Towards Knowledge Discovery from WWW Log Data

Feng Tao tao.feng@qub.ac.uk
Fionn Murtagh f.murtagh@qub.ac.uk

School of Computer Science
Queen's University of Belfast
Belfast, BT7 1NN
Northern Ireland, UK

Abstract

As the result of interactions between visitors and a web site, an http log file contains very rich knowledge about users on-site behaviors, which, if fully exploited, can better customer services and site performance. Different to most of the existing log analysis tools which use statistical counting summaries on pages, hosts, etc., we propose a transaction model to represent users access history and a framework to adapt data mining techniques such as sequence and association rule mining to these transactions. In this framework, all transactions are extracted from the raw log file though a series of step by step data preparation phases. We discuss different methods to identify a user, and separate long convoluted sequences into semantically meaningful sessions and transactions. A new feature called interestingness is defined to model user interests in different web sections. With all the transactions being imported into an adapted cube structure with a concept hierarchy attached to each dimension of it, it is possible to carry out multi-dimensional data mining at multi-abstract levels. Using interest context rules, we demonstrate the potentially significant meaning of this system prototype.

1. Introduction

While the Internet has been growing to be an important part of peoples' lives, there are more and more data recorded of users online activities. These data can contain valuable information on a user's behavior, which, if fully used, can guide the resource maintenance, provide personalized services and support target marketing in e-business.

In this paper, we propose a solution to extract knowledge from the HTTP web log which implies large quantities of information on user behavior recorded while using the services in the web site. Web log is a web server mechanism that records each single request coming from the visitors for the services in the server. Each entry in the log file has three important fields: the IP address from which the request originated, the description of the request and a time stamp indicating when this occurs. A popular web site can see its web log growing by hundreds of megabytes every day. Pioneering researchers have done quite good work on investigating the web log data to improve system design, understand the nature of web traffic, and discover user preferences [8,16,19]. Even more novel ideas have been proposed to design auto-adaptive sites [15] and improve proxy cache efficiency [6] based on learning visitors' access patterns from the web log file. On the application side, there are more than 20 commercial web log analysis tools available in the internet such as ilux Edge[10], Marketwave [12] and WebTrends [20]. But most of them reveal only frequency counts at predefined primitive conceptual levels without a systematic underlying data model. The most common reports of these tools are the following: a summary report of the hits over a certain period of time, top n files visited most frequently, hits per domain report, etc. One can easily notice that they only show the statistical results of the raw data recorded in the log file and after that human intervention is still needed to generalize user access behavior at high and abstract concept levels.

Besides, the performance of these tools declines quickly when the size of the dataset increases. The recent progress in data warehouse research has made available powerful data warehousing systems with OLAP (OnLine Analytical Processing) ability [9,17]. We construct a data cube as the intermediate layer between the knowledge discovery modules and the underlying relational database. This layer is designed to meet the OLAP (OnLine Analytical Processing) standard. This data model enables the data mining task to be carried out in a more straightforward manner from various points of view at different aggregation concept levels. We then study how to populate and aggregate data into this data structure as well as that metadata which seeks to embody time relationships of the items so that sequences can be manipulated easily. Based on this data model, we study how to manipulate the cube to apply sequence mining algorithms on it efficiently.

As already noted, many current analysis tools try to describe the user's behavior in a specific web site by a list of hit counts. We believe that the knowledge representing user behavior goes far beyond frequent access to some pages. [7] did quite good work in this direction while our contributions are the following: the idea of using a transaction model to represent visitors traces in web log data; the proposal to use interest context rules to represent inter-relationship among different sections of the web site according to user interest; the definition of interestingness as a measurement of the rules and method for its computation in the data preparation phases.

The rest of this paper is organized as follows: Section 2 proposes the framework that enables the new concept of knowledge discovery on web log data. A breakdown of the data flow is also described there with an emphasis on data preparation phases; Section 3 presents a transaction model to present the user's visited sessions. Due to the space limitation, we are not going to describe the cube structure for OLAP in this paper and refer the readers to an extend version of this paper. Then we present the experimental results with semantic explanation before coming to the conclusion and outlook.

2. Architecture and Data Flow

A web server access log contains a complete history of file accesses by clients. Most WWW access logs follow the Common Log Format specified as part of the HTTP protocol by CERN and NCSA [11]. A log entry, according to that definition, has the structure shown in Figure 2.1. Figure 2.2 is a portion of HTTP log data.

***Client*:** visitor's domain name or IP address that can be resolved to domain name
Auth*: username if registered
Timestamp: Date and time of the access
Request: request method, document path and name, parameters, etc.
Status: status code indicating the result of a request
User agent*: client side browser type
Cookie*: Crookie ID
Referrer*: previews link address

Figure 2.1 HTTP Server's Log Structure
(*:- optional fields)

```
atz.cube.net  [01/Jun/1999:00:01:29 +0200] "GET /garching-
info/computing/weatherdir/europe_meteo.jpg HTTP/1.0" 200
55540
eu.ansp.br  [01/Jun/1999:00:01:55 +0200] "GET /garching-
info/computing/weather.html HTTP/1.0" 304 177
200.247.224.105  [01/Jun/1999:00:02:04 +0200] "GET /
HTTP/1.0" 200 5953
200.247.224.105  [01/Jun/1999:00:02:09 +0200] "GET
/icons/redball.gif HTTP/1.0" 200 398
eu.ansp.br  [01/Jun/1999:00:02:14 +0200] "GET
/icons/blueball.gif HTTP/1.0" 304 436
200.247.224.105  [01/Jun/1999:00:02:09 +0200] "GET
/icons/redball.gif HTTP/1.0" 200 398
```

Figure 2.2 A fraction of sample log data

The primary objective of this log mining is to apply generic data mining techniques to transaction data in order to discover interesting patterns in user accesses to various sections within a specific web site.

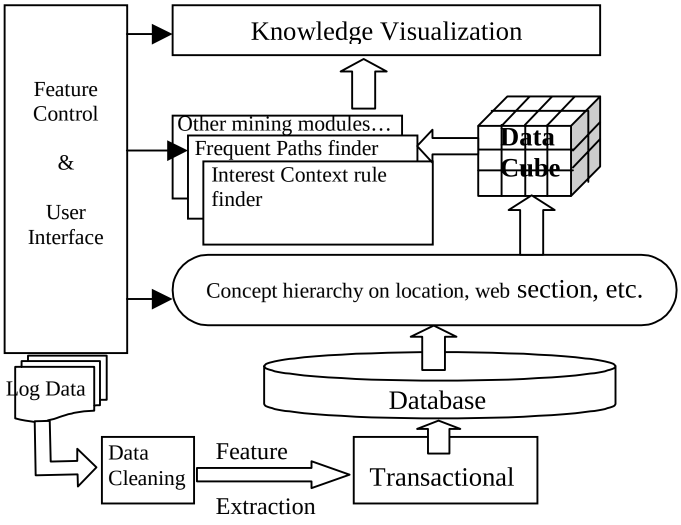

Figure 2.3 System Architecture

The architecture of the system which we propose, as shown in Figure 2.3, is roughly comprised of four components: the data cleaning/preparation module that extracts relevant information from the raw log data by filtering, translating, field encoding, measurement selecting and calculating. All of these results are then interpreted by a transaction model to generate different transactions and sessions of the site browsing trace with some measurements integrated in. With all the necessary data collected in the first phase and stored in a database, we start the data aggregation phase adapting OLAP technology [14] in order to adjust the data to the desired level where data mining algorithms can be applied later on. In this module, a series of concept hierarchy trees is designed and a cube structure is defined to store transaction data with a hierarchy tree attached to each of the cube dimensions if applicable. The data handling component comprises data mining approaches such as interest context rule mining, sequence pattern discovery, etc. The last component is the interface part, which can be described as feature control and knowledge visualization. The former takes care of the finer mining parameters inside the rest of the data processing modules which can be controlled via a query mechanism. The latter presents the discovered knowledge in an efficient and straightforward manner.

2.1 Data Cleaning

As we have seen in Figure 2.2, web access information is stored in a flat file which is not immediately interpretable. Many entries are considered as irrelevant to our final goal, thus being removed at this step. This filtering is application dependent, but usually those accesses to image icon files are filtered out since most of them are triggered automatically on loading an HTML page, and it is the latter which, on the whole, constitutes the important factor describing user traversal patterns and their context interest within the web site. Data cleaning is performed by checking the extension of the URL name. For instance, all the log entries with filename extension such as GIF, JPEG, JPG and map are removed from further processes. One exception is an art gallery site where images are the main concern. In that case, we can configure the filter parameters in the feature selecting interface module as shown in Figure 2.3.

2.2 Transactional Model and Data

Unfortunately, unlike the classical “basket” data mining solutions [18] where transaction is defined as a list of itemsets, there is no natural definition of a user transaction in the site navigation scenario where the HTTP server’s multi-thread and multi-user features make user navigational traces nest together. At one extreme, we can consider each log entry to be a separate transaction, while on the other hand we can define a time window to crop a transaction to be a series of adjacent accesses from the same host. Therefore, a semantic transaction model must be defined with extra effort. In the following section, we define this model along with the concept of sessions and discuss how to form transactions from the log entries.

Definition 2.1 Let e∈E, E is the set of all log entries after data cleaning and e is a log entry object with direct attributes such as host, document, timestamp etc, and derived features such as dwell time, interestingness and session ID, etc. (for example, e.interestingness, interestingness∈D.) We denote all these attributes as D and the Domain of D as Dom(D).

A transaction E[n] can be defined in the following model:

$$E[n] = \{\ E[n][s]\ \big|\ E[n][s].host = E[n].host \text{ AND } E[n][s+1].timestamp - E[n][s].timestamp < W.\Delta T,\ 1 \le s \le |E[n]|\ \}$$

where n is the transaction ID number, s is the session ID number within a specific transaction, ΔT is the time interval configured in the window object W, which will be defined in Definition 2.2. The following illustrates these transaction and session concepts. We can notice that all the log data have been transformed to transactions and sessions, the session is an object with various encapsulated measurements contributing to the user’s overall behavior.

Transactions	Sessions			
E[i]	E[i][1]	E[i][2]	E[i][3]	E[i][4]
E[j]	E[j][1]	E[j][2]	E[j][3]	
E[k]	E[k][1]	E[k][2]	E[k][3]	
… …	… …	… …		

Figure 2.4 Transactions and Sessions

Definition 2.2 A window W in a transaction set E is a mechanism to categorize sessions into different transactions according to its integrated attribute window-size ΔT, which can be adjusted to simulate the time-span of the browsing transaction semantic. We set it as 30 minutes in our case.

Definition 2.3 Session s = E[n][i] is the ith element in a specific transaction E[n], where 1 ≤ i ≤ |E[n]|, s is an object encapsulating all the attributes and features that are possibly derived from either a single log entry or the context entries within a single transaction. Figure 2.5 shows a list of features for a session object. We denote these features as D ∈ Dom(D).

Since the research work on association mining [2, 3] and sequence mining [4], most subsequent researches have been taking it for granted that the transaction is only a set of literals called items which do not have any attributes themselves. Therefore the association rule mining algorithms applied to this transaction model were constrained to the counting of the item occurrences and finding dominant inter and intra co-occurrence relationships among the items set. In our research, we improved that model to be a multi-feature session model that makes it possible to carry out more detailed intra association rule mining, i.e., the encapsulated features inside the items are also involved. For instance, using this model with interestingness integrated as a feature of the items, we can discover the hidden interest context rules within the site.

Feature D	Description
Visit ID	Transaction.visit ID
Session	Its sequence number in a transaction
Action	Derived from the document name
Time	Time stamp
Interestingness	Function of relative dwell time and volume of the content

Figure 2.5 Features in a session

It can be noticed from Figure 2.4 and Figure 2.5 that the transaction-session model is somewhat like a matrix with features integrated for each of the matrix elements. This is extremely useful when traversing the transaction-session structure to collect the data and to feed the mining algorithms.

By applying this model to the log data, we can easily transform it to a transaction set, upon which data mining algorithms can be applied to discover interest context rules, etc. Next, we give the definition of interestingness and the method to compute it.

3. Feature selection

There are lots of mining targets in the web log data that have been exploited by recent research in this area. The web page accessed, the domain name resolved from the host IP address, all of these, after being aggregated, can produce significant patterns indicating the user's distribution and the web sections of interest based on hit count and the frequency. Besides these, however, some more useful yet hidden measurements can be obtained by analyzing the visiting context and combining other attributes. We studied the patterns of Internet surfers and found that a user usually spends longer time gazing at an interesting page than they do on those less interesting pages. Noticing that the dwell time is also dependent on the information volume and the network speed in different areas, we firstly calculate the relative dwell time within a single transaction to exclude the factor of different network speed. We then divide this relative dwell time by the size of the page and thus get a new measurement termed *relative interestingness* to pinpoint the user preference issue in web log data mining more effectively. We call this interestingness for brevity. After grouping the log entries according to the host IP and timestamp, the following equation is applied to compute the interestingness for each of the log entries *E[Transaction_ID][Session_ID]* in a transaction *E[Transaction_ID]*.

$$E[k][i].Interestingness = \frac{E[k][i].dt \Big/ \frac{\sum_{j=1}^{n} E[k][j].dt}{n}}{E[k][i].size}$$

$$E[k][i] \in E[k],$$

$$E[k][i].dt = E[k][i+1].timestamp - E[k][i].timestamp,$$

$$1 \le i \le |E[Session\_ID]|$$

where *E[k][i]* is the *i*th session within a single transaction *E[k]* and the *dt* is the dwell time of the session, one of the features or attributes of *E[k][i]* that can be obtained by calculating the time interval between two adjacent accesses from the same host.

But before that, however, we need to store this transaction data in a data structure such that flexible mining can be carried out on various features, as well as at different abstraction levels. In the following sections, we introduce the approach to this issue.

4. A Multi-Dimensional Data Structure

In order to accommodate the multi-feature transactional data transformed from the log data, a web log data cube is proposed here for further retrieval and mining activities. The generic cube model was proposed in [9], improved from the multidimensional database [14,1] to support the increasingly popular On-Line Analytical Processing (OLAP) standard. Our system uses a customized cube structure to meet the requirement of flexible mining on different features and abstract levels of the transaction data. Due to the space limit, we omit this section and refer interested readers to the extended version of this paper[21].

5. Mining approaches and the experiments

Up to this point, we have successfully reorganized the web access data to a state where it is convenient to apply classic data mining algorithms. Two mining approaches can be made on these transaction data, one is the adjusted association rule mining emphasizing the inter-relationship of the interested pages. We call the results interest context rules revealing interest-related sections in the site from a visitor's point of view; the other approach tends to discover the frequent sequence patterns of the user's browsing traces.

5.1 Finding interest context rule

Let T be a set of transactions and X,Y be the visitor's focuses which are the objects encapsulating web sections and its average interestingness (ai) to visitors, e.g., X is a set of [X.section, X.ai]. X.section,Y.section$\subset$ Dom(Features.section) are non-empty subsets of web sections, count(X) is the number of times that the transaction includes X. A context rule is an expression of the form where Max(X.timestamp)< Min(Y.timestamp), the support of the rule s=count(X.section$\cup$Y.section)/|T|, the confidence of the rule c= count(X.section$\cap$Y.section)/count(X.section). The task of discovering a context rule is to find all rules whose s and c are higher than their respective thresholds. We get a three dimensional cube denoted as C (TRANSACTION, SESSION, SECTION, interestingness), and by integrating with a step of computing the relative interestingness and the average value of them, the algorithm described in [13] can be easily adapted to search for the interest context rules in this web scenario.

Given a sample web log file of a commercial site and a time window of size 30 minutes, we list part of the interest context rule found with relatively high confidence.

$$\{[\text{/product/}, 0.27],[\text{/products/product1/}, 0.23]\} \xrightarrow{0.04\quad 0.80} \{\text{/products/product2/}, 0.24\}$$

$$\{[\text{/}, 0.17]\} \xrightarrow{0.11\quad 0.59} \{\text{/}\textit{news}\text{/}, 0.22\}$$

$$\{[\text{/news/on-Sale/}, 0.17],[\text{/products/product1.html}, 0.23]\} \xrightarrow{0.03\quad 0.65} \{\text{/products/product3.html/}, 0.02\}$$

$$\{[\text{/products/}, 0.21],[\text{/company/history/}, 0.33]\} \xrightarrow{0.01\quad 0.76} \{\text{/company/employment/}, 0.39\}$$

Figure 5.1 Sample Result of interest context Rules

These results improve results of other tools that provide only statistical summary about the access log. They are also different to classical association rules in that the latter do not have any ordering and any interestingness measurements which involve the item attribute into the process of knowledge discovery. Among these interest context rules, we can find interesting knowledge about the interaction between the site and its visitors. For example, the third interest context rule indicates that 65% of the visitors who browsed the on-sale news section and product1 section with relatively high interest also visited the product3 section later with average interest of 0.02, and this context rule applied in 3% of all the transactions. This rule revealed some hidden flaws in the site topology which shows that although visitors are interested in the on-sale news and the content of the product1, they have little interest in the product3 section. However they were cheated or forced into the product3 area by the links. This finding results in the decision to adjust the links among these sections and improve the relevance of the sections and effectiveness of the links. It also provides a perspective solution of dynamic organized web site based on user-interest-oriented web page clustering.

6. Summary and Future Work

In this paper, we have presented a framework for data mining and knowledge discovery on the HTTP server's log data. The aim of the research is to find a solution to model the log data and discover user navigation behavior in a specific web site. The knowledge is potentially useful to e-business and the efficient maintenance of the web site based on the preference of the visitors. To implement this, we designed a flexible log mining architecture and discussed the data flow from data preparation, transaction modeling, multi-dimensional storing structure with concept hierarchy to the mining approaches. According to the interaction between site visitors and the web log mechanism, we proposed a novel transaction model to interpret user access records as transactions and sessions based on the time window threshold and the users identification. Various features have been extracted from the log data including the modeling of visitor's interests in specific web sections as a relative interestingness feature. We then proposed the idea to integrate the interestingness feature with the sequence association rule finding algorithm to discover the interest context rules in the web site navigation scenario.

Although we have proposed several new ideas in the direction of knowledge discovery in web site log data, it is not enough to use only the web log data especially in the more and more complicated WWW community. For example, for an internet based supermarket, it will be more efficient to combine the user registration information and the ordering information with the log data. We plan to study how to integrate more heterogeneous data sources into the current framework. The other direction is to develop more generic models to cope with transactions of other application data such as bank transactions and DNA sequence data in biochemical areas.

Reference

[1] R. Agrawal, A. Gupta, and S. Sarawagi. Modeling Multi-dimensional Databases. In *Proc. of the 13th Int'l Conference on Data Engineerin*g, Birmingham, UK., April 1997

[2] R. Agrawal, T. Imielinski, and A. Swami. Mining Association Rules Between Sets of Items in Large Databases. In *Proc. 1993 ACM-SIGMOD Int. Conf. Management of Data*, page 207-216, Washington, D.C., May 1993

[3] R. Agrawal and R. Srikant. Fast Algorithms for Mining Association Rules. In *Proc. 1994 Int. Conf. Very Large Data Bases*, page 487-499, Santiago, Chile, September 1994

[4] R. Agrawal and R. Srikant. Mining Sequential Patterns. In *Proc. 1995 Int. Conf. Data Engineering*, pages 3-14, Taipei, Taiwan, March 1995

[5] A. G.Büchner, M. D. Mulvenna. Discovering Internet Marketing Intelligence though Online Analytical Web Usage Mining. *SIGMOD Record 27(4)* page 54-61, 1998

[6] E. Cohen and H. Kaplan Exploiting Regularities in Web Traffic Patterns for Cache Replacement. *31th STOC*, 1999

[7] R. Cooley, B. Mobasher and J. Srivastava. Data Preparation for Mining World Wide Web Browsing Patterns, *Journal of Knowledge and Information Systems*, Vol. 1, No. 1, 1999

[8] R. Fuller and J. de Graaff. Measuring User Motivation from Server Log Files. In *http://www.microsoft.com/usability/webconf/fuller/fuller.htm*, 1997

[9] J.Gray, A. Bosworth, A. Layman, and H. Pirahesh, Data cube: A Relational Aggregation Operator Generalizing Group-by, Cross-tabs and Sub-totals. In *Proc. of the 12th Int'l Conference on Data Engineering*, page 152-159,1996

[10] iLux Edge E-business company, *http://www.ilux.com/products/ent/index.html* ,1999

[11] A. Luotonen. The common log file format. *http://www.w3.org/pub/WWW/,* 1997

[12] Marketwave Corporation, *http://www.marketwave.com/hitlist/newreports/complete.htm,* 1998

[13] M. J. Zaki, Efficient Enumeration of Frequent Sequences, *7th International Conference on Information and Knowledge Management*, Washington DC, November 1998

[14] The OLAP Council. MD-API and OLAP Application Program Interface Version 0.5 Specification, September 1996

[15] M. Perkowitz and O. Etzioni. Adaptive site: Automatically Learning from User Accesses Patterns. In *Proc. 6th Int. World Wide Web Conf., Santa California*, April 1997

[16] T. Sullivan. Reading Reader Reaction: A Proposal for Inferential Analysis of Web Server Log Files. In *Proc. 3rd Conf. Human Factors & the Web*, Denver, Colorado, June 1997

[17] S. Sarawagi, R. Agrawal, N. Megiddo. Discovery-driven Exploration of OLAP Data Cubes, *Proc. of the Sixth Int'l Conference on Extending Database Technology,* Valencia, Spain, March 1998

[18] C. Silverstein, S. Brin, R. Motwani: Beyond Market Baskets: Generalizing Association Rules to Dependence Rules. *Data Mining and Knowledge Discovery 2(1)* page 39-68. 1998)

[19] L. Tauscher and S. Greenberg. How People Revisit Web Pages: Empirical Findings and Implications For the Design of History Systems. *International Journal of Human Computer Studies, Special issue on World Wide Web Usability, 47*. Page 97-138, 1997

[20] WebTrends Corporation *http://www.webtrends.com/SampleReports/Cluster_03_b.html*, 1999

[21] Feng T, Fionn M. Towoards Knowledge Discovery from WWW Log Data, Extended version http://www.qub.ac.uk/~F.Tao/research/extended.ps