Computer Engineering and Intelligent Systems
ISSN 2222-1719 (Paper) ISSN 2222-2863 (Online)
Vol.6, No.7, 2015

REVIEW PAPER ON WEB PAGE PREDICTION USING DATA MINING

Smriti Pandya
Mtech Scholar, Department Of CSE, RGPV, Bhopal, India

Mr. Rajesh Nigam
Department Of CSE, Bhopal, India

Abstract –
The continuous growth of the World Wide Web imposes the need of new methods of design and determines how to access a web page in the web usage mining by performing preprocessing of the data in a web page and development of on-line information services. The need for predicting the user's needs in order to improve the usability and user retention of a web site is more than evident now a day. Without proper guidance, a visitor often wanders aimlessly without visiting important pages, loses interest, and leaves the site sooner than expected. In proposed system focus on investigating efficient and effective sequential access pattern mining techniques for web usage data. The mined patterns are then used for matching and generating web links for online recommendations. A web page of interest application will be developed for evaluating the quality and effectiveness of the discovered knowledge.

Keyword: Webpage Prediction, Web Mining, MRF, ANN, KNN, GA.

1. Introduction

Web usage mining is the third category in web mining. This type of web mining allows for the collection of Web access information for Web pages. This usage data provides the paths leading to accessed Web pages. This information is often gathered automatically into access logs via the Web server. CGI scripts offer other useful information such as referrer logs, user subscription information and survey logs. This category is important to the overall use of data mining for companies and their internet/ intranet based applications and information access.

Web usage mining allows companies to produce productive information pertaining to the future of their business function ability. Some of this information can be derived from the collective information of lifetime user value, product cross marketing strategies and promotional campaign effectiveness. The usage data that is gathered provides the companies with the ability to produce results more effective to their businesses and increasing of sales. Web usage data can also be useful for developing marketing skills that will out-sell the competitors and promote the company's services or product on a higher level.

Researchers have recently suggested various schemes that apparently make the visitor interest in the site and guide him or her to important pages. Research has suggested personalization systems to alleviate this problem by causing a web site to be more responsive to the unique and individual needs of each user. Personalization involves the creation of user profiles and tailoring the information received by the user to his profile. Some personalization systems get the users; interests by asking them to complete a form or answer a questionnaire in an intrusive manner, whereas others use data that users generate during online transactions. Irrespective of how webmasters accomplish personalization, bamshad Mobasher and others have classified it into four major systems: knowledge-based systems, content based filtering, collaborative filtering and web usage mining.

Knowledge-based systems make use of knowledge about users and products to generate recommendations. They use a reasoning process to determine what products meet a user's requirements. Content based systems establish users' interest profiles by analyzing the features of their preferred items. They compare the features of recommendable items to those of the preferred items of the user. CF-based systems provide personalized recommendations according to user preference. They maintain data about target users purchasing habits or interests and use this data to identify groups of similar users. They then recommend items liked by others; similar users.CF tries to identify users that have relevant interests and preferences by calculating similarities among user profiles.web usage mining based system incorporate data mining techniques such as classification, clustering, association rule, sequential pattern to discover interesting pattern in web usage mining.

2. MOTIVATION

In the current era, we are witnessing a surge of web usage around the globe. A large volume of data is constantly being accessed and shared among a varied type of users; both humans and intelligent machines. Thus, taking up a structured approach to control this information exchange, has made Web Mining one of the hot

topics in the field of information technology. This reason motivated to take up this topic. Another consideration was the fact there has been a minimal amount of research work done on Web Usage Mining.

3. RELATED WORK

Preprocessing of web log file is first necessary and must step for web usage mining. Cleaned data after preprocessing is solid base for pattern mining and pattern analysis. Quality of pattern mining and pattern analysis is fully dependent on preprocessing step. Some of preprocessing techniques are applied but we can use less or even ignored preprocessing techniques to improve the quality of preprocessed data. Some new techniques can provide the user with the opportunity to analyze the log file at different level of abstraction such as user sessions. In order to gain better understanding of log file we need hierarchical clustering by using proposed clustering technique. The user or data mining expert can have more knowledge of log file since unbiased grouping of data exists. With this enhanced information, the web log user can be more focused and guided.

There are some benefits of these concept in which exploring preprocessing techniques and use them with the combination of existing techniques to make the whole process more robust and there are some advantages & disadvantages inside there in which Web log preprocessing improves the quality and efficiency WUM. It is the beneficial step of WUM such as pattern discovery and pattern analysis. Each step of preprocessing depends upon the other and after summarizing gives the result hence the process takes longer time. Large amount of "irrelevant information" in the web log, the original log file cannot be directly used in the web usage mining (WUM) procedure. Therefore the preprocessing of web log file becomes imperative [1].

It determines a novel approach for next page access by web user using Zipf estimator. It has been found that by using the Zipf estimator, the probability of accessing the next page has been computed efficiently. Depending upon the probability of the next page, the web pages are prefetched locally on the proxy server so as to reduce the retrieval latency. In this paper Prediction Perfecting Engine has been proposed to processes the past references to deduce the probability of future access for the documents accessed so far. There are some advantages & disadvantages inside there in which, it reduces latency and filter the unwanted data. Web performance is improved by caching; the benefit of using it is limited to filling the cache with documents without any prior knowledge [2].

Dr. A.K. Sharma et al [3] proposed a novel approach for predicting user behavior for improving web performance. In this prediction and prefetching is done both by collaborating information from user access log and website structure repository. This work overcomes the limitation of path completion. Application of Petri Nets for extracting web site structure helps in path completion process, better prediction, decreasing web latency and improving web performance.

There are some benefits of these concept in which all the preprocessing is done, the cleaner version of the log is formed called Data mart. Data mart acts as a database on which various data mining operations operate for generating the rules. In this, the combination of Clustering and association rule mining is proposed. The clustering of web user sessions is done so as to cluster user with similar behavior together. Further association rules mining is applied on clustered sets. Association Rule mining are a major pattern discovery technique. There are some advantages & disadvantages inside there in which, log file helps in analyzing user access Patterns and in predicting next page likely to be accessed by the web user. Using association rules for web access prediction involves dealing with too many rules and it is not easy to find a suitable subset of rules to make accurate and reliable predictions [3].

The web is a most important medium to conduct business and commerce. Therefore the design of web pages is very important for the system administrator and web designers. These features have great impact on the number of visitors. So the web analyzer has to analyze with the data of server log file for detecting pattern. In this paper they tried to give a clear understanding of the data preparation process and pattern discovery process. Web usage patterns and data mining can be the basis for a great deal of future research. More research needs to be done in Ecommerce, Bioinformatics, Computer Security, Web Intelligence, Intelligent Learning, Database Systems, Finance, Marketing, Healthcare and Telecommunications by using Web usage mining.

There are some benefits of these concept in which offers some techniques such as statistical analysis, association rules, sequential pattern analysis, clustering and so on. It determines the visitor's location converting the IP address into its domain name. Using the path analysis technique, the information offers a valuable imminent of user navigation problems. To analyze the path the administrator can understand what pages the visitors like most or how long path they like to visit in a web site. For e.g. If 65% of visitors who accessed/sustTube/video.php by starting at / sustTube and proceeding through /sustTube/view_video.php, or /sustTube/video.php, decided to make a decision after seeing the sample video. From server log file's user agent portion we get the browsers name and the number of users uses a particular browser. So the system can decide from which browsers most number of visitors hit the site. It's also suitable to determine the operating system of the visitors there are some advantages & disadvantages-use of preprocessing and pattern discover has been used for the preprocessing of

data and mining hidden pattern for the data. It provides a comprehensive idea about the pattern discovery of web usage mining [4].

Because of the huge quantity of data of web pages on many portal sites, for convenience, are to assemble the web page based on category. In this paper user's browsing behavior has been observed. One is category stage and the other is web page stage. In stage one is to predict category. The unnecessary categories can be excluded. The scope of calculation is massively reduced.

There are some benefits of these concept in which Web usage mining focuses on discovering the potential knowledge from the browsing patterns of users and to find the correlation between the pages. With exponential growth of web log, the conventional data mining techniques were proved to be inefficient. As web log is incremental in nature, it becomes a crucial issue to predict exactly the ways how users browse websites. It is necessary for web miners to use predictive mining techniques to extract the usage patterns and study the visiting characteristics of user. The data on the web log is heterogeneous and non scalable, hence to reduce the operation scope an increase the accuracy precision significantly an improved hybrid model is required. There are some advantages & disadvantages inside there in which, it introduces an efficient hybrid predictive model, which is a combination of Markov model and Bayesian theorem. This two stage predictive model to enables the web miner to identify and analyze web user navigation patterns. Web log is incremental in nature; it becomes a crucial issue to predict exactly the ways how users browse websites. This model is used to enable the identification of user navigation patterns and also used to foresee the next link choice of a user [5].

When a client requests for a web page before accessing the web page a prediction is made for accessing that web page. All the web objects are brought from server to the client. The access to the web objects are on the basis of the data prefetched from the server. The data of three web logs of servers are tested on both existing algorithm and the proposed model. In the real environment, the results showed that our proposed mechanism performed better than the existing algorithm for web page prediction. Major advantage of the Sequential Rank Selection algorithm is that It selects only one web page of a website for prefetching purposes of user; hence consumed much less memory space of users and utilizes much less bandwidth of the network. The proposed architecture reduced the user's latency due to the efficient prediction of web pages by the Sequential Rank Based Selection algorithm from the cluster. There are some advantages & disadvantages inside there research focused on when a user requests for a web page, how to improve the overall performance of web perfecting mechanism. The proposed mechanism provides the pages locally available to a user or group of users by utilizing bandwidth of the network. The server contains an algorithm for the prediction of web pages and the prediction of a webpage is based on counting the number of times a page is accessed by a user from each cluster.

There are some advantages & disadvantages inside there in which, it improves the performance of the Web Caching techniques due to prediction of the user pages in advance before the user requests. Web caching is limited due to its size. The prefetching of data is done which is very efficient method and also reduces the user latency [6].

Analysis is performed directly on the log files and no separate data warehouse is required. Web log analyzer has identified several Web usage access pattern by applying well known data mining techniques to the access logs. This includes descriptive statistic and Association Rules including support and confidence to represent the Web usage and user behavior. The size NITR log file is 57.7MB which consists of more than one lakh entries of particular duration. Other analysis about trends of visitor's access patterns, bandwidth consumption, referrals, user agents and different types of errors that occurred in web surfing. There are some benefits of these concept in which it discovers useful patterns from the web server log file of an academic institute and results can be used in different applications like web traffic analysis, efficient website administration, site modifications, system improvement and personalization and business intelligence etc.

There are some advantages & disadvantages inside there in which describe the common web log content which need to be analyzed. It gives idea about different Web log it provides the overview of Web usage pattern analysis process and gives various applications of analyzed patterns [7].

We analyzed the all-Kth Markov model and formulated its general accuracy and PR. Moreover, we proposed and presented the modified Markov model to reduce the complexity of original Markov model. The modified Markov model successfully reduces the size of the Markov model while achieving comparable prediction accuracy. Additionally, we proposed and presented a two-tier prediction framework in Web prediction. We showed that our two-tier framework contributed to preserving accuracy (although one classifier was consulted) and reducing prediction time. We conducted extensive set of experiments using different prediction models, namely, Markov, ARM, all-Kth Markov, all-Kth ARM, and a combination of them. We performed our experiments using three different data sets, namely, NASA, UOFS, and UAEU, with various parameters such as rank, partition percentage, and the maximum number of N-grams. There are some benefits of these concept in which framework can improve the prediction time without compromising prediction accuracy. We have used standard benchmark data sets to analyze, compare, and demonstrate the effectiveness of our techniques using variations of Markov models and association rule mining. Our experiments show the effectiveness of our

modified Markov model in reducing the number of paths without compromising accuracy. Additionally, the results support our analysis conclusions that accuracy improves with higher orders of all-Kth Model.
There are some advantages & disadvantages inside there in which smaller N-gram models perform better than higher N-gram models in terms of accuracy. This is because of the small number of experiences/sessions obtained during data processing of large N-grams. We have also applied ranking to improve the prediction accuracy and to enhance its applicability. Increases the rank improves prediction accuracy and individual higher ranks have contributed Less to the prediction accuracy [8].
One of the most important internet challenges in coming years will be the introduction of intelligent services and a more personalized environment for user. In this paper web page prediction is presented. We use several classification techniques, namely, Support Vector Machines (SVM), Association Rule mining (ARM), and Markov model in WWW prediction. We proposed a hybrid model by combining two or more of them using Dempster's rule to enhance the efficiency of prediction [9].
This paper presents our experimental work on applying K-means, heuristic K-means and fuzzy C-means algorithms for clustering text documents. We have experimented with different representations (tf, tf.idf & Boolean) and different feature selection schemes (with or without stop word removal & with or without stemming). We ran our implementations on some standard datasets and computed various performance measures for these algorithms. The results indicate that tf.idf representation, and use of stemming obtains better clustering. Moreover, fuzzy clustering produces better results than both K-means and heuristic K-means on almost all datasets, and is a more stable method. Document clustering refers to unsupervised classification (categorization) of documents into groups (clusters) in such a way that the documents in a cluster are similar, whereas documents in different clusters are dissimilar. The documents may be web pages, blog posts, news articles, or other text files [10].
Here author introduce the use of default rule in resolving web access ambiguous predictions. This method could provide better prediction than using the individual traditional models. The results have shown that the default rule increases the accuracy and model-accuracy of web page access predictions. It also applies to association rules and the other combined models. Mining user patterns of web log files can provide significant and useful informative knowledge. A large amount of research has been done in trying to predict correctly the pages a user will most likely request next. Markov models are the most commonly used approaches for this type of web access prediction. Web page prediction requires the development of models that can predict a user's next access to a web server. Many researchers have proposed a novel approach that integrates Markov models, association rules and clustering in web site access predictability. The low order Markov models provide higher coverage, but these are couched in ambiguous rules [11].
Ming Syan Chen et al. introduced the notion of "maximal forward reference (MFR)" to identify users' transactions and employed data mining techniques (such as association rules discovery) to mine frequently-accessed paths and make predictions. They first converted the original log data sequence into a set of maximal forward references and eliminated the effect of some backward references, and then they presented algorithms to recognize the frequent traversal patterns from the maximal forward references obtained, which can be used to predict the user's future requests [12].
T. I. Ibrahim et al introduced a neural networks model to implement the semantics-based Web page prediction. This model extracts the semantics of a Web page according to the keywords offers URL anchor text. It employs these keywords as the input of the neural network to construct the semantic network of URLs, and predicts user's future requests based on the output of the neural network. In order to reduce the influence of the ambiguity of key words, this model builds a predictor for every different category of Web pages, which enhances the prediction accuracy but also decreases the applicability of this mode [13].
M. Eirinaki et al. proposed a novel Web personalization approach: Usage-based Page Rank (UPR), which combines both Web usage information and Web link structure information to conduct Web page ranking and prediction. This approach employs UPR to rank the Web pages in a relevant personalized navigational graph and predicts the probable pages in terms of their ranking values [14].
Yong Zhen Guo et al extended the UPR approach by introducing the access time duration of each Web page as another biasing factor, which will yield more accurate prediction [15].
Schechter constructed an access path tree for the current user and used the longest-match method to find a history path which matched the user's current navigational path. In this way the user's following access requests can be predicted, but the construction of path trees and the match of history paths are expensive in terms of both computing and storage [16].
Sarukkai considered one state can transfer to another state with a certain probability according to previous users' access paths. After all transition probabilities are computed from training Web logs, the model can predict the most probable next page for the current user in terms of the transition probability matrix. However, when making predictions, this approach only takes users' current access requests into consideration but not the whole access paths, which will influence the prediction accuracy [17].

In order to deal with this problem, higher-order Markov models are proposed, which take into account more states when computing the transition probability, and thus improve the prediction accuracy. However, the increase of the order will increase the state space complexity [18].
M. Deshpande et al discussed the shortcomings of higher-order Markov models in predicting Web users' browsing behaviors, and presented three schemes to eliminate the state space complexity of higher-order Markov models without influencing the performance [19].
A Hidden Markov Model (HMM) is a dual-stochastic process which is very popular for labeling sequences; one stochastic process is an invisible Markov chain that describes the transition between states (labels) while the other reflects the statistical relationship between states and observations [20].
Xin Jin et al. proposed a HMM- based prefetching model in which they employed HMM to capture and mine the latent concepts of information requirement implied by Web users' access paths, and then used the obtained information to make semantic-based prefetching decisions [21].

4. PROBLEM STATEMENT

Web page prediction is the web usage mining by performing preprocessing of the data from a web site. The need for predicting the user's needs in order to improve the usability and user maintenance of a web site is more than marked now a day's lacking proper guidance, a visitor often wanders aimlessly without visiting significant pages, loses attention, and leaves the site earlier than expected; in previous work with MRF and KNN for web page prediction quality and effectiveness of the discovered knowledge not much improve.

5. PROPOSED METHODOLOGY

In proposed system focus on investigating competent and efficient sequential admittance model mining techniques for web usage data. The mined patterns are then used for matching and generating web links for online recommendations. GA with KNN can be used for web page prediction with greater speed and accurately.

6. CONCLUSION

Web usage mining allows the collection of web access information for Web pages prediction. This usage data provides the paths leading to accessed Web pages. Mainly focus on investigating efficient and effective sequential access pattern mining techniques for web usage data. The mined patterns are then used for matching and generating web links for online recommendations. The proposed system will than compared for good performance with high satisfaction and applicability with the existing systems. A web page of interest (web recommendation) application will be developed for evaluating the quality and effectiveness of the discovered knowledge.

REFERENCES

[1] Tasawar Hussain, Dr. Sohail Asghar, Dr. Nayyer Masood, "Web Usage Mining: A Survey on Preprocessing of Web Log File"

[2] Payal Gulati, "A Novel Approach for Determining Next Page Access", 2008, IEEE.

[3] Priyanka Makkar, Payal Gulati, Dr. A.K. Sharma. "A Novel Approach for Predicting User Behavior for Improving Web Performance", (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 04, 2010, 1233-1236

[4] Shahnaz Parvin Nina, Md. Mahamudur Rahaman, Md. Khairul Islam Bhuiyan, "Pattern Discovery of Web Usage Mining", Vol. 02, No. 04, 2010, 1233-1236

[5] V.V.R. Maheswara Rao, Dr. V. Valli Kumari, "An Efficient Hybrid Predictive Model to Analyze the Visiting Characteristics of Web User using Web Usage Mining", International Conference on Advances in Recent Technologies in Communication and Computing, 2010.

[6] Naveed Ahmad, Owais Malik, Mahmood ul Hassan, Muhammad Shuaib Qureshi, Asim Munir , "Reducing User Latency in Web Prefetching Using Integrated Techniques", 2011 IEEE conference

[7] Dilip Singh Sisodia, Shrish Verma, "Web Usage Pattern Analysis through Web Logs: A Review" Ninth international joint conference on computer science and software engineering (JCSSE) – 2012.

[8] Mamoun A. Awad and Issa Khalil, "Prediction of User's Web-Browsing Behavior: Application of Markov Model", IEEE Transactions on Systems, Man, And Cybernetics—Part B: Cybernetics, Vol. 42, And No. 4, August 2012.

[9] Shreya Dubey et al. "Web Page Prediction using Hybrid Model", International Journal on Computer Science and Engineering (IJCSE) ISSN: 0975-3397

[10] Vivek Kumar Singh , Nisha Tiwari, Shekhar Garg "Document Clustering using K-means, Heuristic K-means and Fuzzy C-means", International Conference on Computational Intelligence and Communication Systems, 978-0-7695-4587-5/11, 2011 IEEE.

[11] Thanakorn Pamutha, Chom Kimpan, "Improving Web Page Prediction Using Default Rule Selection", (IJACSA) International Journal of Advanced Computer Science and Applications, .Vol. 3, No.11, 2012.

[12] Ming Syan Chen, Jong Soo Park, "Data Mining for Path Traversal Patterns in a Web Environment", Proceedings of the 16 Conference on Distributed Computing Systems, 385-392, 1996.

[13] T. I. Ibrahim, Cheng Zhong Xu. "Neural Nets Based Predictive Prefetching to Tolerate WWW Latency", Proceedings of the 20 Conference on Distributed Computing Systems, 636-643, 2000.

[14] M. Eirinaki, M. Vazirgiannis, "Usage-based Page Rank for Web Personalization", Proceedings of the Data Mining (ICDM'05), 2005

[15] Yong Zhen Guo, Kotagiri Ramamohanarao, Laurence A. F. Park. "Personalized Page Rank for Web Page Prediction Based on Access Time-Length and Frequency", Proceedings of the 2007 IEEE/WIC/ACM International Conference on Web Intelligence (WI-07), 2007

[16] Schechter S, Krishnan M, Michael DS, "Using Path Profiles to Predict Http Requests", Proceedings of the 7Web Conference, 1998.

[17] Ramesh R. Sarukkai, "Link Prediction and Path Analysis Using Markov Chains", Computer Networks, 33(1-6), 377-386, 2000

[18] Jaideep Srivastava, Robert Cooley, Mukund Deshpande, Pang-NingTan, "Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data", SIGKDD Explorations, Vol. 1, Issue 2, 12-23, 2000.

[19] Mukund Deshpande, George Karypis, "Selective Markov Models for Predicting Web-Page Accesses", Proceedings SIAM International Conference on Data Mining (SDM2001), 2001

[20]Lawrence R. Rabiner, "A Tutorial on Hidden Markov Models and Selected Applications in Speech Recognition", Proceedings of the IEEE, 257-286, 1989.

[21] Xin Jin, Huanqing Xu, "An Approach to Intelligent Web Prefetching Based on Hidden Markov Model", Proceedings of the 42th Conference on Decision and Control, 2003

The IISTE is a pioneer in the Open-Access hosting service and academic event management. The aim of the firm is Accelerating Global Knowledge Sharing.

More information about the firm can be found on the homepage:
http://www.iiste.org

CALL FOR JOURNAL PAPERS

There are more than 30 peer-reviewed academic journals hosted under the hosting platform.

Prospective authors of journals can find the submission instruction on the following page: http://www.iiste.org/journals/ All the journals articles are available online to the readers all over the world without financial, legal, or technical barriers other than those inseparable from gaining access to the internet itself. Paper version of the journals is also available upon request of readers and authors.

MORE RESOURCES

Book publication information: http://www.iiste.org/book/

Academic conference: http://www.iiste.org/conference/upcoming-conferences-call-for-paper/

IISTE Knowledge Sharing Partners

EBSCO, Index Copernicus, Ulrich's Periodicals Directory, JournalTOCS, PKP Open Archives Harvester, Bielefeld Academic Search Engine, Elektronische Zeitschriftenbibliothek EZB, Open J-Gate, OCLC WorldCat, Universe Digtial Library , NewJour, Google Scholar