International Journal of Enhanced Research in Management & Computer Applications
ISSN: 2319-7471, Vol. 4 Issue 8, August-2015

Review Paper on Webpage Prediction Based on Mining Techniques

Ankita Gulati[1], Lokesh Kumar[2]

[1]Dept. of Computer Science Rohtak Institute of Engineering and Management, Rohtak, Haryana, India
[2]Assistant Professor & HOD, Dept. of Computer Science & Engineering, Rohtak Institute of Engineering & Management, Rohtak, Haryana, India

ABSTRACT

The wide adoption of Internet has fundamentally altered the ways in which we communicate, gather information, conduct businesses and make purchases. As the use of the World Wide Web, computer scientists and physicists rushed to characterize this new phenomenon. The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized in such a way that they can be accessed by different users efficiently. The objective of this review paper is to study the various data mining techniques to improve the efficiency in predicting the next page. The Markov Model is studied to determine the probability of next page access and that pages are prefetched in the cache.

Keywords: Web Mining, Markov Model, Data Mining Techniques.

1. INTRODUCTION

Web page access prediction gained its importance from the ever increasing number of e-commerce Web information systems and e-businesses. Web page prediction, that involves personalizing the Web users' browsing experiences, assists Web masters in the improvement of the Website structure and helps Web users in navigating the site and accessing the information they need. The most widely used approach for this purpose is the pattern discovery process of Web usage mining that entails many techniques like Markov model, association rules and clustering. Implementing pattern [1] discovery techniques as such helps predict the next page to be accessed by the Web user based on the user's previous browsing patterns. However, each of the aforementioned techniques has its own limitations, especially when it comes to accuracy and space complexity.

We can integrate low -order Markov model and clustering. The data sets are clustered and Markov model analysis is performed on each cluster instead of the whole data sets. The outcome of the integration is better accuracy than the combination with less state space complexity than higher order Markov model.

II. LITRATURE REVIEW

It was Etzioni [3] who first coined the term web mining. Etzioni starts by making a hypothesis that the information on the web is sufficiently structured and outlines the subtask of web mining. Cooley et al; Srivastava et al. [4] define Web usage mining as a three-phase process, consisting of preprocessing, pattern discovery, and pattern analysis. Their prototype system, Web SIFT, first performs intelligent cleansing and preprocessing for identifying users, server sessions, and inferring cached page references through the use of the referrer field, and also performs content and structure preprocessing [5]. Wang et al.in [6] surveys the caching studies taking into account many issues such as caching architectures, replacement policies, cache routing, dynamic caching, fault tolerance, security, etc. Padmanabhan [2] use dependency graph for prediction and prefetching. Their prediction algorithm construct a dependency graph that depicts the pattern of accesses to different file stored at the server.

M. Eirinaki [7] use Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between groups of users with specific interest. MOBASHER [8] showed the evidence that web request follow a Zipf-like distribution. He first investigates the page request distribution seen by web proxy cache using traces from a variety of sources. Lei Shi [9] presents the Web object popularity based

model on zipf–like law, introduces the stability concept of the web system, and calculate the upper bound for the minimum length of the request stream in order to get stability. It was Etzioni [3] who first coined the term web mining .Etzioni starts by making a hypothesis that the information on the web is sufficiently structured and outlines the subtask of web mining. Cooley et al; Srivastava et al. [4] define Web usage mining as a three-phase process, consisting of preprocessing, pattern discovery, and pattern analysis. Their prototype system, Web SIFT, first performs intelligent cleansing and preprocessing for identifying users, server sessions, and inferring cached page references through the use of the referrer field, and also performs content and structure preprocessing [5]. Wang et al. in [6] surveys the caching studies taking into account many issues such as caching architectures, replacement policies, cache routing, dynamic caching, fault tolerance, security, etc. Padmanabhan [2] use dependency graph for prediction and prefetching. Their prediction algorithm construct a dependency graph that depicts the pattern of accesses to different file stored at the server. M. Eirinaki [7] use Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between groups of users with specific interest. MOBASHER [8] showed the evidence that web request follow a Zipf-like distribution. He first investigates the page request distribution seen by web proxy cache using traces from a variety of sources. Lei Shi [9] presents the Web object popularity based model on zipf–like law, introduces the stability concept of the web system, and calculate the upper bound for the minimum length of the request stream in order to get stability.

III. WEB MINING COMPONENT AND METHODOLOGY

The web is a vast collection of completely uncontrolled heterogeneous documents. Thus, it is huge, diverse, and dynamic, and raises the issues of scalability, heterogeneity, and dynamism, respectively. Web mining can be broadly defined as the discovery and analysis of useful information from the WWW. In web mining data can be collected at the server side, client side, proxy servers, or obtained from an organization's database. Depending on the location of the source, the type of collected data differs [10].It also has extreme variation both in its content (e.g., text, image, audio, symbolic) and meta information that might be available. This makes the techniques to be used for a particular task in web mining widely varying. Some of the characteristics of web data are : 1) Unlabeled; 2) Distributed; 3) Heterogeneous (mixed media); 4) Semi structured; 5) Time varying; 6) High dimensional

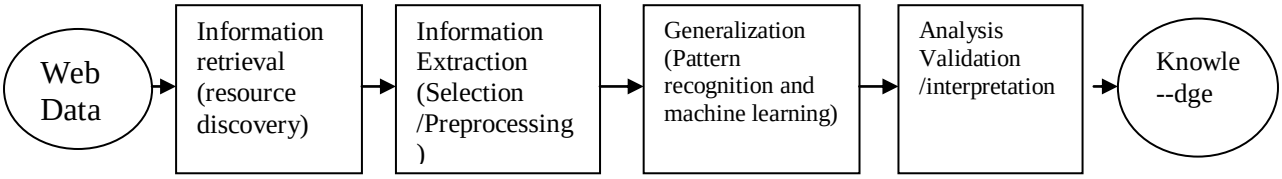

Figure. 1: Web mining subtask

Information Retrieval (IR) (Resource Discovery): Resource discovery or IR deals with automatic retrieval of all relevant documents, while at the same time ensuring that the non relevant ones are fetched as few as possible. The IR process mainly includes document representation, indexing, and searching for documents.

Information Selection/Extraction and Preprocessing: Once the documents have been retrieved the challenge is to automatically extract knowledge and other required information without human interaction. Information extraction (IE) is the task of identifying specific fragments of a single document that constitute its core semantic content.

Generalization: In this phase, pattern recognition and machine learning techniques are usually used on the extracted information. Most of the machine learning systems, deployed on the web, learn more about the user's interest than the web itself. A major obstacle when learning about the web is the labeling problem: data is abundant on the web but it is unlabeled. Many data mining techniques require inputs labeled as positive (yes) or negative (no) examples with respect to some concept.

Analysis: Analysis is a data-driven problem which presumes that there is sufficient data available so that potentially useful information can be extracted and analyzed. Humans play an important role in the information or knowledge discovery process on the web since the web is an interactive medium. This is especially important for validation and or interpretation of the mined patterns which take place in this phase.

WEB MINING CATEGORIES: Web mining can be categorized in to three area of interest based on which part of the web to mine:

1. **Web content mining** describes the discovery of useful information from the web contents/data /documents. However, what consist of web content could encompass a very broad range of data. Previously the internet consists of different type of services and the data sources such as Gopher, FTP and Usenet. Now most of those

data are either ported to or accessible from the web. Basically the web content consists of several types of data such as textual image, audio, video, meta data as well as hyperlinks.

2. **Web structure mining** tries to discover the model underlying the link structure of the web. The model is based on the topology of hyperlinks with or with out the description of the link. This model can be used to categories web pages and is useful to generate information such as similarity and relationship between different web sites.
3. **Web usage mining** tries to make sense of the data generated by the web surfer's sessions or behaviors. While the web content and structure mining utilize the real or primary data on the web .web usage mining mines the secondary data derived from the interaction of the users while interacting with web. The web usage data include the data from the web server logs, proxy server logs, browser logs, user profiles, registration data, user session or transactions, cookies, user queries and any other data as the result of interaction.

Typically, the usage mining is defined as a three-phase process: data preprocessing, pattern discovery, and pattern analysis. Figure 4.2 demonstrates such architecture. In this section, we present an overview of the detailed process.

- **Data Preprocessing**: retrieves raw data from the Web resources, and automatically selects and preprocesses the retrieved data. It includes any kind of transformation of the original raw data.

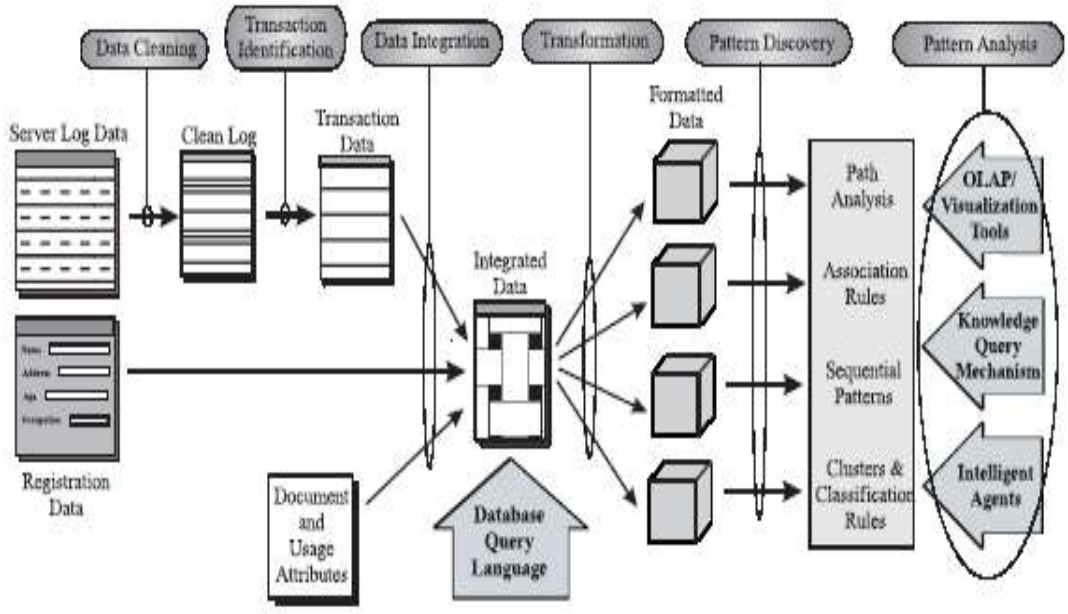

Figure. 2: Data Preprocessing

- **Pattern Discovery**: discovers knowledge from the pre-processed data. Machine learning and data mining procedures are carried out at this stage.
- **Pattern Analysis and Applications**: validates and post-processes the discovered patterns.

IV. WEB MINING PROS AND CONS

PROS

Web mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. This technology has enabled ecommerce to do personalized marketing, which eventually results in higher trade volumes. The government agencies are using this technology to classify threats and fight against terrorism. The predicting capability of the mining application can benefits the society by identifying criminal activities. The companies can establish better customer relationship by giving them exactly what they need. Companies can understand the needs of the customer better and they can react to customer needs faster. The companies can find, attract and retain customers; they can save on production costs by utilizing the acquired insight of customer requirements. They can increase profitability by target pricing based on the profiles created. They can even find the customer who might default to a competitor the company will try to retain the customer by providing promotional offers to the specific customer, thus reducing the risk of losing a customer or customers.

CONS

Web mining, itself, doesn't create issues, but this technology when used on data of personal nature might cause concerns. The most criticized ethical issue involving web mining is the invasion of privacy. Privacy is considered lost when information concerning an individual is obtained, used, or disseminated, especially if this occurs without their knowledge or consent. The obtained data will be analyzed, and clustered to form profiles; the data will be made anonymous before clustering so that there are no personal profiles. Thus these applications de-individualize the users by judging them by their mouse clicks. De-individualization, can be defined as a tendency of judging and treating people on the basis of group characteristics instead of on their own individual characteristics and merits.

Another important concern is that the companies collecting the data for a specific purpose might use the data for a totally different purpose, and this essentially violates the user's interests. The growing trend of Selling personal data as a commodity encourages website owners to trade personal data obtained from their site. This trend has increased the amount of data being captured and traded increasing the likeliness of one's privacy being invaded. The companies which buy the data are obliged make it anonymous and these companies are considered authors of any specific release of mining patterns. They are legally responsible for the contents of the release; any inaccuracies in the release will result in serious lawsuits, but there is no law preventing them from trading the data. Some mining algorithms might use controversial attributes like sex, race, religion, or sexual orientation to categorize individuals. These practices might be against the anti-discrimination legislation. The applications make it hard to identify the use of such controversial attributes, and there is no strong rule against the usage of such algorithms with such attributes. This process could result in denial of service or a privilege to an individual based on his race, religion or sexual orientation, right now this situation can be avoided by the high ethical standards maintained by the data mining company. The collected data is being made anonymous so that, the obtained data and the obtained patterns cannot be traced back to an individual. It might look as if this poses no threat to one's privacy, actually many extra information can be inferred by the application by combining two separate unscrupulous data from the user.

V. WEB PAGE PREDICTION TECHNIQUE

After dividing user sessions into a number of clusters, Markov model [11] analyses are carried out on each of the clusters. Markov models are used in the identification of the next page to be accessed by the Web site user based on the sequence of previously accessed pages. Let $P = \{p1, p2, \ldots\ldots\ldots, pm\}$ be a set of pages in a Web site. Let W be a user session including a sequence of pages visited by the user in a visit. Assuming that the user has visited l pages, then prob $(p_i \mid W)$ is the probability that the user visits pages p_i next. Page P_{l+1} the user will visit next is estimated by:

$$P_{l+1} = \text{argmax}_{p \in IP}\{P(P_{l+1}=p|W)\}$$
$$P_{l+1} = \text{argmax}_{p \in IP}\{P(P_{l+1}=p|p_l,p_{l-1},\ldots.p_1)\} \quad (1)$$

This probability, prob $(p_i \mid W)$ is estimated by using all sequences of all users in history (or training data), denoted by W. Naturally, the longer l and the larger W, the more accurate prob $(p_i \mid W)$. However, it is infeasible to have very long l and large W and it leads to unnecessary complexity. Therefore, a more feasible probability is estimated by assuming that the sequence of the Web pages visited by users follows a Markov process that imposes a limit on the number of previously accessed pages k. In other words, the probability of visiting a page *pi* does not depend on all the pages in the Web session, but only on a small set of k preceding pages,

where $k << l$.

The equation become

$$= \text{argmax}_{p \in IP}\{P(P_{l+1}=p|p_l,p_{l-1},\ldots.p_{l-(k-1)})\} \quad (2)$$

Where k denotes the number of the preceding pages and it identifies the order of the Markov model. The resulting model of this equation is called the all *kth* order Markov model. Of course, the Markov model starts calculating the highest probability of the last page visited because during a Web session, the user can only link the page he is currently visiting to the next one.

Let S^k_j be a state containing k pages, $S^k_j = (p_{l-(k-1)}, p_{l-(k-2\_}, \ldots. p_l)$ The probability of $P(p_i \mid S^k_j)$ is estimated as follows from a history (training) data set.

$$P(p_i \mid S^k_j) = \frac{\text{Frequency}(<S^k_j, p_i>)}{\text{Frequency}(S^k_j)} \quad (3)$$

The fundamental assumption of predictions based on Markov models is that the next state is dependent on the previous k states. The longer the k is, the more accurate the predictions are.

VI. CONCLUSIONS AND FUTURE SCOPE

Prefetching is characterized as one of the most efficient schemes to further reduce the user access latency, but runs the risk of increasing network traffic. Most of the approaches attempt to prefetch web objects according to some kind of criteria. The Web page access prediction accuracy can be improved by integrating various prediction models: Markov model, Clustering and association rules according to certain constraints. After that Zipf Estimator can be applied on the rule generated from the previous phase. Efficient prefetching is very important for prefetching the next page. Usage patterns discovered through Web usage mining are effective in capturing item-to-item and user-to-user relationships and similarities at the level of user sessions. However, without the benefit of deeper domain knowledge, such patterns provide little insight into the underlying reasons for which such items or users are grouped together. Furthermore, the inherent and increasing heterogeneity of the Web has required Web-based applications to more effectively integrate a variety of types of data across multiple channels and from different sources. Thus, a focus on techniques and architectures for more effective integration and mining of content, usage, and structure data from different sources is likely to lead to the next generation of more useful and more intelligent applications, and more sophisticated tools for Web usage mining that can derive intelligence from user transactions on the Web.

REFERENCES

[1]. Virgilio Augusto F. Almeida, Marcio Anthony G. Cesirio, Rodrigo Fonseca Canado, Wagner Meira Junior, and Cristina Duarte Murta, "Analyzing the behavior of a proxy server in the light of regional and cultural issues." http://www.anades.dcc.ufmg.br/paperSubmetidos/ web cache/cultural/, 1998.

[2]. Padmanbhan, "Using Predictive prefetching to Improve World wide web Latency", V.N, 1996 Comput. Comm. Rev, 26(3):22-36.

[3]. O. Etzioni, "The World Wide Web: Quagmire or gold mine". Communication of the ACM, 39(11): 65-68, 1996.

[4]. COOLEY, R., TAN, P-N., AND SRIVASTAVA, J. 1999b, "WebSIFT: The web site information filter system. In Proceedings of the Web Usage Analysis and User Profiling" Workshop (WEBKDD'99), Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (Boston, August).

[5]. Carlos Cunha, Azer Bestavros, and Mark Crovella, "Characteristics of WWW client-based traces.", Technical Report TR-95-010, Boston University, Computer Science Dept., Boston, MA 02215, USA, April 1995.

[6]. J. Wang, "A survey of web caching schemes for the internet," ACM Computer Communication Review, vol. 29, no. 5, pp. 36– 46, 1999.

[7]. M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," ACM Trans. Inter. Tech., Vol. 3, No. 1, pp. 1-27, 2003.

[8]. COOLEY, R., MOBASHER, B., AND SRIVASTAVA, J. 1999a. " Data preparation for mining world wide web browsing patterns", Knowl. Inf. Syst., 1, 1 (Feb).

[9]. Lei Shi,Zhimin Gu, Lin Wei, and Yun Shi, "An applicative study of zipf's law on web cache" ,In International Jounal of information Technology, Vol. 12 No.4 2006.

[10]. Steven Glassman, " A caching relay for the World Wide Web", In First International Conference on the World Wide Web, CERN, Geneva, Switzerland, May 1994.

[11]. Faten Khalil, Jiuyong Li and Hua Wang "A Framework of Combining Markov Model with Association Rules for Predicting Web Page Accesses" ,Proc. Fifth Australasian Data Mining Conference (AusDM2006), CRPIT Volume 61,177-184.