Volume 4, Issue 3, March 2014 ISSN: 2277 128X
International Journal of Advanced Research in Computer Science and Software Engineering
Research Paper
Available online at: www.ijarcsse.com

A Survey on: Methods of Web Behavior Prediction by: Utilizing Different Features

J.P. Maurya, Sandeep Singh
A.P.(CSE)
ICOT, Bhopal, India

Harish Patil, Pinki Jain
M.Tech [CSE]
ICOT, Bhopal, India

Abstract— 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. There are many technique which 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. In this paper, the study of web features has been done which are acting as tool in different methods for next page prediction.

Index Terms— Information Extraction, Text Analysis, Ontology, feature extraction, text categorization, clustering

I. INTRODUCTION

To facilitate web page access by users, web recommendation model is needed. So the Interest in the analysis of user behavior on the Web has been increasing rapidly. This increase stems from the realization that added value for Web site visitors is not gained merely through larger quantities of data on a site,but through easier access to the required information at the right time and in the most suitable form. Estimates of Web usage expect the number of users to climb up to 945million by 2004 (Computer Industry Almanac, May 2003). The majority of these users are non-expert and it difficult to keep up with the rapid development of computer technologies, while at the same time they recognize that the Web is an invaluable source of information for their everyday life. The increasing usage of the Web also accelerates the pace at which information becomes available online. In various surveys of the Web, it is estimated that roughly one million new pages are added every day and over 600 GB of pages change per month. A New Web server providing Web pages is emerging every two hours. Nowadays, more than three billion Web pages are available online almost one page for every two people on the earth (Usa Today, April 2003). In the above, one notices the emergence of a spiral e!ect, i.e. increasing number of users causing an increase in the quantity of online information, attracting even more users, and so on. This pattern is responsible for the 'explosion' of the Web, which causes the frustrating phenomenon known as 'information overload to Web users.

With the growing popularity of the World Wide Web, A large number of users access web sites in all over the world. When user access a websites, a large volumes of data such as addresses of users or URLs requested are gathered automatically by Web servers and collected in access log which is very important because many times user repeatedly access the same type of web pages and the record is maintained in log files. These series of accessed web pages can be considered as a web access pattern which is helpful to find out the user behavior. Through this behavior information, we can find out the accurate user next request prediction that can reduce the browsing time of web page thus save the time of the user and decrease the server load.In recent years, there has been a lot of research work done in the field of web usage mining „ Future request prediction". The main motivation of this study is to know the what research has been done on Web usage mining in future request prediction.

In Web prediction, main challenges are in both preprocessing and prediction. Preprocessing challenges include handling large amount of data that cannot fit in the computer memory, choosing optimum sliding window size, identifying sessions, and seeking/extracting domain knowledge. Prediction challenges include long training/prediction time, low prediction accuracy, and memory limitation.

II. Web Features

Web data mining is the process of applying data mining techniques to Web data. Research in this area has the objectives of helping e-commerce businesses in their decision making, assisting in the design of good Web sites and assisting the user when navigating the Web. The World Wide Web data mining focuses on three issues: Web structure mining, Web content mining and Web usage mining. The classification is based on two aspects: the purpose and the data sources. Mining research concentrates on finding new information or knowledge in the data. On the basis of above

information, Web mining can be divided into web structure mining, web content mining, and web usage mining as shown in Figure 1.

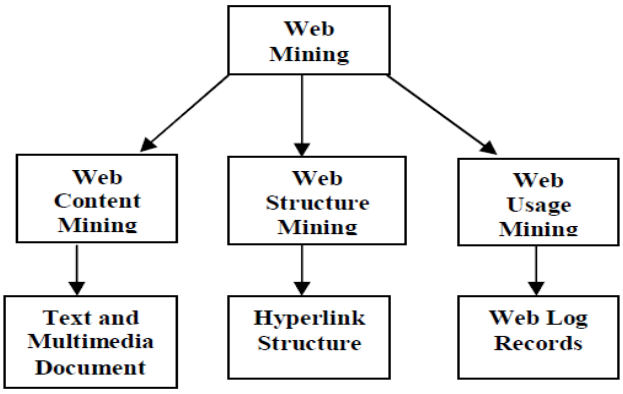

Fig.1 Types of Web Mining

Web Structure Mining

Most of the Web information retrieval tools only use the textual information, while ignore the link information that could be very valuable. The goal of Web structure mining is to generate structural summary about the Web site and Web page. Technically, Web content mining mainly focuses on the structure of inner-document, while Web structure mining tries to discover the link structure of the hyperlinks at the inter-document level. Based on the topology of the hyperlinks, Web structure mining will categorize the Web pages and generate the information, such as the similarity and relationship between different Web sites.

Web structure mining can also have another direction – discovering the structure of Web document itself. This type of structure mining can be used to reveal the structure (schema) of Web pages; this would be good for navigation purpose and make it possible to compare/integrate Web page schemes. This type of structure mining will facilitate introducing database techniques for accessing information in Web pages by providing a reference schema. What is on earth the structural information, and how to discover, [9] gave a detailed description about how to discover interesting and informative facts describing the connectivity in the Web subset, based on the given collection of interconnected web documents. The structural information generated from the Web structure mining includes the follows: the information measuring the frequency of the local links in the Web tuples in a Web table; the information measuring the frequency of Web tuples in a Web table containing links that are interior and the links that are within the same document; the information measuring the frequency of Web tuples in a Web table that contains links that are global and the links that span different Web sites; the information measuring the frequency of identical Web tuples that appear in a Web table or among the Web tables.

In general, if a Web page is linked to another Web page directly, or the Web pages are neighbours, we would like to discover the relationships among those Web pages. The relations maybe fall in one of the types, such as they related by synonyms or ontology, they may have similar contents, and both of them may sit in the same Web server therefore created by the same person. Another task of Web structure mining is to discover the nature of the hierarchy or network of hyperlinks in the Web sites of a particular domain. This may help to generalize the flow of information in Web sites that may represent some particular domain; therefore the query processing will be easier and more efficient.

Web structure mining has a nature relation with the Web content mining, since it is very likely that the Web documents contain links, and they both use the real or primary data on the Web. It's quite often to combine these two mining tasks in an application. Web structure mining aims at generating structured summary about web sites and web pages in order to identify relevant documents. The focus here is on link information, which is an important aspect of Web data. Web structure mining can be used to reveal the structure or schema of Web pages which would facilitate Web document classification and clustering on the basis of its structure.

Web Content Mining

Web content mining describes the automatic search of information resource available online [1], and involves mining web data contents. In the Web mining domain, Web content mining essentially is an analogous of data mining techniques for relational databases, since it is possible to find similar types of knowledge from the unstructured data residing in Web documents. The Web document usually contains several types of data, such as text, image, audio, video, metadata and hyperlinks. Some of them are semi-structured such as HTML documents or a more structured data like the data in the tables or database generated HTML pages, but most of the data is unstructured text data. The unstructured characteristics of Web data force the Web content mining towards a more complicated approach.

The Web content mining is differentiated from two different points of view: Information Retrieval View and Database View [6] summarized the research works done for unstructured data and semi-structured data from information retrieval view. It shows that most of the researches use bag of words, which is based on the statistics about single words in isolation, to represent unstructured text and take single word found in the training corpus as features. For the semi-structured data, all the works utilize the HTML structures inside the documents and some utilized the hyperlink structure between the documents for document representation. As for the database view, in order to have the better information management and querying on the Web, the mining always tries to infer the structure of the Web site of to transform a Web site to become a database.

Multimedia data mining is part of the content mining, which is engaged to mine the high-level information and knowledge from large online multimedia sources. Multimedia data mining on the Web has gained many researchers' attention recently. Working towards a unifying framework for representation, problem solving, and learning from multimedia is really a challenge, this research area is still in its infancy indeed, many works are waiting to be done. For the details about multimedia mining, please refer [8] to find the related resource information.

Web Usage Mining

Web usage mining tries to discover the useful information from the secondary data derived from the interactions of the users while surfing on the Web. It focuses on the techniques that could predict user behavior while the user interacts with Web. In [5] abstract the potential strategic aims in each domain into mining goal as: prediction of the user's behavior within the site, comparison between expected and actual Web site usage, adjustment of the Web site to the interests of its users. There are no definite distinctions between the Web usage mining and other two categories. In the process of data preparation of Web usage mining, the Web content and Web site topology will be used as the information sources which interacts Web usage mining with the Web content mining and Web structure mining. Moreover, the clustering in the process of pattern discovery is a bridge to Web content and structure mining from usage mining.

The pre-fetching engine can be implemented in any of the elements that receive the predictions results. The objects pre-fetched are stored in a cache waiting to be demanded. The limitations on the available user's bandwidth constrained the benefits of pre-fetching in the past because pre-fetching can increase network traffic if its predictions are not accurate enough. These facts together with the difficulty in implementing these techniques, without modifying the massively used protocols have left a gap between academic results and available products. But the current user's bandwidth opens again new possibilities for pre-fetching to improve web performance with a reasonable cost.

III. Web Log and Data Pre-processing

The web log is a registry of web pages accessed by different users at different times, which can be maintained at the server-side, client-side or at a proxy server, each having its own benefits and drawbacks on finding the users' relevant patterns and navigational sessions. A web log stored at the client-side captures only web accesses by that particular client/user and could be beneficial in mining access sequences for a particular user as part of a web personalization system. This requires that a remote agent be implemented or a modified used to collect single-user data, thus eliminating caching and session identification problems. Whereas a proxy-side web log captures access sequences of the clients of a certain service provider company and could be used in applications like page pre-fetching [Pitkow and Pirolli, 1999] and caching to enhance the performance of the proxy server.

A proxy server can also reveal the actual HTTP requests from multiple clients to multiple web servers, thus, characterizing the browsing behaviour of a group of anonymous users sharing a common server [Shrivastava [8] et al., 2000], that can serve the current trend of Group Recommendation Systems. Web access sequences stored on the server-side represent webpage accesses of all users who visit this server at all times, which is good for mining multiple users' behaviour and for web recommender systems. Log while server logs may not be entirely reliable, due to caching as cached page views are not recorded in a server log. But this problem is solved by referring to logged information to infer and reconstruct user paths, filling out missing pages. The knowledge of user's history of navigation within a period of time is referred to as a *session*. These sessions, which provide the source of data for training, are extracted from the logs of the Web servers, and they contain sequences of pages that users have visited along with the visit date and duration

Work in this thesis focuses on server-side web logs used by e-commerce web sites. Such

Web logs are usually raw registries of URLs visited by different users.

1	2006-02-01 00:08:43 1.2.3.4 - GET /classes/cs589/papers.html - 200 9221 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://dataminingresources.blogspot.com/
2	2006-02-01 00:08:46 1.2.3.4 - GET /classes/cs589/papers/cms-tai.pdf - 200 4096 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1;+.NET+CLR+2.0.50727) http://maya.cs.depaul.edu/~classes/cs589/papers.html
3	2006-02-01 08:01:28 2.3.4.5 - GET /classes/ds575/papers/hyperlink.pdf - 200 318814 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1) http://www.google.com/search?hl=en&lr=&q=hyperlink+analysis+for+the+web+survey
4	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/announce.html - 200 3794 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/
5	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/styles2.css - 200 1636 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html
6	2006-02-02 19:34:45 3.4.5.6 - GET /classes/cs480/header.gif - 200 6027 HTTP/1.1 maya.cs.depaul.edu Mozilla/4.0+(compatible;+MSIE+6.0;+Windows+NT+5.1;+SV1) http://maya.cs.depaul.edu/~classes/cs480/announce.html

One can notice that each line contains the following data:

1) Timestamp of the HTTP request
2) IP address of the requesting client,
3) HTTP request command with requested URL
4) HTTP protocol version and client browser details.

Clearly, there is no representation whatsoever of domain knowledge or any means of describing the requested products. Before starting any mining technique, web data has to be cleaned and pre-processed. Pre-processing prepares data for the pattern discovery stage. It transforms Web log files into Web transaction data that can be processed by data mining tasks. Web data could take many forms. The primary data sources are the server log files that include Web server access logs and application server logs. Also, additional data sources may include operational databases, domain knowledge, site files and meta-data. This additional data can be available from client-side or proxy level data collection as well as from external click stream or demographic data sources.
Client side collection methods can get very complex like using a remote agent, such as JavaScript's or Java applets, or by modifying the source code of the browser. Both of these methods require user cooperation. With usage pre-processing, the data usually needs to be transformed and aggregated at different levels of abstraction. The most basic level of data abstraction is page view which represents a collection of Web objects displayed as a result of a single user action. A collection of page views for a single user during a single visit forms a session. Sessions may be used to analyze the user's behavioural browsing patterns. The second type of pre-processing is content pre-processing. It involves preparing text and multimedia files using classification and clustering techniques. Static Web pages can be easily pre-processed by parsing the HTML and reformatting the information or by running additional algorithms.
However, dynamic web pages that are the result of database accesses or personalization algorithms are usually more difficult to pre-process. Also, limiting pre-processing to certain pages that are generated by a combination of database accesses will not give definitive results. The third type of pre-processing is structure pre-processing. It consists of pre-processing the inter-page structure information or the Hyperlinks that connect one page to another. Again, pages that have a predefined structure are easily pre-processed. However, dynamically structured pages can be more difficult. Dynamic structure creates problems since a different site structure may have to be constructed for each server session.

IV. Techniques of Web Mining

Pattern Discovery: Pattern discovery involves the employment of sophisticated techniques from artificial intelligence, data mining techniques, psychology and information theory in order to extract knowledge from collected and pre processed data. Some of the most widely used pattern discovery approaches are statistical analysis, association rule mining, clustering, classification, sequential patterns and dependency modelling Shrivastava [8] et al. (2000). Statistical analysis techniques are the most common tools used to extract knowledge about Web site users. These tools could provide user information like the most frequently accessed pages, average time of viewing a certain page, average time the user spends browsing a certain site etc... This type of knowledge can never have 100% accuracy as in most cases it is based on incomplete log reports. However, knowledge extracted using statistical analysis could be very useful for improving the system performance and for providing support for marketing purposes especially for e-commerce applications Cooley et al.(1999). Association rule mining refers to the sets of pages that are accessed together in a single server session [2]. They are used to identify items that are likely to be purchased or viewed in a similar session. These rules are very helpful for marketing purposes. They also help Web designers improve their hyperlinks and reduce user latency when downloading a page Shrivastava [8] et al. (2000). Clustering aims at identifying a finite set of categories to describe a data set. It groups together data items with same characteristics.

One type of clustering is a usage cluster which involves finding users with same browsing habits. It is useful in providing personalized Web content to users. Another type of clustering is a page cluster which discovers pages that have related content. This kind of clustering is very useful for Internet search engines. Classification aims at finding common properties among a set of objects and mapping those objects into a set of predefined classes Cooley et al. An example is the classification of the clients of an insurance company according to the probability of submitting a claim. Clients classified in the higher risk classes have to pay higher premiums.
Sequential patterns are another type of pattern discovery. A sequence is defined as an ordered list of item and sets. Sequential pattern discovery attempts to find patterns such that a set of items is followed by another item within a certain period of time and in a certain server session. This can be useful in predicting users' future browsing patterns. Dependency modelling consists of techniques that are aimed at finding a model describing dependencies between variables in the Web domain. This is potentially useful for predicting future Web resources consumption. For example, it could help develop strategies to increase the sales of products offered by Website Shrivastava [8] et al. (2000).

Pattern Analysis:
Not all discovered patterns are useful and this step aims at identifying the patterns which represent new and potentially useful knowledge. Pattern analysis involves filtering out the unneeded patterns or rules discovered through the pattern discovery phase. The most common pattern analysis technique is the use of query language like SQL. Another technique could be the usage of online analytical processing (OLAP) tools Shrivastava [8] et al. (2000).

V. Web Usage Mining Techniques and Web Page Prediction

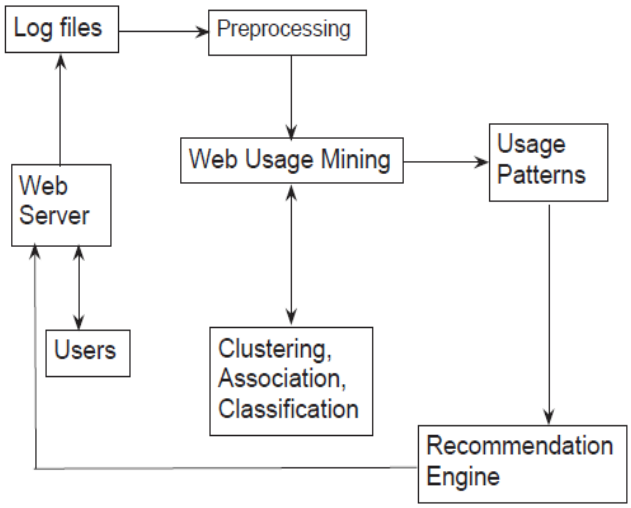

Fig.2 Web usage mining architecture.

Personalizing the Web users' content and recommending appropriate Web page simply that we are able to supply users with what they require based on their previous interactions within the same Web site. This task is viewed as a prediction task for we are trying to predict the users' level of interest in specific pages and rank these pages according to their predicted values shown n Fig.2.

Prediction Techniques

The prediction process forms part of the automatic personalization process that consists of a data collection phase and a learning phase. Data collection phase can take many forms and it is beyond the scope of this dissertation. The learning phase can be classified into memory based learning or model based learning depending on whether the learning is performed online while prediction takes place or offline using training data. Standard user-based and content based personalization systems rely on the memory based approach while item or page based personalization systems rely on the model based approach. Memory based systems simply memorize all the data and generalize from it online real time. Their computational complexity is O (MN) in the worst case where M refers to the number of users and N refers to the number of items. Using memory based systems involves scanning all users to find similar users and then scanning all the items that the similar users have selected. This online computation complexity becomes a problem with typical electronic commerce Web sites. They are, however, extensively used in research and practice. Model based systems, on the other hand, can have heavier computation than user based systems, but their heavy computation is performed offline and their online computations are light. Model based techniques, including those used in the pattern discovery phase of Web usage mining, use at two stage process for prediction. During the first phase, the data collected is mined offline and a model is generated.

During the second phase, prediction takes place online as a new site visitor begins interacting with the Web site. The new visitor session is scored based on the model constructed in the first phase. Model based systems computational complexity could be O (N2M)in the worst case because they first scan the items, then for each item, they scan all the users, and finally, they find similar items by scanning the items again. Their online computational complexity is O (N) in the worst case, but on average, the complexity is O (constant) because the online computation depends on the number of items to look up, not on the total number of users or items. This reduction in computational complexity makes model based systems more suited for the online prediction stage than memory based systems. On the other hand, memory based systems are better at adopting to changes in the data sources. In the case of new data, model based systems have to be either incremented or rebuilt Mobasher et al. (2000).

More reasons why model based systems outperform user based systems for predicting Web pages is that model based are item based and computations are based on the items that are usually easily accessed from a Web server log file. Also, item data is more static than user data that changes with users' circumstances and environment. Since this dissertation focuses on Web personalization where the data source is a repository of Web pages linked together according to some structure in a particular Web site, the prediction process is based on model based systems. In this chapter, we briefly describe a number of data mining algorithms used for offline model building techniques including Markov models, association rules and clustering.

VI. Conclusions

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 user 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. Clustering is one of the best solutions for resolving

the problem of worse prediction accuracy of Markov model. It is a powerful method for arranging users' session into clusters according to their similarity.

References

1. Adami, G., Avesani, P. & Sona, D. (2003), 'Clustering documents in a web directory', WIDM'03, USA pp. 66–73.
2. Agrawal, R., Imielinski, T. & Swami, A. (1993), 'Mining association rules between sets of items in large databases', ACM SIGMOD Conference on Management of data pp. 207–216.
3. 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.
4. The Next Page Access Prediction Using Makov Model "International Journal of Electronics Communication and Computer Technology (IJECCT) Volume 1 Issue 1 | September 2011 ISSN: 2249-7838
5. Silky Makker and R K Rathy. Article:" Web Server Performance Optimization using Prediction Prefetching Engine." International Journal of Computer Applications 23(9):19–24, June 2011.
6. A. Loizou and S. Dasmahapatra, "Recommender Systems for the Semantic Web," in ECAI 2006 Recommender Systems Workshop Trento, Italy, 2006
7. P Ravi Kumar, and Singh Ashutosh kumar, Web Structure Mining Exploring Hyperlinks and Algorithms for Information Retrieval, American Journal of applied sciences, 7 (6) 840-845 2010.
8. J. Shrivastava, R. Cooley, Mukund Deshpande and Pang-Ning Tan, "Web Usage Mining:Discovery and Applications of Usage Patterns form Web Data," *SIGKDD Explorations*, Vol. 1, Issue 2, pp.12-23, Jan 2000.
9. A. Harth, M. Janik, and S. Staab, "Semantic Web Architecture," in Handbook of Semantic Web Technologies, J. Domingue, D. Fensel, and J. A. Hendler, Eds.: Springer-Verlag Berlin Heidelberg, 2011,
10. Robert Cooley, Pang-Ning Tan, and Jaideep Srivastava. Websift: The web site information filter system. In *Proceedings of the Web Usage Analysis and User Profiling Workshop*, 1999
11. Mobasher, B., Dai, H., Luo, T. & Nakagawa, M. (2000), Discovery of Aggregate Usage Prowls for Web Personalization, *in* `Web KDD Workshop 2000', USA, pp. 61-82.
12. Thi Thanh Sang Nguyen, Hai Yan Lu, Web Page Recommendation based on Web Usage & Domain Knowledge, IEEE Transaction,2013

