A complete framework for Web mining.

Andy Tseng[1]
Ilias Petrounias[1]
Panagiotis Chountas[2]

[1] Department of Computation, UMIST
[2] Harrow School of Computer Science, University of Westminster

A Complete Framework for Web Mining*

Andy Tseng[1], Ilias Petrounias[2]
Department of Computation,
UMIST
PO Box 88, Manchester
M60 1QD, UK
[1]y.tseng@postgrad.umist.ac.uk, [2]ilias@co.umist.ac.uk

Panagiotis Chountas
Department of Computer Science,
University of Westminster
Northwick Watford Rd, Northwick Park,
London, HA1 3TP, UK
chountp@wmin.ac.uk

Abstract - *With the rapid growing number of WWW users, hidden information becomes ever increasingly valuable. As a consequence of this phenomenon, mining Web data and analysing on-line users' behaviour and their on-line traversal pattern have emerged as a new area of research. Primarily based on the Web servers' log files, the main objective of traversal pattern mining is to discover the frequent patterns in users' browsing paths and behaviours. This paper presents a complete framework for web mining, allowing users to pre-define physical constraints when analysing complex traversal patterns in order to improve the efficiency of algorithms and offer flexibility in producing the results.*

Keywords: Data mining, web mining, data management, information filtering, data pre-processing.

1 Introduction

With the increasing use of computing for various applications, the importance of data mining is growing with rapid pace recent times. Several data mining capabilities have been explored in the literature. One of the most common and important problem in the research of data mining is the efficiency and accuracy of the algorithm execution.

One of the main challenges for large corporations adopting World Wide Web sites is to discover and rediscover useful information from very rich but also diversified sources in the Web environment. Web log analysis is mainly used in this instance to determine key factors, such as interest in content and usage of Web sites. These become important inputs to design tasks and determine how a Web site is being used. Usage analysis includes straightforward statistics, such as page access frequency, as well as more sophisticated forms of analysis, such as finding the common traversal paths through a Web site. However, most of the work in web mining has focused on web log analysis. Within web log analysis the main interests have been user and session identification and sequences of pages being accessed by users.

This paper presents a complete framework for web mining. Existing proposals in the literature are concerned only with the forward navigation within a web site. In order to filter out the redundant pattern from the log source, [2] introduced the concept of "Maximal Forward Reference or Path" (MFP) as a notion of a maximal forward moving motion in visiting Web documents. The authors assumed that all the backward traversal actions (i.e. Backward Reference) only occur to users in the process of searching for Web pages that really interest them. Hence they assumed that only the forward browsing motion (Forward Reference) contains meaningful information and reflects users' true browsing patterns.

The work in this paper argues that the notion of a "Minimum Backward Path" (MBP) needs to be included since it also provides information about users' navigational patterns and their ability (or not) of navigating easily within a web site. This will demonstrate whether there exists a frequent short backward motion which may show that the structure of a web site is not clear. In addition to this, another important characteristic that is addressed is the notion of time. Within this, one can identify the longest time periods within which frequencies of pages occur and also the periodicity with which these web pages are accessed. The framework also addresses several 'constraint-based' pre-processing mining tasks to be performed prior to applying data mining algorithms to data collected from server logs. These constraints are taken from standard Web log files and are categorised into three main groups based on their nature and relation to user's on-site browsing behaviours:

- *"Traversal Constraints"* which concentrate on factors relating to users' navigating movements. A new method of 'Minimum Backward Path' (MBP) is defined to further reduce less meaningful traversal patterns and it successfully cooperates with the existing method of Maximum Forward Path (MFP) proposed in [2].

* 0-7803-7952-7/03/$17.00 © 2003 IEEE.

- *"Temporal Constraints"* include elements of 'Time', 'Session' and 'Periodicity'. These constraint elements concern factors such as duration of staying on a particular web page, session intervals and periodicity of visits to web pages.

- *"Personal Constraints"*, consist of other available information regarding each individual visiting a web site. For example, the IP address, demographical data and relevant topics, and is recognised as the subset of element 'User'.

Data mining algorithms that incorporate the above set of "Objective Constraints" are an attempt to resolve the shortcomings of existing approaches by introducing more relevant information (MBP, longest interval, periodicity of visits). By applying conditional restrictions with specific patterns, the proposed approach enables data analysts to focus on individual cases with more control while at the same time providing more knowledge about users' patterns. The outcome of any web mining algorithm is then influenced by those conditions. The value of conditional restrictions can be anything within users' traversal patterns, e.g. the length of the traversal movement, the direction of browsing path, designating nodes inside the browsing pattern, etc.

This framework is developed to support and assist existing data mining algorithms in order to first refine browsing pattern with relevant constraints and then aid the discovery tasks in both intra and inter-sessional information retrieval. With such a framework implemented, information retrieval and pattern identification is significantly faster and more accurate than just using standard discovery methods.

2 Infrastructure of the Framework

In order to filter out the redundant pattern from the log source, Chen et al [2] introduced the concept of "Maximal Forward Reference" as a notion of a maximal forward moving motion in visiting Web documents. They assumed that all the backward traversal actions (i.e. Backward Reference) only occur to users in the process of searching for Web pages that really interest them. Hence, they assumed that only the forward browsing motion (Forward Reference) is reflecting users' true browsing patterns and contains the meaningful inflation. For instance, if a user has the following traversal pattern inside a particular Web site:

$$\{ABCDCBEGHGWAOUOV\} \quad (1)$$

By using traditional analysis methods, nodes B and C are showing greater importance than nodes D and E, where it may in fact be that nodes D and E are actually the pages containing the information that the user needs. Nodes B and C might be pages embedded with all the inter-links in that site and as a result, cause an illusion in becoming the most valuable pages. When the "Maximal Forward Reference" method has been taken into consideration, the original traversal pattern will be translated into a new set of patterns as:

$$\{(ABCD)|(ABCDEGH)|(ABEGW)|(AOU)|(AOV)\} \quad (2)$$

M.F.R. successfully redefines the traversal data into a more meaningful manner by ignoring the continuous repetition of backward browsing actions.

In the "Maximal Forward Reference" [2] [4], one considers users' onward browsing flow as the only mean for measuring users' browsing behaviours and completely ignores the backward browsing paths. However, on-line browsing movement is not a simple single-directional action, but rather a "dual-directional" action. Although the conversing direction of the traversal paths only exist because of users' convenience, if it is paired with the result of the onward path analysis it offers better insights into users' actual travelling intentions. For instance, the Minimum Backward Path (BMP) demonstrates groups of nodes in the shortest-length combination. This presents a good indication of how well the infrastructure of a site is constructed and arranged. The longer the combination of nodes MBP holds the less organised a site appears to be.

This can be interpreted as users having difficulties in finding their desired nodes and hence they are forced to browse each link one after another in order to narrow down the possibilities. If the MBP contains many same combinations then this can inform the Webmaster that this particular reference of linkages is well constructed.

This paper proposes a new approach for data processing by adapting a constraint-based technique. These constraints are based on users' on-site browsing behaviours, for instance the maximum forward-browsed nodes (MFP) and minimum visited nodes in the reverse direction (MBP).

Furthermore the duration users have taken in visiting a site and their demographical records such as IP address and the time interval (and also periodicity) during which they access the site are also being used as factors in deciding the mining section of raw data. These records hold valuable information that can determine specific requirements and further focus on particular data sectors in order to obtain a refined data analysis.

'Objective constraints' apply additional restrictions onto the existing traversal data. The new inclusion allows data analysts to apply personalised conditional restrictions, e.g. to include some sort of path sequence or designated starting and finishing nodes. As a result, it allows execution of algorithms to improve the speed of generating candidate

sets in order to produce a more efficient analysis that suits individual needs.

As indicated earlier the essential part of this proposal is to introduce the constraints that are able to refine the data source in order to lessen the processing time with a better return of outcome. These objective constraints are introduced as below and are classified into three main categories: Traversal, Temporal and Personal. The overall framework is presented in Figure 1.

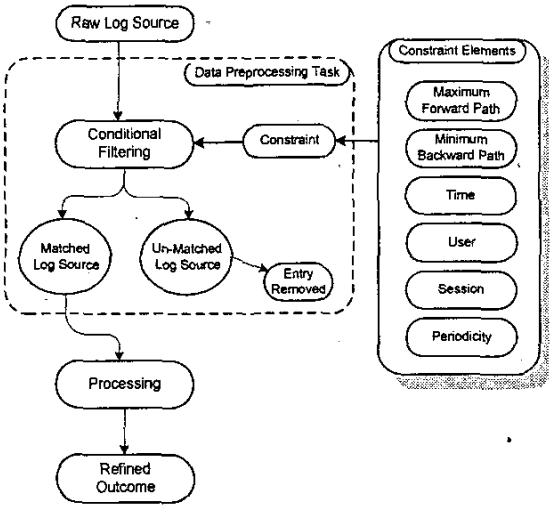

Figure 1: Proposed Data Processing Framework

2.1 Traversal Constraints

MFP and MBP are the main elements in this category. Each denotes the significance pattern of forward or backward directions of a user.

2.1.1 MFP (Maximum Forward Path)

Definition of MFP: Given a set of inter-linked nodes arranged in a hierarchy fashion, the action starts from the highest node (the root node) and follows its way down. When the first reverse movement occurs the forward movement is terminated. This results in a collection of nodes which is marked as maximum forward path [2].

For instance, taking the following as a template, the whole traversal pattern from the root node (i.e. node A) to the node R is shown as:

$$\{ABDGDBEHJMQSQMJNRTVXVTR\} \quad (3)$$

According to the definition given above, the maximum forward path for this instance will be extracted as below:

$$\{ABDGDBEHJMQSQMJNRTVXVTR\} \Downarrow \begin{Bmatrix} (ABDG) \\ (ABDGEHJMQS) \\ (ABDGEHJMQSNRTVX) \end{Bmatrix} \quad (4)$$

This presents that the MFP is confirmed on the nodes G, S and X where reverse movement starts taking place. Hence travelling from node A to R produces three maximum forward paths listed above. Since MFP omits all the reversing directional travelling, it will contain purely the nodes captured during the forward visits.

2.1.2 MBP (Minimum Backward Path)

Definition of MBP: In a set of hierarchal inter-linked nodes and during a particular session in time, the MBP starts at a node when a reverse behaviour occurs and returns back to the node where a new forward movement was invoked.

Minimum Backward Path is "not" necessary the reverse order of a maximum forward path. Again using (3) as an example, the MBP for travelling from node A to R, is listed as follows:

$$\{ABDGDBEHJMQSQMJNRTVXVTR\} \Downarrow \begin{Bmatrix} (BDG) \\ (JMQS) \\ (RTVX) \end{Bmatrix} \quad (5)$$

The difference between the two sets is noticeable after comparing (4) and (5). MBP contains the nodes covered by bidirectional movements (that is both forward and backward travelling), whereas MFP contains only single directional nodes.

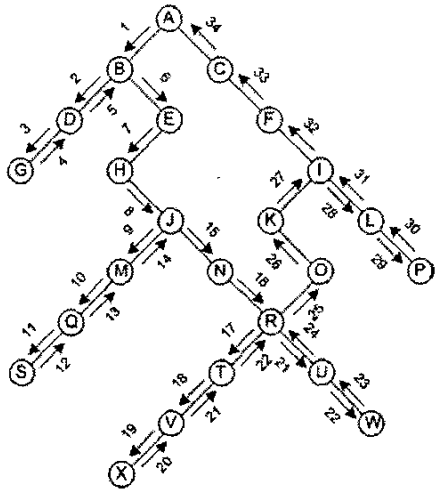

Figure 2: An Simulative Traversal Patterns for the W.W.W.

2.2 Temporal Constraints

2.2.1 Time

Definition of Time: Indefinite continued progress of existence, events, etc., in the past, present, and future, regarded as a whole (taken from Oxford Dictionary).

A time domain is a pair where a non-empty is set of chronons and is total order on [3]. As it can be seen from above definition, time is constructed by a set of chronons arranged in a total order manner.

2.2.2 Session

Definition of Session: Given a time-stamped starting point SS, on a particular visitor V, the session time ST remains until a visitor's onsite presence disappears at ending point SE.

A session is the time presence of a completed visit of a user with a specific IP address. This is normally achieved by setting a transient cookie. Transient cookies are only stored in temporary memory and are erased when the browser is closed (unlike persistent cookies which are stored in the user's hard disk and only removed when past the expiration date or deleted by the user manually).

2.2.3 Periodicity

Definition of Periodicity: A time consisting of a series of periodic intervals based on a time cycle unit.

Each periodic interval appears within an interval of the cycle unit and all these periodic intervals have the same position in their correspondent cycles.

Nevertheless it is important to distinguish between the concepts of "Session" and "Periodicity". A session, as described above, is created based on individual temporal attributes, such as the beginning and end of access time to a site. On the other hand, periodicity represents general time intervals during a bounded period over the time domain. Hence it is possible for multi sessions to occur over the same periodic time. As demonstrated in figure 2, a user can visit two or more sites simultaneously at the same or different time intervals.

2.3 Personal Constraints

2.3.1 User

Definition of User: A user is determined by a specific IP (Internet Protocol) address, which is assigned to each individual access when connected to the Internet.

This IP address is usually unique apart from the case of using a shared proxy server, in which case all users will be counted as using the same IP. Every web request made will then be processed and recorded according to their IP.

The IP address usually presents a good geographical indication of a user's origin. It also suggests the information to determine certain place or region with the characteristic qualities of users that originates therein. It is therefore important that the data analyst derives and identifies each user group's qualities and relation from that place. Since those qualities depend on different places, a specific "connection" might exist between the "products" of the site and the user group that may be of interest.

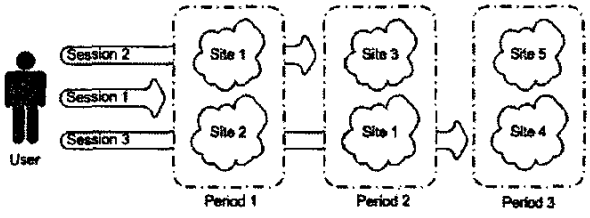

Figure 4: Session and Periodicity

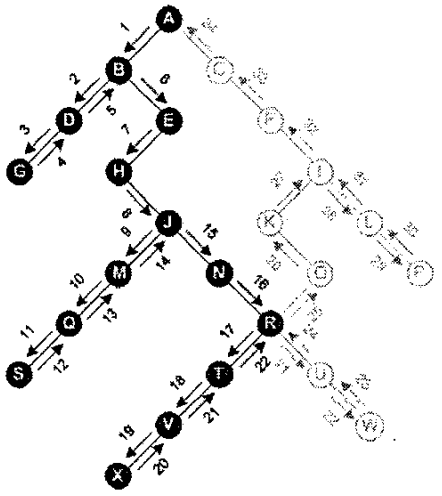

Figure 3: M.F.P. Illustration

3 Scheme of Data Preparation

The scheme is then formed based on two main parts, that is, primary and secondary entities. The primary entity deals with the type of data mining algorithm to be executed for discovering or rediscovering data patterns from the post-selected data source. The secondary entity which is based on the constraints mentioned in the previous section, on the other hand, prunes the pre-selected data source, refines the outcome and passes i8nformation back to the primary entity.

The structure and commands of this data preparation scheme can be described in a specialised syntax manner similar to the SQL (Structured Query Language) as Figure 5 demonstrates below:

```
SELECT   Mining_Rule_1 ( < rule_condition_1 > )
  WHERE  Cons_Type_1 (<type_condition_1> )
  WHERE  Cons_Type_2 (<type_condition_2> )
                 ⋮
  WHERE  Cons_Type_n (<type_condition_n>)
SELECT   Mining_Rule_2 (<rule_condition_2> )
                 ⋮
SELECT   Mining_Rule_n (<rule_condition_n> )
IN     ( < data_mining_algorithm > )
```

Figure 5: Mining Task

The constraints can be applied interchangeably and can be modified in accordance to analysts' preferences.

3.1 Mining_Rule(<rule_condition>)

It denotes the type of data mining technique to be used in the task for seeking a particular pattern. The task can be implemented with popular mining techniques such as Associations Rule, Classification, Clustering, Summarisation etc. Each technique has its own merits towards addressing different problems and this scheme provides the flexibility of adopting different techniques with customised conditions.

As each of the algorithms can be used to seek for potential patterns, it is possible to define specific variables in order to limit the range of the discovery process. This can be set to each algorithm accordingly.

3.2 Cons_Type(<type_condition>)

It is composed of elements discussed in the previous section. Constraints are inserted into an ordered list, where the position of a constraint's type determines the priority of the execution. Several constraints can be combined to form a complex constraint's type for setting a more detailed filter. This secondary entity is essential to this proposed framework as it is here that the selected data is being analysed and processed before any mining algorithms.

For instance, the case of attempting to find frequent itemsets using Association Rules (based on, for example, the 'Apriori' algorithm [1]) to find users' online MFP patterns (A, B, C and D) that have a period of over 2 hours between every Friday and Saturday afternoons after 15:30 hours and with the IP address ranging from 192.168.0.1 to 192.168.0.255, can be expressed as shown in Figure 6.

After execution of the mining task, the data will be pruned accordingly with the constraint values and then passed on to the selected data mining algorithm.

```
SELECT   Mining_Rule ("association_rule")
 WHERE  MFP ("threshold = '5 :: {A→B→C→D}' ")
  AND  USER ("ip = '192.168.0.1' TO '192.168.0.255' ")
  AND  TIME ("duration_hr = '2' ")
  AND  PERIODICITY ("str_day = 'friday'
                     AND end_day = 'saturday'
                     AND str_time = '1500'
                     AND end_time = '1530' ")
IN     ("apriori_gen( )")
```

Figure 6: Example of Mining Task

4 Conclusions

This paper has presented the details of a framework which consists of pre-processing tasks that are essential for applications performing various tasks of knowledge discovery such as data mining, web mining of content and usage as well as applications using data mining techniques to process web server access logs.

During this paper, a constraint-based scheme for the mining tasks was presented for the purpose of illustrating the possible compatibility when integrated with other applications or algorithms (Section 3). The proposed approach performs independently prior to the main algorithm thus being capable of handling an overwhelmingly large set of data and its customisable constraint-based syntax increases the efficiency and accuracy of algorithms.

Future work will include further tests to verify the model traversal of users' online browsing behaviours discussed in (Section 2.1) and a more rigorous analysis of the temporal constraints concerning periodicity aspects.

References

[1] Agrawal, R. and Srikant R., "Fast Algorithms for Mining Association Rules", Proc. of the 20th Int'l Conference on Very Large Databases, Santiago, Chile, p.487-499, 1994.

[2] Chen, M.S., Park, J.S., and Yu, P.S. "Data Mining for path traversal patterns in a Web environment", Proc. 16th Int. Conf. on Distributed Computing Systems, p.385-392, 1996.

[3] Etzion, O., Jajodia, S., and Sripada, S., "Temporal Database: Research and Practice", Springer-Verlag, Berlin, Germany, 1998.

[4] Park, J.S., Chen, M.S., and Yu, P.S., "An Efficient Hash-Based Algorithm for Mining Association Rules", pp. 175-186, Proceedings of SIGMOD 1995.