ISSN No. 0976-5697

Volume 2, No. 2, Mar-Apr 2011

International Journal of Advanced Research in Computer Science

RESEARCH PAPER

Available Online at www.ijarcs.info

Designing an Intelligent Web Browser Using Web Usage Mining Techniques

Manoj pandia
SIT, BBSR RCMA, BBSR,TAT,BBSR KIIT,BBSR Utkal University,BBSR,India
manoj_pandia@yahoo.com

Subhendu pani*
SIT, BBSR RCMA, BBSR,TAT,BBSR KIIT,BBSR Utkal University,BBSR, India
subhendu_pani@rediffmail.com

Satya Biswal
SIT, BBSR RCMA,BBSR,TAT,BBSR KIIT,BBSR Utkal University,BBSR,India
satyabiswal@gmail.com

Santosh ku.
SIT, BBSR RCMA, BBSR,TAT,BBSR KIIT,BBSR Utkal University,BBSR, India
swainsantosh@yahoo.co.in

Swain Bikram ke. Ratha
SIT, BBSR RCMA, BBSR, TAT, BBSR KIIT,
BBSR Utkal University,BBSR, India
vkramus@yahoo.com

Abstract:- In the context of data mining the feature size is very large and it is believed that it needs a bigger population. Hence, this translates directly into higher computational load. With the huge amount of information available online, the web mining is a fertile area of research which applies the data mining techniques. It relates to several research communities such as Database, Information Retrieval and Visualization. We have categorized web data mining into threes areas; web content mining, web structure mining and web usage mining. In this research area that is receiving increasing attention from the data mining community.In this paper, We discuss some web mining techniques that could be used to design an efficient web browser .

Key Words: Data Mining, Web Mining, Web Data, Information Retrieval.

I. INTRODUCTION

In the era of Information Technology, accessing information is the most frequent task. Every day we have to go through several kind of information that we need and what we do? Just browse the web and the desired information is with us on a single click. Today, internet is playing such a vital role in our everyday life that it is very difficult to survive without it. The World Wide Web (WWW) has influenced a lot to both users (visitors) as well as the web site owners. The web site owners are able to reach to all the targeted audience nationally and internationally. They are open to their customer 24X7. On the other side visitors are also availing those facilities.

In the last fifteen years, the growth in number of web sites and visitors to those web sites has increased exponentially. Due to this growth a huge quantity of web data has been generated. To mine the interesting data from this huge pool, data mining techniques can be applied [7]. But the web data is unstructured or semi structured. So we can not apply the data mining techniques directly. Rather another discipline is evolved called web mining which can be applied to web data. Web mining is used to discover interest patterns which can be applied to many real world problems like improving web sites, better understanding the visitor's behavior, product recommendation etc.

Web mining [1] is the use of data mining techniques to automatically discover and extract information from Web documents/services (Etzioni, 1996). Web mining is categorized into 3 types. 1. Content Mining (Examines the content of web pages as well as results of web Searching) 2. Structure Mining (Exploiting Hyperlink Structure) 3. Usage Mining (analyzing user web navigation).

Web usage mining [2] is a process of picking up information from user how to use web sites. Web content mining is a process of picking up information from texts, images and other contents. Web structure mining is a process of picking up information from linkages of web pages, such as Table 1.

Table: 1 The Releationship Among Different Areas of Web Minig

Type	Structure	Form	Object	Collection
Usage	Accessing	Click	Behavior	Logs
Content	Pages	Text	Index	Pages
Structure	Map	Hyperlinks	Map	Hyperlinks

These 3 approaches attempts to extract knowledge from Web generate some useful result from that knowledge and apply the result to certain real world problems. Web Usage Mining is the process of applying data mining techniques to the discovery of usage patterns from data extracted from Web Log files.

Why Usage Mining

A. One can keep track of previously accessed pages of a user. These pages can be used to identify the typical behavior of the user and to make prediction about desired pages. Thus personalization for a user can be achieved through web usage mining.
B. Frequent access behavior for the users can be used to identify needed links to improve the overall performance of future accesses. Prefetching and

caching policies can be made on the basis of frequently accessed pages to improve latency time.

C. Common access behaviors of the users can be used to improve the actual design of web pages and for making other modifications to a Web site.

D. Usage patterns can be used for business intelligence in order to improve sales and advertisement by providing product recommendations.

Five major steps followed in web usage mining are

A. Data collection – Web log files, which keeps track of visits of all the visitors

B. Data preprocessing – Cleaning and structuring data to prepare for pattern extraction

C. Pattern extraction – Extracting interesting patterns

D. Pattern analysis and visualization – Analyze the extracted pattern

E. Pattern applications – Apply the pattern in real world problems

II. BASIC CONCEPTS

We will be using usage mining techniques on the back of a web browser to design an intelligent web browser. In basic concepts we will be discussing about the how a web browser works and how we can make it intelligent.

A. Web Browser

A web browser is a computer program which takes an URL as input, sends the requested URL to web server, receives the response in the form of HTML document from web server, renders the web document and displays the output to user. Examples of web browsers are, Microsoft Internet Explorer, Mozilla Firefox, Google Chrome, Safari, Netscape Navigator etc. Most of the web browsers are having common features and some browsers are having some special features like in opera the text content can be selected and can be read by computer.

Every web page on a web site is having a fixed URL, which is requested by the user. But in the web page some other resources are embedded like a) JavaScript files, b) CSS files, c) Images etc. Due to these resources inside the web page, even if user has entered only one page URL, requests for these resources also send from the browser.

As these resources i.e. images, css and javascript files are static in nature, browser does not sends request for these resources every time. Rather it keeps them in some temporary area called cache. Also browser maintains cache for web pages. But for how long the web pages will stay in cache depends on the web server configuration in header of the web page.

Web browser stores objects - for example, images, HTML documents, style sheets downloaded over the network in a special area called the browser cache. The way the cache works is simple: when the user navigates to a page, the web browser will first check if the browser cache already contains the content for that page. If the content is still fresh in the cache, another download is unnecessary.

The HTTP/1.1 protocol - the communications protocol allows specifying what content is cacheable, and for how long the downloaded content can be considered fresh by the browser cache. This information is specified in the response headers returned by the web server. Response headers are lines of text describing the page being sent (and the server that's sending it). The parts of the response header relating to cache control are called the cache control directives. Using either one of the following headers in the server response will tell the browser that the content is cacheable:

Cache-Control: max-age=specify a duration in seconds

or

Expires: a GMT date in the format specified by RFC 1123

Only one of these is needed, but if both headers are present in the server response for some inexplicable reason, the Cache-Control header takes priority over the Expires header. If you use the Cache-Control header, the cache entry will be considered fresh until the duration that you specified (in seconds) has elapsed. If you use the Expires header, on the other hand, the cache entry is considered to be fresh until the expired date arrives. The RFC 1123 standard specifies the following date time format: Thu, 01 Jan 2008 13:37:41 GMT. To specify an expiration time in the near future, it's better to use the max-age directive in the Cache-Control header, to avoid clock synchronization errors between the browser and the server. For expiration times far into the future, the Expires header is a safer but—it's more readable to humans and less error prone.

When web server sends the response to client along with the requested web page content it also sends the status code [9] of the response in headers. The status codes and their meaning are described in Table 2.

Table: 2 The Status Codes and Their Meaning

Informational 1xx	
100	Continue
101	Switching Protocols
Successful 2xx	
200	OK
201	Created
202	Accepted
203	Non-Authoritative Information
204	No Content
205	Reset Content
206	Partial Content
Redirection 3xx	
300	Multiple Choices
301	Moved Permanently
302	Found
303	See Other
304	Not Modified
305	Use Proxy
306	(Unused)
307	Temporary Redirect
Client Error 4xx	
400	Bad Request
401	Unauthorized
402	Payment Required
403	Forbidden
404	Not Found
405	Method Not Allowed
406	Not Acceptable

407	Proxy Authentication Required
408	Request Timeout
409	Conflict
410	Gone
411	Length Required
412	Precondition Failed
413	Request Entity Too Large
414	Request-URI Too Long
415	Unsupported Media Type
416	Requested Range Not Satisfiable
417	Expectation Failed
Server Error 5xx	
500	Internal Server Error
501	Not Implemented
502	Bad Gateway
503	Service Unavailable
504	Gateway Timeout
505	HTTP Version Not Supported

B. Pattern Extraction Techniques

There are two classes [3] of data mining namely i) to summarize or characterize general properties of data in repository which is called Descriptive and ii) to perform inference on current data, to make predictions based on the historical data which is called Prescriptive. There are various data mining techniques available which also can be applied to web data mining. Few techniques are listed below.

[a] Association Rules Mining: When the book Data Mining Concepts and Techniques is bought, 40% of the time the book Database System is bought together, and 25% of the time the book Data Warehouse is bought together. Those rules discovered from the transaction database of the book store can be used to rearrange the way of how to place those related books, which can further make those rules more strong

[b] Sequential Pattern Mining: Association rule mining does not take the time stamp into account, the rule can be Buy A=>Buy B. If we take time stamp into account then we can get more accurate and useful rules such as: Buy A implies Buy B within a week, or usually people Buy A every week. As we can see with the second kind of rules, business organizations can make more accurate and useful prediction and consequently make more sound decisions. A database consists of sequences of values or events that change with time, is called a time-series database, a time-series database records the valid time of each dataset. For example, in a time-series database that records the sales transaction of a supermarket, each transaction includes an extra attribute indicate when the transaction happened. Time-series database is widely used to store historical data in a diversity of areas such as, financial data, medical data, scientific data and so on. Different mining techniques have been designed for mining time-series data, basically there are four kinds of patterns we can get from various types of time-series data:1) Trend analysis, 2) Similarity search, 3) Sequential patterns and 4) Periodical patterns. Sequential patterns: sequential pattern mining is trying to find the relationships between occurrences of sequential events, to find if there exists any specific order of the occurrences. We can find the sequential patterns of specific individual items; also we can find the sequential patterns cross different items. Sequential pattern mining is widely used in analyzing of DNA sequence. An example of sequential patterns is that every time Microsoft stock drops 5%, IBM stock will also drops at least 4% within three days.

[c] Classification: Classification is to build (automatically) a model that can classify a class of objects so as to predict the classification or missing attribute value of future objects (whose class may not be known). It is a two-step process. In the first process, based on the collection of training data set, a model is constructed to describe the characteristics of a set of data classes or concepts. Since data classes or concepts are predefined, this step is also known as supervised learning (i.e., which class the training sample belongs to is provided). In the second step, the model is used to predict the classes of future objects or data. A decision tree for the class of buy laptop, indicate whether or not a customer is likely to purchase a laptop. Each internal node represents a decision based on the value of corresponding attribute, also each leaf node represents a class (the value of buy laptop=Yes or No). After this model of buy laptop has been built, we can predict the likelihood of buying laptop based on a new customer's attributes such as age, degree and profession. That information can be used to target customers of certain products or services, especially widely used in insurance and banking.

[d] Clustering: Classification can be taken as supervised learning process, clustering is another mining technique similar to classification. However clustering is a unsupervised learning process. Clustering is the process of grouping a set of physical or abstract objects into classes of similar objects [Han and Kamber 2000], so that objects within the same cluster must be similar to some extent, also they should be dissimilar to those objects in other clusters. In classification which record belongs which class is predefined, while in clustering there is no predefined classes. In clustering, objects are grouped together based on their similarities. Similarities between objects are defined by similarity functions, usually similarities are quantitatively specified as distance or other measures by corresponding domain experts. For example, based on the expense, deposit and draw patterns of the customers, a bank can clustering the market into different groups of people. For different groups of market, the bank can provide different kinds of loans for houses or cars with different budget plans. In this case the bank can provide a better service, and also make sure that all the loans can be reclaimed.

Several algorithms are used for Association Rule Mining (ARM) by different authors which can be summarized below.

Table: 3The Research Work Done on Arm

Algorithm Used	Authors	Year
Maximal forward references [5]	Ming-Syan Chen, Jong Soo Park, Philip S. Yu	1998
Markov Chains [6]	Jianhan Zhu, Jun Hong, and John G. Hughes	2002
Improved AprioriAll [7]	WANG Tong, HE Pi-lian	2005
Fpgrowth and Prefixspan [8]	Hengshan Wang, Cheng Yang, Hua Zeng	2006
Custom Built APRIORI Algorithm [4]	Sandeep Singh Rawat, Lakshmi Rajamani	2010

III. PROBLEM DEFINITION

In order to browse a web site, user enters the URL for the homepage of the web site. The browser sends the request to web server to get the requested web page. Some time is elapsed in that process. Once the page is received and loaded in the browser, user goes through the web page then clicks on another link which he wants. And this process continues.

In this process user will be waiting for the main page to be completely displayed so that he will be able to see the next link to be clicked. When he is able to get the link, he clicks the link and again few seconds are elapsed to get that web page. Ex. User visits www.kiit.ac.in. Then clicks Faculty link. Then selects School of Computer Engg. Even If this browsing pattern is frequent, browser does not aware about this, so user has to wait for the pages to be fetched and displayed.

There are possibilities, when the user enters an URL and the URL is invalid or the request URL does not exist. In that case browser handles it in different ways. Some browsers i.e. IE and Chrome, takes this string as a search string, passes it to search engines and displays the result accordingly. In some other browsers they simply sends the request to web server, but as by that name no domain exists user may get another web site displaying the requested URL domain name.

Generally when we visit any secure website like banking sites it displays our last login date and time. From this we are able to know that nobody else except us has logged in since our last visit if we have remembered our last login date and time. The website does not displays the history of our activities for last month or year.

IV. MOTIVATION

As described above, the existing web browsers are having these difficulties, which can be addressed and a new web browser can be designed where the waiting time can be reduced by pre-fetching predicted web pages, along with the request for invalid URL can also be handled.

V. PROPOSED SOLUTION

As described earlier the difficulties with the existing browser can be summarized like this.

A. Waiting time to get the link and page load time when the link is selected.
B. Waiting time for the invalid URL and URL which does not exist.
C. History of login activities for secured (banking) URLs.

One approach to handle all the difficulties is to use web usage mining techniques on client side.

When user enters an URL or selects a hyperlink a request is generated. These requests can be stored in a log file which is maintained by the browser. Then the ARM algorithms can be applied to predict the next URL and request can be send in advance also the response will be kept locally. So when the user follows that link instead of send it to web server the page can be delivered from local so that the waiting time can be reduced.

The proposed architecture is given in figure-1

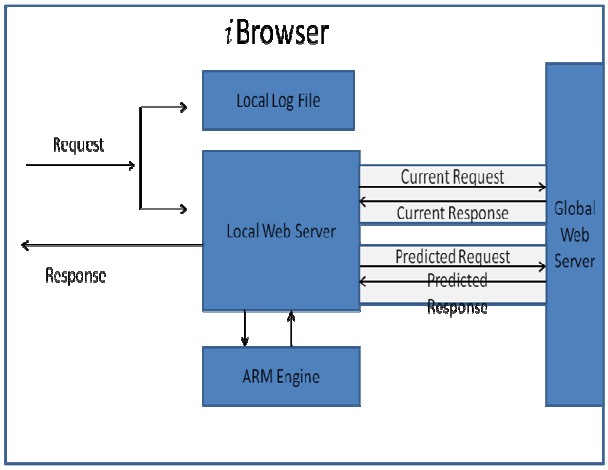

Figure 1: Proposed Architecture for iBrowser

[a] Description of the Terms Used

Local Log File – In order to apply the ARM algorithms we need a log file. So we are maintaining a local log file on the client side. Whenever user enters an URL or follows a hyperlink, a log record is stored in this file. Unlike log format of general web servers like Apache, which contains IP address, userid, useragent etc. we are not keeping all these fields. Because in our case this is the only browser from which the request will go, and as the browser will be running on the local machine which is used by a specific user so no need to keep track of the user, browser (user agent) and the IP address. Our local lof file format will be like

ReqID URL Date Time Referer StatusCode

Where ReqID is auto-generated, URL is the requested URL, Date is the requested date, Time is the requested time, Referer is the URL from which user has requested this URL, and the status code of that request

Also another log file is maintained which stores only those records whose status code is 4XX. The format is

URL StatusCode

This is used to check for the URL against these codes later so as to provide response to user quickly without sending to server again

[i] Local Web Server – We are maintaining a local web server which will be basically providing two features. It will work as a File Repository as well it maintains a File Table. Whenever the response comes from the Global Web Server as web pages, these web pages will be stored in the File Repository and a record will be maintained in File Table about the URL and the data and time when that web page is fetched. This date and time is used to know whether the page is older one or new one according to some amount of time.

[ii] ARM Engine - Engine which implements any ARM algorithms on the Local Log File. The algorithm takes an URL and gives the predicted URL if any.

[b] How it Works

The work flow of the browser is described below

[i] User enters an URL or follows a hyperlink
[ii] The request is recorded in Local Log File
[iii] Search the URL in error log file. If invalid or URL does not exist send response to user and exit.
[iv] Requested URL is searched in File Table to check whether the web page present in File Repository is valid or not.

[v] If valid sends the web page as response to user and pass the URL to ARM Engine to get the predicted URL.

[vi] If any URL predicted send the request to Global Web Server and stores the response in File Repository, as well update that record in File Table

VI. CONCLUSION

The web usage mining techniques are applied on web servers as well as proxy servers to extract interesting patterns. As mentioned earlier the patterns can be extracted out of classification or clustering. Also the user's next request can also be predicted by applying association rule mining (ARM) techniques. We have proposed here is to apply ARM algorithms on the background of client's web browser so as to predict the next movement and pre-fetch that webpage so that the waiting time for webpage can be reduced. Also this approach has a limitation like if the pre-fetched webpage is not visited by the user then there is wastage of bandwidth. But if the prediction done by ARM algorithm is quite successful then it will be of great use. This approach is applicable only to frequently visited web pages. For others, it will behave like any other browser.

VII. REFERENCES

[1] Chen Hu, Xuli Zong, Chung-wei Lee and Jyh-haw Yeh, "World Wide Web Usage Mining Systems and Technologies", Journal of SYSTEMICS, CYBERNETICS AND INFORMATICS Vol. 1, No. 4, Pages53-59, 2003.

[2] Florent Masseglia, Pascal Poncelet, Rosine Cicchetti, **"**An efficient algorithm for Web usage mining", Networking and Information Systems Journal. Volume X, 2000

[3] Qiankun Zhao, Sourav S. Bhowmick, "Sequential Pattern Mining: A Survey", Technical Report, CAIS, Nanyang Technological University, Singapore, No. 2003118 , 2003.

[4] Sandeep Singh Rawat, Lakshmi Rajamani, "Discovering Potential User Browsing Behaviors Using Custom-Built APRIORI Algorithm", International journal of computer science & information Technology (IJCSIT) Vol.2, No.4, August 2010

[5] Ming-Syan Chen, Jong Soo Park, Philip S. Yu, "Efficient Data Mining for Path Traversal Patterns", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 10, NO. 2, MARCH/APRIL 1998

[6] Jianhan Zhu, Jun Hong, John G. Hughes, "Using Markov Chains for Link Prediction in Adaptive Web Sites", Soft-Ware 2002, LNCS 2311, pp. 60–73, 2002

[7] WANG Tong, HE Pi-lian, "Web Log Mining by an Improved AprioriAll Algorithm", World Academy of Science, Engineering and Technology 4 2005

[8] Hengshan Wang, Cheng Yang, Hua Zeng, " Design and Implementation of a Web Usage Mining Model Based On Fpgrowth and Prefixspan", Communications of the IIMA 2006 Volume 6 Issue 2

[9] http://www.w3.org/Protocols/rfc2616/rfc2616-sec10.html