Global Journal of Computer Science and Technology
Volume 11 Issue 7 Version 1.0 May 2011
Type: Double Blind Peer Reviewed International Research Journal
Publisher: Global Journals Inc. (USA)
ISSN: 0975-4172 & Print ISSN: 0975-4350

Web Page Prediction for Web Personalization: A Review

By R. Khanchana, Dr. M. Punithavalli
Karpagam University

Abstract- This paper proposes a survey of Web Page Ranking for web personalization. Web page prefetching has been widely used to reduce the access latency problem of the Internet. However, if most prefetched web pages are not visited by the users in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently and may worsen the access delay problem. Therefore, it is critical that we have an accurate prediction method during prefetching. The technique like Markov models have been widely used to represent and analyze user's navigational behavior (usage data) in the Web graph, using the transitional probabilities between web pages, as recorded in the web logs. The recorded users' navigation is used to extract popular web paths and predict current users' next steps.

Keywords: *Web Personalization, Page Ranking, User Browsing, Markov Model.*

GJCST Classification: *H.3.5*

Strictly as per the compliance and regulations of:

Web Page Prediction for Web Personalization: A Review

R. Khanchana[α], Dr. M. Punithavalli[Ω]

Abstract- This paper proposes a survey of Web Page Ranking for web personalization. Web page prefetching has been widely used to reduce the access latency problem of the Internet. However, if most prefetched web pages are not visited by the users in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently and may worsen the access delay problem. Therefore, it is critical that we have an accurate prediction method during prefetching. The technique like Markov models have been widely used to represent and analyze user's navigational behavior (usage data) in the Web graph, using the transitional probabilities between web pages, as recorded in the web logs. The recorded users' navigation is used to extract popular web paths and predict current users' next steps.

Keywords- Web Personalization, Page Ranking, User Browsing, Markov Model.

I. Introduction

A portion of Data mining, which resolves around the assessment of the World Wide Web, is known as web mining. Data mining, Internet Technology, World Wide Web as well as semantic web, are incorporated in web mining. Web mining refers to the use of data mining techniques to automatically discover and extract information form world wide web documents and services. Web mining has been classified into three areas such as Web content mining, Web structure mining and Web usage mining. The most common applications include the ranking of the results of a web search engine and the provision of recommendations to users of (usually commercial) web sites, known as web personalization. Even with the speed of today's Internet, web latency is still one of the major concerns of its users. Reducing latency is particularly important for online businesses, since if their web pages cannot be opened within about eight seconds, they might lose customers. Web servers collect huge amount of data everyday. Users search any information, that relevant data is prefetched from web server.

About[α]- Research Scholar, Karpagam University, Coimbatore, Tamilnadu,India-641 021.
E-mail : kanchusri@gmail.com
About[Ω]- Director and Head, SNS college, Bharathiar University, Coimbatore, Tamilnadu, India.
E-mail : mpunitha_srcw@yahoo.co.in

However, if most prefetched web pages are not visited by the user in their subsequent accesses, the limited network bandwidth and server resources will not be used efficiently and may worsen the access latency problem. The objective of a Web personalization system is to "provide users with the information they want or need, without expecting from them to ask for it explicitly" [14]. The most common approaches used for web user browsing pattern prediction are Markov model, sequential association rules and clustering. PageRank is used in order to rank web pages based on the results returned by a search engine after a user query. The ranking is performed by evaluating the *importance* of a page in terms of its connectivity to and from other important pages. In the context of navigating a web site, a page/path is *important* if many users have visited it before, we propose a novel approach that is based on a personalized version of PageRank, applied to the navigational tree created by the previous users' navigations. A new technique was proposed by Page and Brin called PageRank to compute the importance of Web pages. PageRank [2] determines the significance of Web pages and helps a search engine to choose high quality pages more efficiently.

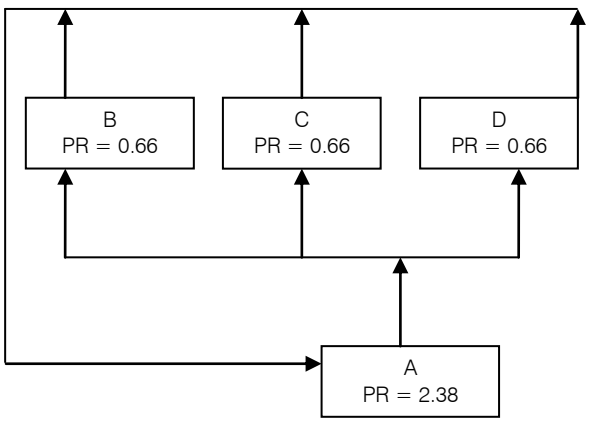

Fig.1 Web Page Linking and their Page Ranks

In the above example we consider the average page rank for the site is 1. The page B, C and D has the page ranks like 0.66 but A has 2.38.The links are well done such that we can navigate in and out of every page. The PR has also been distributed in favour of page A which has PR 2.38. This knowledge is then used

from the system in order to personalize the site according to each user's behavior and profile.

Fig 2 shows a multi-tiered Web site and the caching and personalization techniques suitable for each Web site component. The caching levels show that performance is maximized when cache hits occur close to the browsers. For example, at the ISP and router levels, rule-based and simple filtering may offer sufficient personalization capabilities for a relatively small investment of effort. When more is needed or wanted, more complex techniques can be implemented. When data mining is needed to develop business intelligence and offer highly sophisticated personalization, the processing occurs at the database layer.

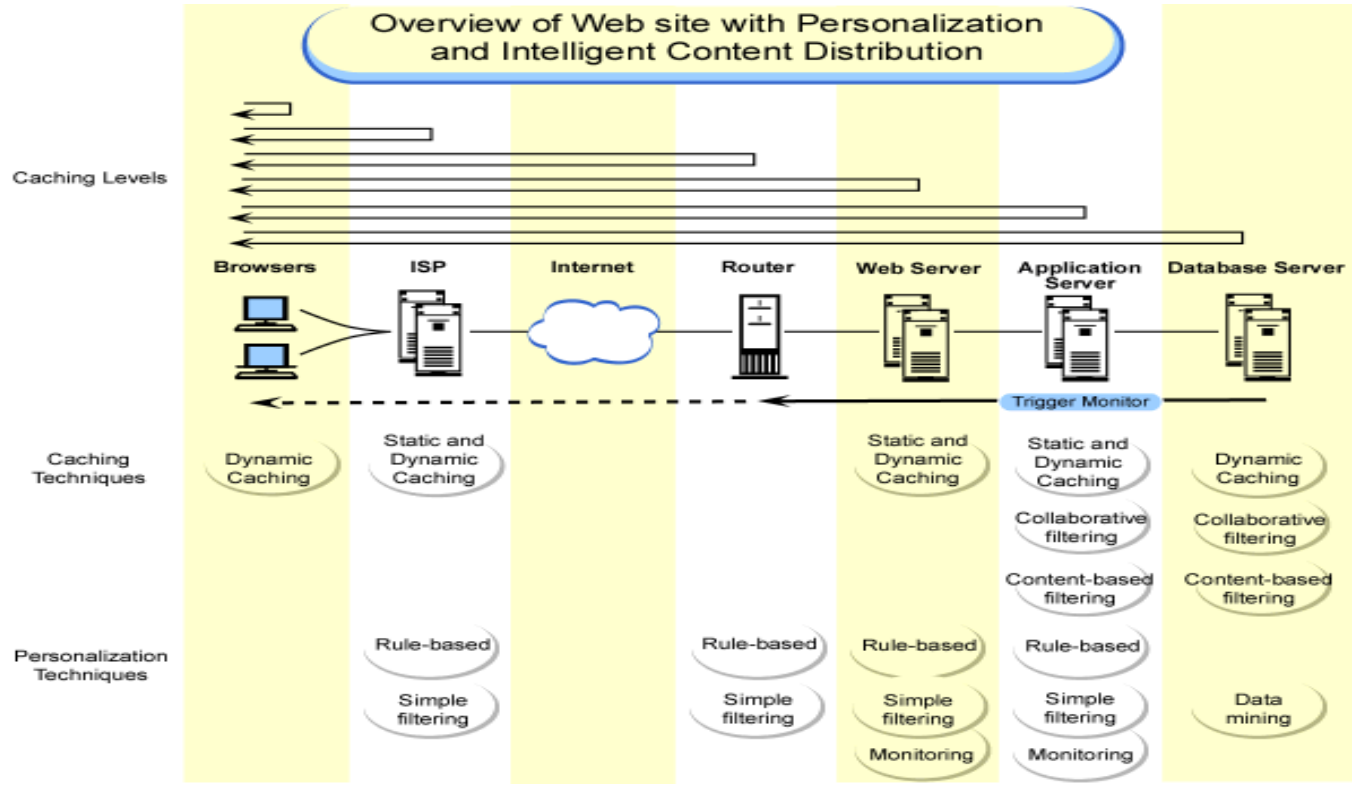

Fig 2. Overview of Web site with personalization and intelligent content distribution

The basic personalization techniques are

1. Rule Based
2. Simple Filtering
3. Content- Based Filtering
4. Collaborative Filtering

Rule Based: Rule-based techniques provide a visual editing environment for the business administrator to specify business rules to drive personalization.

Simple Filtering: Simple filtering relies on predefined groups, or classes, of visitors to determine what content is displayed or what service is provided.

Content- Based Filtering: Content-based filtering works by analyzing the content of the objects to form a representation of the visitor's interests.

Collaborative Filtering: Collaborative filtering collects visitors' opinions on a set of objects, using either explicit or implicit ratings, to predict a particular visitor's interest in an item.

Caching techniques have long been used to improve the system performance. With content caching, frequently accessed pages do not need to be retrieved remotely or materialized at the server for each access. This can significantly reduce the latency for obtaining Web pages, as well as reduce the load on the server and network. In the Web environment, frequently accessed Web pages can be cached at the client browser, proxy servers, and server caches. For caching to be effective, data needs to be reused frequently. Personalization is a process of gathering and storing information about site visitors, analyzing the information, and, based on the analysis, delivering the right information to each visitor at the right time. It is a key technology needed in various e-business applications. The elements of Personalization system includes

- Identify site visitor
- Retrieve visitor's profile(Id, Password, interest, role, business needs etc)
- Select content that matches visitor's preferences
- Retrieve content and assemble page for display to visitor

Principal elements of Web personalization include (a) the categorization and preprocessing of Web data, (b) the extraction of correlations between and across different kinds of such data, and (c) the determination of the actions that should be recommended by such a personalization system [13]. In this work we focus on Web usage mining. This process

relies on the application of statistical and data mining methods to the Web log data, resulting in a set of useful patterns that indicate users' navigational behavior. The data mining methods that are employed are: association rule mining, sequential pattern discovery, clustering, and classification. This knowledge is then used from the system in order to personalize the site according to each user's behavior and profile. Today, personalization is increasingly used as a means to expedite the delivery of information to a visitor, making the site useful and attractive to return to.

a) *Data Preprocessing*

An extensive description of data preparation and preprocessing methods can be found in. The data preprocess includes three basic steps like

- Data Cleaning
- User Identification
- Session Identification

The first issue in the preprocessing phase is data cleaning. Depending on the application, Web log data may need to be cleaned from entries involving pages that returned an error or graphics file accesses. In some cases such information might be useful, but in others such data should be eliminated from a log file.

Most important of all is the user identification issue. There are several ways to identify individual visitors. The most obvious solution is to assume that each IP address (or each IP address/client agent pair) identifies a single visitor. Nonetheless, this is not very accurate because, for example, a visitor may access the Web from different computers, or many users may use the same IP address (if a proxy is used). A further assumption can then be made, that consecutive accesses from the same host during a certain time interval come from the same user. More accurate approaches for a priori identification of unique visitors are the use of cookies or similar mechanisms or the requirement for user registration. However, a potential problem in using such methods might be the reluctance of users to share personal information.

The next step is to perform session identification, by dividing the click stream of each user into sessions. The usual solution in this case is to set a minimum timeout and assume that consecutive accesses within it belong to the same session, or set a maximum timeout, where two consecutive accesses that exceed it belong to different sessions.

b) *User Profiling*

User profiling is the process of collecting information about the characteristics, preferences, and activities of a Web site's visitors. This can be accomplished either explicitly or implicitly. Explicit collection of user profile data is performed through the use of online registration forms, questionnaires, and the like. The methods that are applied for implicit collection of user profile data vary from the use of cookies or similar technologies to the analysis of the users' navigational behavior that can be performed using Web usage mining techniques.

The extraction of information concerning the navigational behavior of Web site visitors is the objective of Web usage mining. Nevertheless this process can also be regarded as part of the creation of user profiles; it is therefore evident that those two modules overlap and are fundamental in the Web personalization process. A user profile can be either static, when the information it contains is never or rarely altered (e.g., demographic information), or dynamic when the user profile's data change frequently. Such information is obtained either explicitly, using online registration forms and questionnaires resulting in static user profiles, or implicitly, by recording the navigational behavior and/or the preferences of each user, resulting in dynamic user profiles. In the latter case, there are two further options: either regarding each user as a member of a group and creating aggregate user profiles, or addressing any changes to each user individually. When addressing the users as a group, the method used is the creation of aggregate user profiles based on rules and patterns extracted by applying Web usage mining techniques to Web server logs.

II. Related Works

Lamberti *et al.* proposed a relation-based page rank algorithm for Semantic Web search engines [10]. With the incredible increase of data available to end users through the Web, search engines come to play ever a more critical role. However, due to their general-purpose approach, it is always less uncommon that obtained result sets provide a burden of useless pages. The next-generation Web architecture [8], characterized by the Semantic Web, provides the layered architecture possibly allowing overcoming this limitation. Many search engines have been proposed, that allow increasing data retrieval accuracy by exploiting a key content of semantic Web resources, that is, relations. On the other hand, in order to rank results, the majority of the existing solutions need to work on the whole annotated knowledge base. In this paper, the author proposed a relation-based page rank algorithm to be used in conjunction with semantic Web search engines that simply relies on information that could be extracted from user queries and on annotated resources. Relevance is calculated as the probability that a retrieved resource actually contains those relations whose existence was assumed by the user at the time of query definition.

Ranking web pages using machine learning approaches is put forth by Sweah *et al* [19]. One of the key components which guarantee the acceptance of web search service is the web page ranker - a

component which is said to have been the main contributing factor to the early successes of Google. It is well recognized that a machine learning method such as the Graph Neural Network (GNN) can be able to learn and estimate Google's page ranking algorithm. This paper demonstrates that the GNN can successfully learn many other Web page ranking methods [7] e.g. TrustRank, HITS and OPIC. Experimental results illustrate that GNN may be suitable to learn any arbitrary web page ranking scheme, and hence, may be more flexible than any other existing web page ranking scheme. The significance of this inspection lies in the fact that it is possible to learn ranking schemes for which no algorithmic solution exists or is known.

Shohel Ahmed *et al.* proposed a personalized URL re-ranking method based on psychological characteristics of users browsing like “common-mind,” “uncommon-mind,” and “extremely uncommon-mind” [17]. This personalization method constructs an index of the anchor text retrieved from the web pages that the user has clicked during his/her past searches. Our method provides different weights to the anchor text according to the psychological characteristics for re-ranking URLs.

Srour *et al.* provided a personalized Web Page ranking using trust and similarity. Search engines, like Google, utilize link structure to rank web pages [18]. Although this technique offers an objective global estimate of the web page importance, it is not targeted to the specific user preferences. This paper presents a new technique for the personalization of the results of a search engine based on the user's taste and preferences. The idea of trust and similarity, obtained from explicit user input and implicit user behavioral patterns, are used to compute personalized page rankings [6].

Shiguang *et al.* given the improvement of page ranking algorithm based on timestamp and link [16]. The conventional ranking technique favors the old pages, which makes old pages always emerge on the top of the ranking results when pages are ranked according to the dynamic web by the static ranking algorithm. Therefore, this paper proposes a temporal link - analysis technique to overcome the problem. This technique uses the last variation time that returned by the HTTP response as the timestamp of nodes and links concerned. And the weight of the in-link and out-link are also combined to calculate the overall weight of the pages. Using the WTPR technique can make the old pages decline and new pages rise in the ranking result, meanwhile it can help the old pages which have high-quality get higher rank value than common old pages.

Kritikopoulos *et al.* proposed Wordrank: a method for ranking web pages based on content similarity [9]. This paper presents WordRank, a new page ranking system, which utilize similarity between interconnected pages. WordRank establishes the model of the biased surfer which is based on the following hypothesis: "the visitor of a Web page have a tendency to visit Web pages with similar content rather than content irrelevant pages". This technique modifies the random surfer model by biasing the probability of a user to follow a link in favor of links to pages with similar content. It is the perception that WordRank is most suitable in topic based searches, since it prioritizes strongly interconnected pages, and in the same time is more robust to the multitude of topics and to the noise produced by navigation links. This paper provides preliminary experimental verification from a search engine developed for the Greek fragment of the World Wide Web.

Magdalini Eirinaki *et al.* present a hybrid probabilistic predictive model extending the properties of Markov models by incorporating link analysis methods [12]. More specifically, we propose the use of a PageRank-style algorithm for assigning prior probabilities to the web pages based on their importance in the web site's graph. We prove, through experimentation, that this approach results in more objective and representative predictions than the ones produced from the pure usage-based approaches.

III. Markov Models

The 1st-order Markov models (Markov Chains) provide a simple way to capture sequential dependence [3], but do not take into consideration the “long-term memory” aspects of web surfing behavior since they are based on the assumption that the next state to be visited is only a function of the current one. Higher-order Markov models [11] are more accurate for predicting navigational paths, there exists, however, a trade-off between improved coverage and exponential increase in statespace complexity as the order increases. Moreover, such complex models often require inordinate amounts of training data, and the increase in the number of states may even have worse prediction accuracy and can significantly limit their applicability for applications requiring fast predictions, such as web personalization. There have also been proposed some mixture models that combine Markov models of different orders. Such models, however, require much more resources in terms of preprocessing and training. It is therefore evident that the final choice that should be made concerning the kind of model that is to be used, depends on the trade-off between the required prediction accuracy and model’s complexity/size.

Hidden markov model (HMM) are generative, directed graphical models, which describe the joint probability over a state sequence and output sequence. Such generative models make limiting independence assumptions over the output sequence.

Mixed Markov models are based on the selection of parts from Markov models of different order,

so that the resulting model has reduced state complexity as well as increased precision in predicting the user's next step. Deshpande and Karypis *et al.* [5] propose the All-Kth-Markov models, presenting 3 schemas for pruning the states of the All-K*th*-Order Markov model. Cadez *et.al* [4] as well as Sen and Hansen *et al.* [15] also proposed the use of mixed Markov models. A different approach is that of [1] Acharyya and Ghosh *et al*, who use concepts, to describe the web site. Each visited page is mapped to a concept, imposing a tree hierarchy on these topics. A semi-Markov process is then defined on this tree based on the observed transitions among underlying visited pages. They prove that this approach is computationally much less demanding compared to using higher order Markov models.

IV. Conclusion

The explosive growth of information sources available on the World Wide Web has necessitated the users to make use of automated tools to locate desired information resources and to follow and asses their usage pattern. Web page prefetching has been widely used to reduce the access latency problem of the internet, its success mainly relies on the accuracy of web page prediction. Markov model is the most commonly used prediction model because of its high accuracy. Low order markov models have higher accuracy and lower coverage. The higher order models have a number of limitations associated with

i) Higher state complexity,
ii) Reduced coverage,
iii) Sometimes even worse prediction accuracy.

To overcome these limitations of higher order markov model by Hidden markov model. It is a powerful method for labeling sequence data but it has two major drawbacks such as one stemming from its independence assumptions and the other from its generative nature.

We have discussed some of the techniques to overcome the issues of web page change ranking. As the web is going to expand, web usage in web databases will become more and more and the rank prediction is also more difficult. The above findings will become will be good guide to rank the web pages effectively. In this paper, we have presented a comprehensive survey of up-to-date researchers of ranking web pages for web personalization. Besides, a brief introduction about web mining, web personalization and web page change ranking have also been presented. However, research of the web page ranking is just at its beginning and much deeper understanding needs to be gained.

V. Future Work

This survey paper intends to aid upcoming researchers in field of web page ranking for web personalization to understand the available methods and help to perform their research in right direction. For future work, there are some improvements that can be implemented. First, the first-order Markov models (Markov Chains) provide a simple way to capture sequential dependence, but do not take into consideration the "long-term memory" aspects of web surfing behavior since they are based on the assumption that the next state to be visited is only a function of the current one. Higher-order Markov models and hidden markov models are more accurate for predicting navigational paths, there exists, however, a trade-off between improved coverage and exponential increase in state space complexity as the order increases. Secondly, to predict web page ranks efficiently by doing the preprocessing phase effectively.

References Références Referencias

1. S. Acharyya and J. Ghosh. "Context-Sensitive Modeling of Web- Surfing Behaviour Using Concept Trees, *in Proc. Of the 5th WEBKDD Workshop,* Washington DC, August 2003.
2. M. S. Aktas, M.A. Nacar and F. Menczer. Personalizing Page Rank Based on domain Profiles, *Processing of WEBKDD 2004* Workshop, 2004.
3. S. Brin and L.Page. The anatomy of a large-scale hypertextual Web search engine, *in Proc. of the 7th International World Wide Web Conference (WWW7),* Brisbane, 1998.
4. I. Cadez, S. Gaffney and P.Smyth. A general probabilistic framework for clustering individuals and objects*, in Proc. Of the 6th ACM SIGKDD Conference*, Boston, 2000.
5. M. Deshpande, and Karypis. Selective Markov Models for Predicting Web-Page Accesses, *Proc. of the 1st SIAM International Conference on Data Mining*, 2001.
6. T. H Haveliwala. Topic-sensitive PageRank, *Processing of WWW,* 2002.
7. H. Y kao and S.Flin. A Fast PageRank Convergence Method based on the Cluster rediction, *Proc. Of IEEE/WIC/ACM International Conference on Web Intelligence*, 2007.
8. M. Y Kan and H.O.N Thi. Fast webpage classification using URL features, Processing of CIKM, 2005.
9. A.Kritikopoulos, M. Sideri and I. Varlamis. WordRank: A Method for Ranking Web Pages Based on Content Similarity, *British National Conference on Databases,* pp. 92-100, 2007.

10. F.Lamberti, A. Sanna, and C. Demartini. A Relation-Based Page Rank Algorithm for Semantic Web Search Engines, *IEEE Transactions on Knowledge and Data Engineering*, vol. 21, pp. 123-136, 2009.
11. M. Levene and G. Loizou. Computing the Entropy of User Navigation in the Web, *in Intl. Journal of Information Technology and Decision Making,* 2:459-476, 2003.
12. Magdalini Eirinaki, Michalis Vazirgiannis, Dimitris Kapogiannis. Web path Recommendations Based on Page Ranking and Markov Models, *WIDM'05*, November 05, 2005.
13. B. Mobasher, R. Cooley and J.Srivastava. Automatic personalization based on web usage mining,. *Commun. ACM,* 43, 8 (August), 142–151, 2000a.
14. M. D. Mulvenna, S. S. Anand and A.G Buchner. Personalization on the net using web mining,. *Commun*. ACM, 43, 8 (August), 123–125, 2000.
15. R. Sen and M. Hansen. Predicting a Web user's next accessbased on log data, *in Journal of Computational Graphics and Statistics,* 12(1):143-155, 2003.
16. Shiguang Ju, Zheng Wang and Xia Lv. Improvement of Page Ranking Algorithm Based on Timestamp and Lin, *International Symposiums on Information Processing* (ISIP), pp. 36-40, 2008.
17. Shohel Ahmed, Sungjoon Park, Janson, J. Jung and Sanggil Kang. A Personalized URL Re-ranking ethod using Psychological User Browsing Characteritics, *Journal of Universal Computer Science*, vol.15, no.4,2009.
18. L. Srour, A. Kayssi and A. Chehab. Personalized Web Page Ranking Using Trust and Similarity, *in Intl. Conference on Computer Systems and Applications*, pp. 454-457, 2007.
19. Sweah Liang Yong, M. Hagenbuchner. M and ah chung Tsoi. Ranking Web Pages Using Machine Learning Approaches*, in Intl. Conference on Web Intelligence and Intelligent Agent Technology,* vol. 3, pp. 677-680, 2008.