Analysis of User Behaviour by Hybrid Technique

Poonam Kaushal¹, Jitendra Dangra², Dr. M.K. Rawat³

¹ M. Tech, CSE Department, LNCT, Indore, Madhya Pradesh, India. poonamkaushal14@gmail.com
² Asst. Professor, CSE Department, LNCT, Indore, Madhya Pradesh, India. jitendra.dangra@gmail.com
³ Professor & HOD, CSE Department, LNCT, Indore, Madhya Pradesh, India. cshodlnct@gmail.com

Abstract

Web page prediction problem of predicting user’s next web page request is gaining importance due to the increase in the demand of world wide web; recently Markov models are widely used for this purpose. Markov models found major application in this area different order markov models are used for providing different level of accuracy to the prediction. This paper focuses on different techniques used for predicting user’s request and also introduced the hybrid technique for web page request prediction that is the combination of markov model and the nearest neighbor model. The web is a large source of information that can be turned into knowledge. That part of knowledge concerns the usage of web itself and is valuable to website and organization of websites. Web mining is the application of data mining techniques to discover patterns from the web. Web logs contains information about web server request and response. In this paper we propose a Hybrid web page prediction technique by which we improve the performance of web page access time. In this combination technique of nearest neighbour algorithm and markov model are applied. Moreover we propose a technique for log writing and managing the logs, by which we easily find the pattern of required data.

Keywords: Hybrid Technique, Markov model, PLSA, SVR, SVM, World wide web

1. INTRODUCTION

Analysis of useful information from the world wide web require web mining. That is the application of data mining techniques to discover patterns from the web. According to analysis targets, web mining divided into three categories web usage mining, web content mining, web structure mining. Web structure mining is the technique of using graph theory to analyze the node and connection structure of a web site. Web usage mining is that area of web mining which deals with the extraction of interesting knowledge from logging information produced by the web servers. web servers maintain this information in the form of log file that contain user name which identifies who has visited the site this identification basically perform on the basis of the IP address, visiting path that determines the path that user take while visiting the particular website, path traversed .time stamp that includes the time spent by the user in each web page while surfing through the website, page last visited this shows the page visited by the user before leaving any website, success rate determines the number of downloads made and copying activity performed by the user on the particular website, user agent is the browser from which the particular user send the request to the web server, URL is the resource accessed by the user it may be HTML page, CGI program or any script, request type determines the kind of request made by the user either GET request or POST request. Log files are generally found in web servers, web proxy servers and client browsers. We focus on mining web access patterns. A web log can be regarded as a sequence of pairs of user identifier and event. Several techniques are used for web usage mining for the purpose web page request prediction some of them are support vector machine, rough set theory, neural network, genetic algorithm etc.

2. LITERATURE SURVEY

2.1 Related Work

In Web usage mining several data mining techniques can be used. 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. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together. Association rules in Web logs are discovered in. Sequence mining can be used for discover the Web pages which are accessed immediately after another. Using this knowledge the trends of the activity of the users can be determined and predictions to the next visited pages can be calculated. Sequence mining is accomplished in, where a so-called WAP-tree is used for storing the patterns efficiently. Tree-like topology patterns and frequent path traversals are searched by. Web usage mining is elaborated in many aspects. Besides applying data mining techniques also other approaches are used for discovering information. For example uses probabilistic grammar-based approach, namely an Ngram model for capturing the user navigation behavior patterns. The Ngram model assumes that the last N pages browsed affect the probability of identifying the next page to be visited. Uses Probabilistic Latent Semantic Analysis (PLSA) to discover the navigation patterns. Using PLSA the hidden semantic relationships among users and between users and Web pages can be detected. In Markov assumptions are used as the basis to mine structure of browsing patterns. For Web prefetching uses Web log mining techniques and uses a Markov predictor. Log management and intelligence. Log management (LM) comprises an approach to dealing with large volumes of computer-generated log messages (also...
known as audit records, audit trails, event-logs, etc.). LM covers log collection, centralized aggregation, long-term retention, log analysis (in real-time and in bulk after storage) as well as log search and reporting. Log management is driven by reasons of security, system and network operations (such as system or network administration) and regulatory compliance. Effectively analyzing large volumes of diverse logs can pose many challenges — such as huge log-volumes (reaching hundreds of gigabytes of data per day for a large organization), log-format diversity, undocumented proprietary log-formats (that resist analysis) as well as the presence of false log records in some types of logs (such as intrusion-detection logs). Users and potential users of LM can build their own log management and intelligence tools, assemble the functionality from various open-source components, or acquire (sub-) systems from commercial vendors. Log management is a complicated process and organizations often make mistakes while approaching it; 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.

3. WEB MINING PROS AND CONS

3.1 Pros
Web mining essentially has many advantages which makes this technology attractive to corporations including the government agencies. This technology has enabled e-commerce 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 benefit 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 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. 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 user’s interests.

4. BACKGROUND SVM
Generally, the SVM regression technique is used for identifying user’s prefetching SVM is support vector machine used for analyzing the data and recognize patterns, it is a supervised learning technique. The SVM request. It predicts the user future web page request by taking measures such as hit rate and byte hit rate into consideration. Prefetch hit (PH) maintain the number of objects that are required later by the user where as prefetch miss (PM) denotes the number of objects that are never requested by the user. User request denotes the total number of objects requested by the user. SVM implements an prefetching system server contain the various web log files such as web pages. The request first sent from client to the server and waits for response from the server. At the start cache remains empty but for each session request the system looks for that request is in the cache. SVR prediction involves or applies with request page on server and then fetch next page, with this request the web server sends the requested web page such as predicted page to the cache. If the prediction is correct sends a request if the page in the cache then calculate hit rate for the next new session of request then the cache will again send the request to the server, which applies SVR prediction model to send the next future page. User’s session based and link based. Both of these classes consider the data from the web log that contain information about the every action of the user on the web.
In this fig, data preprocessing includes tasks of data cleaning, correctness, attribute creation, selection of attribute and discretization. Data conversion is performed based on initial data by application of specific rough set technique. Here the prediction technique also involves sequential markov model and sequential association rules. The all k-th order markov model has been proposed to intelligently combine all the k-th order markov rules extracted from historical training data set and provide predictions with a partial match method. In this approach predictions are provided with all k-th order markov models by first matching the test instance against the higher order rules. If no matching rules are found additional predictions are attempted by another aspect of the artificial neural networks is that there are different architectures, which consequently requires different types of algorithms, but despite to be an apparently complex system, a neural network is relatively neural system. Why would be necessary the implementation of artificial neural networks? Although computing these days is truly advanced, there are certain tasks that a program made for a common microprocessor is unable to perform even so a software implementation of a neural network can be made with their advantages and disadvantages.

5. ROUGH SET CLUSTERING AND MARKOV MODEL

Rough set theory introduce the new way or means of clustering the web data. It clusters the data into classes so that similar objects will be in same class or dissimilar objects will be in different class. Its main aim to group the web data clustering into two classes including systems. It can be implemented in hardware, software, or a combination of both.

6. NEURAL NETWORK

Neural network is a network of neurons connected in the form of layers to accomplish some task. They basically work on the training mechanism but the learning process in neural network is slow they learn the pattern slowly and require to adjust the weights between inputs and outputs layer of the connecting links, this process continues until it obtains the correct response to the training patterns. Neural network find applications in engineering and business. ART structure used for huge widely evolving, hypertext information repository of world wide web. In this way web sites automatically improve their organization and presentation by self learning. An artificial neural network is a system based on the operation of biological neural networks, in other words, is an emulation of biological Artificial neural networks are among the newest signal processing technologies nowadays. The field of work is very interdisciplinary, but the explanation I will give you here will be restricted to an engineering perspective. In the world of engineering, neural networks have two main functions: Pattern classifiers and as non linear adaptive filters. As its biological predecessor, an artificial neural network is an adaptive system. By adaptive, it means that each parameter is changed during its operation and it is deployed for solving the problem in matter. This is called the training phase. An artificial neural network is developed with a systematic step-by-step procedure which optimizes a criterion commonly known as the learning rule. The input/output training data is fundamental for these networks as it conveys the information which is necessary to discover the optimal operating point. In addition, a non linear nature make neural network processing elements a very flexible system.

Advantages:
- A neural network can perform tasks that a linear program cannot.
- When an element of the neural network fails, it can continue without any problem by their parallel nature.
- A neural network learns and does not need to be reprogrammed.
- It can be implemented in any application.
- It can be implemented without any

Disadvantages
The neural network needs training to operate.
- The architecture of a neural network is different from the architecture of microprocessors therefore needs to be emulated.
- Requires high processing time for large neural networks.

7. FUZZY LOGIC

Fuzzy logic is also used for the purpose of prediction, prediction involves the knowledge about the past performance. Historical data provides the information on the specific pattern of data learning and it is possible to have the knowledge about the future to some extent from the past learning, fuzzy logic prediction system basically works on this concept. It basically carries two steps include clustering and specification of input and output relations by making use of IF-THEN rules. In this context, FL is a problem-solving
control system methodology that lends itself to implementation in systems ranging from simple, small, embedded micro-controllers to large, networked, multi-channel PC or workstation-based data acquisition and control. FL provides a simple way to arrive at a definite conclusion based upon vague, ambiguous, imprecise, noisy, or missing input information. FL’s approach to control problems mimics how a person would make decisions, only much faster.

8. CONCLUSIONS

As we have discussed too many techniques for web page prediction and used for the web prediction all these techniques associated with some limitations therefore we are going to propose a new hybrid technique for web page prediction in order to enhance the accuracy of prediction by 75% to 99%.

9. PROPOSED AND FUTURE WORK

In this research work, the concept of web usage mining is used for the prediction of the next web page access by the user by analyzing the user previously accessed web pages. Each web page sequence accessed by the user is stored in the form of the session 1, 2, 3…, n. Suppose user access the web page of any web site in sequence home -> About Us -> Department then this sequence will be stored as session 1 and so on. Therefore according to the behavior the model will predict the probability of next frequently accessed web page.

10. REFERENCES


[6]. Poonam kaushal –prediction of user’s next web page request by hybrid technique, IJETAE, Vol 2, No3.