User Navigation Pattern Prediction using Longest Common Subsequence

Samir S. Shaikh
SRESCOE
Kopargaon

Pravin B. Landage
SRESCOE
Kopargaon

D. B. Kshirsagar
SRESCOE
Kopargaon

ABSTRACT

Web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data and traces users' visiting characteristics, and then extracts the users' using pattern. Web mining technologies are the right solutions for knowledge discovery on the Web. The knowledge extracted from the Web can be used to raise the performances for Web information retrievals, question answering, and Web based data warehousing. In this paper, I provide an introduction of Web mining as well as a review of the Web mining categories. Web mining applies the data mining, the artificial intelligence and the chart technology and so on to the web data. And traces users' visiting characteristics, and then extracts the users' navigation pattern.

Web mining has quickly become one of the most important areas in Computer and Information Sciences because of its direct applications in ecommerce, e-CRM, Web analytics, information retrieval and filtering, and Web information systems.

General Terms

Longest common subsequence algorithm, Graph partitioning algorithm.

Keywords

Web usage mining, longest common subsequence, graph partitioning, navigation pattern.

1. INTRODUCTION

As many believe, it is Oren Etzioni first proposed the term of Web mining in his paper 1996 [5]. In that paper, he claimed the Web mining is the use of data mining techniques to automatically discover and extract information from World Wide Web documents and services. Many of the following researchers cited this explanation in their works. In the same paper, Etzioni came up with the question: Whether effective Web mining is feasible in practice? Today, with the tremendous growth of the data sources available on the Web and the dramatic popularity of e-commerce in the business community, Web mining has become the focus of quite a few research projects and papers. Some of the commercial consideration has presented on the schedule. Way to decompose Web mining into the following subtasks:

a) Resource Discovery: the task of retrieving the intended information from Web.

b) Information Extraction: automatically selecting and pre-processing specific information from the retrieved Web resources.

c) Generalization: automatically discovers general patters at the both individual Web sites and across multiple sites.

d) Analysis: analysing the mined pattern.

In brief, Web mining is a technique to discover and analyses the useful information from the Web data. The Web involves three types of data: data on the Web (content), Web log data (usage) and Web structure data. Data type as content data, structure data, usage data, and user profile data. The Web mining categorized into Web usage mining, Web text mining and user modeling mining; while today the most recognized categories of the Web data mining are Web content mining, Web structure mining, and Web usage mining. It is clear that the classification is based on what type of Web data to mine [4].

2. RELATED WORK

Identifying Web browsing strategies is a crucial step in Website design and evaluation, and requires approaches that provide information on both the extent of any particular type of user behavior and the motivations for such behavior [6].Pattern discovery from web data is the key component of web mining and it converge algorithms and techniques from several research areas. Baraglia and Palmerini [7] proposed a WUM system called SUGGEST that provide useful information to make easier the web user navigation and to optimize the web server performance. Liu and Keselj [8] proposed the automatic classification of web user navigation patterns and proposed a novel approach to classifying user navigation patterns and predicting users' future requests and Mobasher [9] presents a Web Personalizer system which provides dynamic recommendations, as a list of hypertext links, to users. Jalali et al. (2008) [10] proposed a system for discovering user navigation patterns using a graph partitioning model. An undirected graph based on connectivity between each pair of Web pages was considered and weights were assigning to edges of the graph. Dixit and Gadge (2010) [11] presented another user navigation pattern mining system based on the graph partitioning. An undirected graph based on connectivity between Referrer and URI pages was presented along with a pre-processing method to process unprocessed web log file and a formula for assigning weights to edges of the undirected graph.

3. PROPOSED SYSTEM

The general architecture of the PUCC (Predicting User Navigation Patterns Using Clustering and Classification) system is given in Figure 1. The heart of the PUCC system is the web log data, which stores all the successful hit made in the Internet. A hit is defined as a request to view a HTML document or image or any other document. The web log data are automatically created and can be obtained from either client side server or proxy server or from an organization database. Each entry in the web log data includes details like the IP address of the computer making the request, user ID, date and time of the request, a status field indicating if the request was successful, size of the file transferred, referring

URL (URL of the page which contains the link that generated the request), name and version of the browser being used [1].

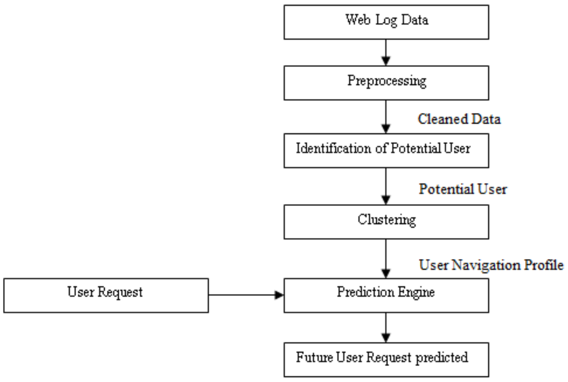


Fig 1: PUCC System Overview

3.2 Preprocessing

The first step of PUCC is the pre-processing of web log data, where the unformatted log data is converted into a form that can be directly applied to mining process. The pre-processing steps include cleaning, user identification and session identification. Cleaning is the process which removes all entries which will have no use during analysis or mining.

3.1.1 Web Log File

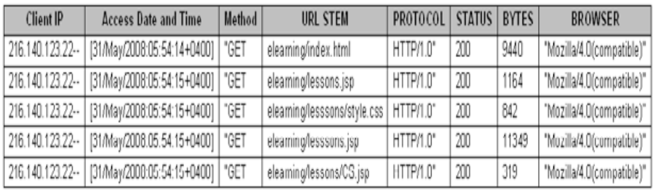
The heading for subsubsections Web log file is log file automatically created and maintained by a web server. Every "hit" to the Web site, including each view of a HTML document, image or other object, is logged. The raw web log file format is essentially one line of text for each hit to the web site. At least two log file formats exists: Common Log File format (CLF) and Extended Log File format, robots do not follow the proposed standard. Thus to delete robot entries the following procedure is used. A sample web log file is shown in Figure 4.2. The information in web log file represent the navigation patterns of different segments of the overall web traffic, ranging from single-user, single-site browsing behaviour to multi-user, multi-site access patterns. Irrespective of the source of collection, the web log file has the following general characteristics [1].

1. The log file is text file. Its records are identical in format.

2. Each record in the log file represents a single HTTP request.

3. A log file record contains important information about a request: the client side host name or IP address, the date and time of the request, the requested file name, the HTTP response status and size, the referring URL, and the browser information.

A browser may fire multiple HTTP requests to Web server to display a single Web page. This is because a Web page not only needs the main HTML document; it may also need additional files, like images and JavaScript files. The main HTML document and additional files all require HTTP requests [1].



Client IP	Access Date and Time	Method	URL STEM	PROTOCOL	STATUS	BYTES	BROWSER
216.140.123.22--	[31/May/2008:05:54:14+0400]	"GET	elearning/index.html	HTTP/1.0"	200	9440	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008:05:54:15+0400]	"GET	elearning/lessons.jsp	HTTP/1.0"	200	1164	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008:05:54:15+0400]	"GET	elearning/lesssons/style.css	HTTP/1.0"	200	842	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2008:05:54:15+0400]	"GET	elearning/lesssons.jsp	HTTP/1.0"	200	11349	"Mozilla/4.0(compatible)"
216.140.123.22--	[31/May/2000:05:54:15+0400]	"GET	elearning/lessons/CS.jsp	HTTP/1.0"	200	319	"Mozilla/4.0(compatible)"

Fig 2: Sample Web Log File

3.2.2 Cleaning

Data cleansing, data cleaning, data wash or data scrubbing is the process of detecting and correcting (or removing) corrupt or inaccurate records from a record set, table, or database. Used mainly in databases, the term refers to identifying incomplete, incorrect, inaccurate, irrelevant, etc. parts of the data and then replacing, modifying, or deleting this dirty data.

a. Parsing: Parsing in data cleansing is performed for the detection of syntax errors. A parser decides whether a string of data is acceptable within the allowed data specification. This is similar to the way a parser works with grammars and languages.

b. Data transformation: Data transformation allows the mapping of the data from its given format into the format expected by the appropriate application. This includes value conversions or translation functions, as well as normalizing numeric values to conform to minimum and maximum values.

c. Duplicate elimination: Duplicate detection requires an algorithm for determining whether data contains duplicate representations of the same entity. Usually, data is sorted by a key that would bring duplicate entries closer together for faster identification.

d. Statistical methods: By analysing the data using the values of mean, standard deviation, range, it is possible for an expert to find values that are unexpected and thus erroneous. Although the correction of such data is difficult since the true value is not known, it can be resolved by setting the values to an average or other statistical value. Statistical methods can also be used to handle missing values which can be replaced by one or more plausible values.

3.3 Identification of Potential User

Data classification is the categorization of data for its most effective and efficient use. In a basic approach to storing computer data, data can be classified according to its critical value or how often it needs to be accessed, with the most critical or often-used data stored on the fastest media while other data can be stored on slower (and less expensive) media. This kind of classification tends to optimize the use of data storage for multiple purposes - technical, administrative, legal, and economic.

In this project we only deals with potential user data so we are classifying log file pre-processed data into two categories:

a. Potential User data

b. Non potential User data

This step of PUCC focuses on separating the potential users from others. Decision tree classification using C4.5 algorithm to identify interested users. They use a set of decision rules for this purpose. The algorithm worked efficiently in identifying potential users, but had the drawback that it completely ignored the entries made by network robots. Search engines normally use network robots to crawl through the web pages to collect information. The number of records created by these robots in a log file is extremely high and has a negative impact while discovering navigation pattern. This problem is solved in this paper by identifying the robot entries first before segmenting the user groups into potential and not-potential users. Entries in web log made by network robots can be identified by their IP address and agents. But this might require knowledge on all type of agents and search engines, which is difficult to obtain. An alternative way is to study the robots.txt file (located at the website's root directory), as a network robot must always read this file before accessing the website. This is because the robots.txt has the access details of the website and each robot is request to know its access right before scrawling. But this cannot be always relied on since compliance to robot exclusion standard is voluntary and most of the

a. Detect and remove all entries which has accessed robots.txt file

b. Detect and remove all entries with visiting time of access as midnight (commonly used as the network activity at that time is light)

c. Remove entry when access mode is HEAD instead of GET or POST

d. Compute browsing speed and remove all entries whose speed exceeds a threshold T1 and number of visited pages exceeds a threshold T2.

The browsing speed is calculated as the number of viewed pages / session time. After handling the network robot entries, a series of decision rules are applied to group the users as potential and not-potential users. Given a set of training data containing valid log attributes, C4.5 classification algorithm is used to classify the users. The attributes selected are time (>30 seconds), number of pages referred in a session (Session time=30 minutes) and the access method used. The decision rule for identifying potential users is "If Session Time > 30 minutes and Number of pages accessed > 5 and Method used is POST then the classify user as "Potential" else classify as "Not-Potential". The purpose of introducing classification is to reduce the size of the log file. This reduction in size will help for efficient clustering and prediction [8].

3.4 Clustering Process

We use a graph partitioned clustering algorithm to group users with similar navigation pattern. An undirected graph based on the connectivity between each pair of web pages is used. Each edge in the graph is assigned a weight, which is based on the connectivity time and frequency. Connectivity Time measures the degree of visit ordering for each two pages in a session [10].

$$TC_{a,b} = \frac{\sum_{i=1}^{N} \frac{T_i}{T_{ab}} \times \frac{f_a(k)}{f_b(k)}}{\sum_{i=1}^{N} \frac{T_i}{T_{ab}}} \qquad (1)$$

T_i is the time duration of i^{th} session that contain both a and b pages, Tab is the difference between requested time of page a and page b in the session, f (k) = k if web page appears in position k.

Frequency measures the occurrence of two pages in each session (Equation 2).

$$FC_{a,b=} \frac{N_{ab}}{Max\{N_a, N_b\}} \qquad (2)$$

Where N_{ab} is the number of sessions containing both page a and page b. N_a and N_b are the number of session containing only page a and page b. Both the formulas normalize all values for time and frequency are between 0 and 1. Both these are considered as two indicators of the degree of connectivity for each pair of web pages and is calculated using Equation (3).

$$W_{a,b} = \frac{2 \times TC_{ab} \times FC_{ab}}{TC_{ab} + FC_{ab}} \qquad (3)$$

The data structure can be used to store the weights is an adjacency matrix M where each entry M_{ab} contains the value W_{ab} computed according to Equation (3) .To limit the number of edge in such graph ,element of M_{ab} whose value is less than a threshold are too little correlated and thus discarded. This threshold is named as MinFreq in this contribution [1], [11].

Graph Partitioning Algorithm:

```
1. L[p] = P; // Assign all URLs to a list of web pages.
2. For each (Pi, Pj) L[p] do // for all pair of web pages
3. M (i, j) = Weight Formula (Pi, Pj); //computing the
   weight based on Equation (3)
4. Edge (i, j_) = M (i, j); End for for all Edge (u, v) Є
   Graph (E, V) do
  // removing all edges that its weight is below than
   MinFreq
5. If Edge (u, v) < MinFreq then Remove (Edge (u,
   v));
 End if End for for all vertices (u) do Cluster[i]=DFS (u);
  // perform DFS If cluster[i] < MinClusterSize
 // remove cluster whose length is below MinClusterSize
6. Remove (Cluster[i]); End if i = i + 1 end for return
   (Cluster)
```

3.5 Prediction Engine

The main objective of prediction engine in this part of architecture is to classify user navigation patterns and predicts users' future requests. This paper uses the Longest Common Subsequence algorithm during prediction. The main aim of LCS is to find the longest subsequence common to all sequences in a set of sequences. This method is discussed in this section. The algorithm works with two features. The first property states that if two sequences X and Y both end with the same element, then their LCS will be found by removing the last element and then finding LCS of the shortened sequence. The second property is used when the two sequences X and Y does not end with the same symbol [1].

Then, the LCS of X and Y is the longest sequence of LCS (X_n, Y_{m-1}) and LCS (X_{n-1}, Y_m). Thus, the LCS can be formulated using Equation (4).

4. CONCLUSION

The system consists of four stages. The first stage is the cleaning stage, where unwanted log entries were removed. In the second stage, cookies were identified and removed. The result was then segmented to identify potential users. From the potential user, a graph partitioned clustering algorithm was used to discover the navigation pattern. An LCS classification algorithm was then used to predict future requests. Experiment will be performed on Web log file its records are identical in format. Experimental result will be based on prediction engine's performance. Performance is get measured by three parameters accuracy, coverage and F1 measure.

5. REFERENCES

[1] V. Sujatha, Punithavalli, "*Improved user navigation pattern prediction technique from web log data*", International Conference on Communication Technology and System Design, 2011, Elsevier publication, Procedia Engineering 30 (2012) pp. 92 – 99

[2] Yue-Shi Lee, Show-Jane Yen, "*Incremental and interactive mining of web traversal patterns*", 2008, Elsevier publication, Information Sciences 178 (2008) pp.287–306.

[3] Neetu Anand, Saba Hilal, "*Identifying the User Access Pattern in Web Log Data*", International Journal of Computer Science and Information Technologies, Vol. 3 (2) , 2012, pp.3536-3539

[4] Yan Wang, "Web Mining and Knowledge Discovery of Usage Patterns" Google Documents, 2000.

[5] Oren Etzioni, "*The world wide Web: Quagmire or gold mine*", Communications of the ACM, 39(11), 1996, pp. 65-68

[6] Kumar, P.R. and Singh, A.K., *"Web Structure Mining: Exploring Hyperlinks and Algorithms for Information Retrieval"*, American Journal of Applied Sciences, 2010, Vol. 7, No.6, Pp. 840-845.

[7] Ranieri Baraglia, Paolo Palmerini, "*SUGGEST : AWeb Usage Mining System*", Proc. of IEEE International Conference on Information Technology: Coding and Computing, 2004.

[8] Haibin Liu, Vlado Kesˇelj, "Combined mining of Web server logs and web contents for classifying user navigation patterns and predicting users' future requests", Data & Knowledge Engineering 61 2007, Elsevier publication, pp. 304–330.

[9] Bamshad Mobasher, Robert Cooley, and Jaideep Srivastava, *"Automatic personalization based on web usage mining"*, Communication of ACM Vol. 43 No.8, 2000, pp.142-151.

[10] Mehrdad Jalali, Norwati Mustapha, Md Nasir Sulaiman, Ali Mamat, *"A Recommender System Approach for Classifying User Navigation Patterns Using Longest Common Subsequence Algorithm"*, American Journal of Scientific Research, ISSN 1450-223X Issue 4 (2009), pp. 17-27.

[11] Dipa Dixit, Jayant Gadage, *"New Approach for Clustering of Navigation Patterns of Online Users"*, International Journal of Engineering Science and Technology, Vol. 2(6), 2010, pp.1670-1676.