, “Computational Intelligence Techniques for Web Personalization,” Web Intelligence and Agent Systems, vol. 6, no. 3, pp. 253-272, 2008.

[3]. X. Jin, Y. Zhou, and B. Mobasher, “Task-Oriented Web User Modeling for Recommendation,” Proc. 10th Int’l Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 109-118, 2005.

[4]. G. Paliouras and C.D. Spyropoulos, “Discovering User Communities on the Internet Using Unsupervised Machine Learning Techniques,” Interacting with Computers J., vol. 14, no. 6, pp. 761-791, 2002.

[5]. G. Xu, Y. Zhang, and Y. Xun, “Modeling User Behaviour for Web Recommendation Using lda Model,” Proc. IEEE/WIC/ACM Int’lConf. Web

Intelligence and Intelligent Agent Technology, pp. 529-532, 2008.

[6]. W. Chu and T.P. Park, "Personalized Recommendation on Dynamic Content Using Predictive Bilinear Models," Proc. 18th Int'l Conf. World Wide Web, pp. 691-700, 2009.

[7]. The Adaptive Web, Methods and Strategies of Web Personalization, P. Brusilovsky, A. Kobsa, and W. Neijdl, eds. Springer, 2007.

[8]. D. Chen, D. Wang, and F. Yu, "A PLSA-Based Approach for Building User Profile and Implementing Personalized Recommendation," Proc. Joint Ninth Asia-Pacific Web Conf. (APWeb '07) and Eighth Int'l Conf. Web-Age Information Management (WAIM '07), pp. 606-613, 2007.

[9]. D. Pierrakos and G. Paliouras, "Exploiting Probabilistic Latent Information for the Construction of Community Web Directories," Proc. 10th Int'l Conf. User Modeling, L. Ardissono, P. Brna, and A. Mitrovic, eds., pp. 89-98, 2005.

[10]. C. Christophi, and G. Paliouras, "Automatically Annotating the ODP Web Taxonomy," Proc. 11th Panhellenic Conf. Informatics (PCI '07), 2007.

[11]. B. Mehta and N. Wolfgang, "Unsupervised Strategies for Shilling Detection and Robust Collaborative Filtering," User Modeling and User-Adapted Interaction, vol. 19, nos. 1/2, pp. 65-97, 2009.

[12]. T. Oishi and M. Koshimura, "Personalized Search Using odp-Based User Profiles Created from User Bookmark," Proc. 10th Pacific Rim Int'l Conf. Artificial Intelligence, pp. 839-848, 2008.

[13]. T. Dalamagas, and T. Sellis, "Mining User Navigation Patterns for Personalizing Topic Directories," Proc. Ninth Ann. ACM Int'l Workshop Web Information and Data Management, pp. 81-88, 2007.

[14]. Dimitrios Pierrakos and Georgios Paliouras "Personalizing Web Directories with the Aid of Web Usage Data" IEEE Transactions on knowledge and data engineering, vol. 22, no. 9, September 2010.