RESEARCH ARTICLE OPEN ACCESS

Strategies for Mining Web Logs: A Survey

Ms. Neha G. Sharma
ME in Computer Science, ECOET, Aurangabad

ABSTRACT
With the expansion of e-commerce and mobile-based commerce, the role of web user on World Wide Web has become pivotal enough to warrant studies to further understand the user's intent, navigation patterns on websites and usage needs. Using web logs on the servers hosting websites, site owners and in turn companies can extract information to better understand and predict user's needs, tailoring their sites to meet such needs. Through this paper, an effort is being made to analyze and survey few of the popular web log mining strategies.
Index Terms – web usage mining, web access pattern, WAP, path traversal, TSP.

I. INTRODUCTION

Initially, the World Wide Web and in-turn, the websites were only created with site owners interest; users' perspective and needs were not considered. Due to constant evolution of commerce from B2B to B2C, users became more important and centrifugal than companies serving or hosting the website. This relative shift from owner-centric to user-centric design has played an important role in improving the access efficiency of web pages by adaptive website system, dynamic re-organization of website, identification of target group of visitors, improving the performance of web search and prediction of user intent in web systems. Hence, the various techniques/strategies of web usage mining were developed to better understand user's intent, preference and interests.

We begin with a discussion on different sources of web usage data obtained from user's web navigation pattern (Section 2). Next, we focus on pre-processing techniques that are applied for filtering web logs (Section 3). Section 4 discusses the basic techniques used to discover web navigation pattern. In Section 5, various algorithms used for obtaining web navigation patterns are surveyed. Privacy issues are discussed in Section 6. Finally, Section 7 presents future research in this field. Section 8 gives references.

II. DATA SOURCES

There are three main sources for web usage mining applications. [2] They are 1. Web Server 2. Proxy Server and 3. Web clients.

2.1 Web server

They are the most common and richest source that contains logs in Common Log Format (CLF) [24], Extended Log Format (ELF) [25] or Log ML[24]. The major issue with the logging information obtained from a web server is identification of a user session. This depends on the type of information stored in log files. A common approach to resolve this is by the use of cookies or packet sniffers.[26,27]

2.2 Proxy server

Proxy servers collect data of group of users accessing a large number of web servers. Session identification is still a difficult task and does not identify the navigation paths of all users because of caching problem.

2.3 Web clients

Usage data can be tracked on the client-side using JavaScript, Java applet or a modified browser. This resolves the problem of session identification and caching but relies heavily on users' co-operation and may raise privacy issues.

III. PRE-PROCRSSING

Pre-processing web logs is a difficult and time-consuming process which comprises of four phases – data cleaning, identification and re-construction of user sessions, content and structure retrieval and, data formatting.

3.1 Data cleaning

This is a process of removal of useless data. Example, request for graphical page content (jpg and gif images), navigation sessions performed by robots and web spiders. Request for graphical contents and files are easy to eliminate while robots and web spiders can be identified by referring remote hostname using user agent or by checking robot.txt.

3.2 Identification and re-construction of user sessions

The problem caused for session identification and re-construction are caused by caching either by proxy servers or browsers. Proxy

caching problem can be resolved by the use of cookies, URL re-writing or by requiring user login.

3.3 Content and structure retrieval

Most of the web usage mining applications use visited URLs as a main source of information, but it does not convey information about the page content. So there was a need to employ content-based information along with a categorization step, wherein pages are classified according to their content type to enrich web log data. Later, concept-based paths were used in which common concepts are extracted by means of intersection of raw user paths and similarity measures.

3.4 Data formatting

This is final step of pre-processing. The extracted data from web logs can be stored into a relational database using a click-fact-schema. Signature tree to index log stored in databases were used for efficient pattern queries. A tree structure called "WAP-tree and FBP-tree" was introduced for sequence pattern discovery.

IV. BASIC TECHNIQUES USED TO DISCOVER WEB NAVIGATION PATTERN

Most of the commercial applications of Web Usage Mining exploit consolidated statistical analysis techniques. In contrast, research in this area is mainly focused on the development of knowledge discovery techniques specifically designed for the analysis of Web usage data. Most of this research effort focuses on three main paradigms: association rules, sequential patterns, and clustering.

4.1 Association rules

These are probably the most elementary data mining technique and, at the same time, the most used technique in Web Usage Mining. Association rules are implications of the form X→ Y where the rule body X and the rule head Y are set of items within a set of transactions. The rule X→Y states that the transactions which contain the items in X are likely to contain also the items in Y. When applied to Web Usage Mining, association rules are used to find associations among Web pages that frequently appear together in users_ sessions. The typical result has the form:

"A.html, B.html → C.html"

which states that if a user has visited page A.html and page B.html, it is very likely that in the same session the same user has also visited page C.html.

4.2 Sequential patterns

These patterns are used to discover frequent sub-sequences among large amount of sequential data. In Web Usage Mining, sequential patterns are exploited to find sequential navigation patterns that appear in users sessions frequently. The typical sequential pattern has the following form [23]: the 70% of users who first visited A.html and then visited B.html afterwards, have also accessed page C.html in the same session. Sequential patterns might appear syntactically similar to association rules; in fact algorithms to extract association rules can also be used for sequential pattern mining. However, sequential patterns include the notion of time, i.e., at which point of the sequence a certain event happened. In the above example, pages A, B, and C appears sequentially, one after another, in the user sessions; in the previous example on association rules, information about the event sequence is not considered.

There are essentially two class of algorithms that are used to extract sequential patterns: one includes methods based on association rule mining; the other one includes methods based on the use of tree structures and Markov chains to represent navigation patterns. Some well-known algorithms for mining association rules have been modified to extract sequential patterns. For instance, [12,13] used AprioriAll and GSP, two extensions of the Apriori algorithm for association rules mining [11]. Ref. [11] argues that algorithms for association rule mining (e.g., Apriori) are not efficient when applied to long sequential patterns, which is an important drawback when working with Web logs. Accordingly, [11] proposes an alternative algorithm in which tree structures (WAP-tree) are used to represent navigation patterns. The algorithm (WAP-mine) [11] and the data structure (WAP-tree), specifically tailored for mining Web access patterns, WAP-mine outperforms other Apriori-like algorithms [11] like GSP.

The GSP Algorithm makes multiple passes over data. The first pass determines the frequent 1-item patterns ($L1$). Each subsequent pass starts with a seed set: the frequent sequences found in the previous pass ($Lk-1$). The seed set is used to generate new potentially frequent sequences, called candidate sequences (Ck). Each candidate sequence has one more item than a seed sequence. In order to obtain *k*-sequence candidate Ck, the frequent sequence $Lk-1$ joins with itself Apriori-gen way. This requires that every sequence s in $Lk-1$ joins with other sequences s in $Lk-1$ if the last *k-2* elements of s are the same as the first *k-2* elements of s. For example, if frequent 3-sequence set $L3$ has 6 sequences as follows: *{*((1, 2) (3)), ((1, 2) (4)), ((1) (3, 4)), ((1, 3) (5)), ((2) (3, 4)), ((2) (3) (5))*}*. Inorder to obtain frequent 4-sequences, every frequent 3-sequence should join with the other 3-sequences that have the same first two elements as its last two elements. Sequence s = ((1, 2) (3)) can

join with s =((2) (3, 4)) to generate a candidate 4-sequence because the last 2 elements of s, (2) (3), are the same as the first 2 elements of s. Then, element (4) can be added to the sequence ((1, 2) (3)). Since element (4) is part of the last element (3, 4) of s, ((2) (3, 4)), the new sequence is ((1, 2) (3, 4)). Also, ((1, 2) (3)) can join with ((2) (3) (5)) to form ((1, 2) (3) (5)). The remaining sequences can not join with any sequence in $L3$. Following the join phase is the pruning phase, when the candidate sequences that have any of their contiguous $(k - 1)$-subsequences having a support count less than the minimum support, are dropped. The supports for the remaining candidate sequences are found next to determine which of the candidate sequences are actually frequent (Lk). These frequent candidates become the seed for the next pass. The algorithm terminates when there are no frequent sequences at the end of a pass, or when there are no candidate sequences generated. The GSP algorithm uses a hash tree to reduce the number of candidates that are checked for support in the database.

4.3 Clustering techniques

It looks for groups of similar items among large amount of data based on a general idea of distance function which computes the similarity between groups. Clustering has been widely used in Web Usage Mining to group together similar sessions.

V. ALGORITHM FOR MINING WEB NAVIGATION PATTERNS

5.1 PrefixSpan

As per [5], Prefixspan is a pattern-growth method like FreeSpan, which reduces the search space for extending already discovered prefix pattern p by projecting a portion of the original database that contains all necessary data for mining sequential patterns grown from p. While FreeSpan supports frequent pattern guided projection, PrefixSpan supports prefix guided projection. Thus, projected database for each f -list prefix pattern α consists of all sequences in the original database D, containing the pattern α and only the subsequences prefixed with the first occurrence of α are included. Although PrefixSpan projects smaller sized databases than FreeSpan, they both still incur non-trivial costs for constructing and storing these projected databases for every sequential pattern in the worst case.

5.1.1 Analysis of prefix span

No candidate sequence needs to be generated by PrefixSpan. Unlike a priori-like algorithms, Prefix-Span only grows longer sequential patterns from the shorter frequent ones. It neither generates nor tests any candidate sequence nonexistent in a projected database. Comparing with GSP, which generates and tests a substantial number of candidate sequences, PrefixSpan searches a much smaller space.

Projected databases keep shrinking. As a projected database is smaller than the original one because only the suffix subsequences of a frequent prefix are projected into a projected database. In practice, the shrinking factors can be significant because 1) usually, only a small set of sequential patterns grow quite long in a sequence database and, thus, the number of sequences in a projected database usually reduces substantially when prefix grows; and 2) projection only takes the suffix portion with respect to a prefix. Notice that FreeSpan also employs the idea of projected databases. However, the projection there often takes the whole string (not just suffix) and, thus, the shrinking factor is less than that of PrefixSpan.

The major cost of PrefixSpan is the construction of projected databases. In the worst case, PrefixSpan constructs a projected database for every sequential pattern. If there exist a good number of sequential patterns, the cost is nontrivial.

The above analysis shows that the major cost of PrefixSpan is database projection, i.e., forming projected databases recursively. Usually, a large number of projected databases will be generated in sequential pattern mining. If the number and/or the size of projected databases can be reduced, the performance of sequential pattern mining can be further improved.

One technique which may reduce the number and size of projected databases is pseudo projection. The idea is outlined as follows: Instead of performing physical projection, one can register the index (or identifier) of the corresponding sequence and the starting position of the projected suffix in the sequence. Then, a physical projection of a sequence is replaced by registering a sequence identifier and the projected position index point. Pseudo projection reduces the cost of projection substantially when the projected database can fit in main memory.

Optimization techniques include (1)bi-level projecting for reducing the number and sizes of projected databases, and (2) Pseudo-projection for projecting memory-only databases, where each projection consists of the pointer to the sequence and offset of the postfix to the sequence.

5.2 Apriori

5.2.1 The general structure of the algorithms

They make multiple passes over the data. In each pass, start with a seed set of large sequences. Then use the seed set for generating new potentially large sequences, called candidate sequences. Find the support for these candidate sequences during the pass

over the data. At the end of the pass, determine which of the candidate sequences are actually large. These large candidates become the seed for the next pass. In the first pass, all 1-sequences with minimum support, obtained in the litemset phase, form the seed set. Then present two families of algorithms, which we call count-all and count-some. The count-all algorithms count all the large sequences, including non-maximal sequences. The non-maximal sequences must then be pruned out (in the maximal phase). present one count-all algorithm, called AprioriAIl, based on the Apriori algorithm for finding large itemsets .Then present two count-some algorithms: Apriori-Some and Dynamicsome. The intuition behind these algorithms is that since they are only interested in maximal sequences, they avoid counting sequences which are contained in a longer sequence if we first count longer sequences. However, not to count a lot of longer sequences that do not have minimum support. Otherwise, the time saved by not counting sequences contained in a longer sequence may be less than the time wasted counting sequences without minimum support that would never have been counted because their subsequences were not large. Both the count-some algorithms have a forward phase, in which find all large sequences of certain lengths, followed by a backward phase, where find all remaining large sequences. The essential difference is in the procedure they use for generating candidate sequences during the forward phase. As AprioriSome generates candidates for a pass using only the large sequences found in the previous pass and then makes a pass over the data to find their support. Dynamicsome generates candidates onthe-fly using the large sequences found in the previous passes.

5.2.2 Drawbacks of Apriori:

The apriori-like sequential pattern mining method,though reducing search space, bears three nontrivial, inherent costs that are independent of detailed implementation techniques

A huge set of candidate sequences could be generated in a large sequence database. Since the set of candidate sequences includes all the possible permutations of the elements and repetition of items in a sequence, the apriori-based method may generate a really large set of candidate sequences even for a moderate seed set. For example, two frequent sequences of length-1, (a) and (b) will generate five candidate sequences of length-2: (aa), (ab), (ba), (bb), and ((ab)), where ((ab)) represents that two events, a and b, happen in the same time slot. If there are 1,000 frequent sequences of length-1, such as (a1); (a2); . . . ; (a1000), an a priori-like algorithm will generate [1000*1000+(1000*999)/2] = 1,499,500 candidate sequences. Notice that the cost of candidate sequence generation, test, and support counting is inherent to the a priori-based method, no matter what technique is applied to optimize its detailed implementation.

Multiple scans of databases in mining. The length of each candidate sequence grows by one at each database scan. In general, to find a sequential pattern of length l, the apriori-based method must scan the database at least l times. This bears a nontrivial cost when long patterns exist.

The apriori-based method generates a combinatorially explosive number of candidates when mining long sequential patterns. A long sequential pattern contains a combinatorial explosive number of subsequences, and such subsequences must be generated and tested in the a priori-based mining. Thus, the number of candidate sequences is exponential to the length of the sequential patterns to be mined. For example, let the database contain only one single sequence of length 100, (a1 a2 . . . a100), and the min_support threshold be 1 (i.e., every occurring pattern is frequent). To (re)derive this length-100 sequential pattern, the a priori-based method has to generate 100 length-1 candidate sequences (i.e., (a1),(a2); . . . ; (a100)),[100*100+(100*99)/2]=14,950 length-2 candidate sequences, ($^{100}C_3$)length-3 candidate sequences,1 and so on. Obviously, the total number of candidate sequences to be generated is
$\sum i=1 \text{ to} 100 \ (^{100}C_i) = 2^{100} - 1 \approx 10^{30}$

5.3 WAP-tree algorithm

In [1] proposed an algorithm using WAP-tree, which stands for web access pattern tree. This approach is quite different from the Apriori-like algorithms. The main steps involved in this technique are summarized next. The WAP-tree stores the web log data in a prefix tree format similar to the frequent pattern tree (FP-tree) for non-sequential data.

5.3.1 The general structure of the algorithms

1. The algorithm first scans the web log once to find all frequent individual events.
2. It scans the web log again to construct a WAP-tree over the set of frequent individual events of each transaction.
3. It finds the conditional suffix patterns.
4. In the fourth step, it constructs the intermediate conditional WAP-tree using the pattern found in previous step.
5. Finally, it goes back to repeat Steps 3 and 4 until the constructed conditional

WAP-tree has only one branch or is empty. Thus, with the WAP-tree algorithm, finding all frequent events in the web log entails constructing the WAP-tree and mining the access patterns from the WAP tree. WAP-tree algorithm scans the original database only twice and avoids the problem of

generating explosive candidate sets as in Apriori-like algorithms. Mining efficiency is improved sharply.

5.3.2 Drawback

WAP-tree mining degrades the performance because of re-construction associated WAP-trees which, cannot be held in main memory.

5.4 UAM algorithm

This algorithm uses user access matrix as a data structure and hence the name "user access matrix" algorithm. It obtains the user navigation pattern from the page-to-page transition probabilities and statistics of all user behaviors.[8] UAM is a URL-URL matrix set up from web logs according to the user navigation pattern where the referrer URL maps to the rows of the matrix whereas navigating URL is mapped as columns. It uses Depth-First Search (DFS) technique to obtain user-preferred navigation path.

5.4.1 Drawback

UAM does not always correctly discriminate different behavioral modes of different users.

5.5 PNT algorithm

It is used to identify user sessions from web logs.[8] PNT is a multi-player tree formed from user session database, where each node is a visited page, each branch is the way to the node along the same route. Each node of PNT records the support along with the number of times a user is visiting the node along the same route, for the entire duration of time the user is viewing the page and the preference which represents in what way user prefers a particular node from one of the previous nodes.

5.5.1 Drawback

This algorithm does not consider browsing actions like "copy", "scroll" or, "save as" as it is not recorded in log files.

5.6 IBD algorithm

It stands for "intentional browsing data" algorithm. IBD takes care of online browsing actions such as "copy", "scroll" or, "save as" that is not recorded in web logs.[21]

5.6.1 Drawback

It requires an online data collection mechanism for capturing IBD.

5.7 Incremental and interactive mining of web traversal patterns

The essence of incremental data mining and interactive data mining is the ability to use previous mining results in order to reduce unnecessary processes when web logs or web site structures are updated, or when the minimum support is changed.

As per [3], they propose two novel incremental web traversal pattern mining algorithms for the maintenance of web traversal patterns when a database is updated or a web site structure is changed. Also present an interactive web traversal pattern mining algorithm to find all web traversal patterns when min_sup is adjusted. This algorithm utilizes previous mining results to find new web traversal patterns such that the total mining time can be reduced.For that used the lattice structure to store the previous mining results for incremental Web traversal patterns. The patterns may be obtained rapidly when the database or the website structure is updated. The problem of choosing an appropriate storage structure to store previous mining results now becomes very important.

5.7.1 Drawback

The size of the lattice structure may become too large to be loaded into the main memory.

VI. PRIVACY

In view of user sensitive data, sourced from various data sources and used in web usage mining concerns are raised on user privacy. This has pushed countries and governments to enact strict laws about user privacy. Another push on this through research and advanced studies is in a form of proposal in world wide web called Platform for Privacy Preferences (P3P). This proposal is very generic and does not impose any specific requirements on the amount of data collected (user profiling) or user privacy information. Another proposal from researchers was to use user models which cannot be linked to individual user and can be used for web usage mining. However, further research and work can be conducted on resolving these privacy issues.

VII.FUTURE WORK

The former mining algorithms that are surveyed in this paper suffer from either repetitive database scans or high memory load. Algorithms that work with a single scan need special data structures to store patterns. However, it may be difficult to hold these patterns in the data structure.

Consequently, a graph traverse algorithm can be used to determine the Throughout Surfing Pattern (TSP) in order to track the intent of website visitors. This does not pose a load on memory as a hyperlink structure of website is stored instead of a sequence database. The algorithm builds on an efficient and effective way to understand the visitor's intent and their navigation path.

REFERENCES

[1] Ezeife, C. I., & Lu, Y. (2005). Mining Web log sequential patterns with position coded pre-order linked WAP-tree. Data Mining and Knowledge Discovery, 10,5–38.

[2] Facca, F. M., & Lanzi, P. L. (2005). Mining interesting knowledge from Weblogs: A survey. Data and Knowledge Engineering, 53, 225–241.

[3] Lee, Y.-S., & Yen, S.-J. (2007). Incremental and interactive mining of Web traversal patterns. Information Sciences, 178(2), 278–306.

[4] Li, H.-F., Lee, S.-Y., & Shan, M.-K. (2006). DSM-PLW: Single-pass mining of path traversal patterns over streaming Web click-sequences. Computer Networks, 50,1474–1487.

[5] Pei, J., Han, J., Mortazavi-Asl, B., Wang, J., Pinto, H., Chen, Q., et al. (2004). Mining sequential patterns by pattern-growth: The prefixspan approach. IEEE Transactions on Knowledge and Data Engineering, 16(10), 1–17.

[6] Tseng, V.-S., & Lin, K.-W. (2006). Efficient mining and prediction of user behavior patterns in mobile Web systems. Information and Software Technology, 48,357–369.

[7] Wang, J., Han, J., & Li, C. (2007). Frequent closed sequence mining without candidate maintenance. IEEE Transactions on Knowledge and Data Engineering, 19(8), 1042–1056.

[8] Xing, D., & Shen, J. (2004). Efficient data mining for Web navigation patterns.Information and Software Technology, 46, 55–63.

[9] Yan, X., Han, J., & Afshar, R. (2003). CloSpan: Mining closed sequential patterns in large datasets. In Third SIAM international conference on data mining (SDM), SanFrancisco. CA (pp. 166–177).

[10] Yen, S.-J., & Chen, A. L. P. (2001). A graph-based approach for discovering various types of association rules. IEEE Transactions on Knowledge and Data Engineering, 13(5), 839–845.

[11] J. Pei, J. Han, B. Mortazavi-asl, H. Zhu, Mining access patterns efficiently from web logs, in: Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2000, pp. 396–407.

[12] X. Huang, N. Cercone, A. An, Comparison of interestingness functions for learning web usage patterns, in:Proceedings of the Eleventh International Conference on Information and Knowledge Management, ACM Press,2002, pp. 617–620.\

[13] B. Mortazavi-Asl, Discovering and mining user web-page traversal patterns, Master_s thesis, Simon Fraser University, 2001

[14] E.S. Nan Niu, M. El-Ramly, Understanding web usage for dynamic web-site adaptation: A case study, in: Proceedings of the Fourth International Workshop on Web Site Evolution (WSE_02), IEEE, 2002, pp. 53–64.

[15] Yao-Te Wang, Anthony J.T. Lee "Mining Web navigation patterns with a path traversal graph" Expert Systems with Applications: An International Journal Volume 38 Issue 6, June, 2011

[16] Bo Wu, Defu Zhang, Qihua Lan, Jiemin Zheng An Efficient Frequent Patterns Mining Algorithm based on Apriori Algorithm and the FP-tree Structure Convergence and Hybrid Information Technology, 2008. ICCIT '08. Third International Conference on (Volume:1)

[17] Suchita A.Chavan " Different Approaches of Mining Web Navigation Pattern: Survey" International Journal of Computer Applications (0975 – 8887) International Conference on Recent Trends in engineering & Technology - 2013(ICRTET'2013)

[18] Rakesh Agrawal and Ramakrishnan Srikant "Mining Sequential Patterns" IBM research paper IEEE.

[19] Jian Pei, Jiawei Han, Behzad Mortazavi-Asl, Jianyong Wang, Helen Pinto, Qiming Chen, Umeshwar Dayal and Mei-Chun Hsu "Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach" IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 16, NO. 11, NOVEMBER 2004

[20] K. Balog P. Hofgesang W. Kowalczyk "Modeling Navigation Patterns of Visitors of Unstructured Websites" Research and Development in Intelligent Systems XXII Proceedings of AI-2005, the Twenty-fifth SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence, Cambridge, UK, December 2005

[21] Yu-Hui Tao , Tzung-Pei Hong , Yu-Ming Su "Web usage mining with intentional browsing data" Expert Systems with Applications: An International Journal Volume 34 Issue 3, April, 2008

[22] Huxing Zhang _, Gang Wu _, Kingsum Chow y, Zhidong Yu y, and XueZhi Xing "Detecting Resource Leaks through

Dynamical Mining of Resource Usage Patterns" DSNW '11 Proceedings of the 2011 IEEE/IFIP 41st International Conference on Dependable Systems and Networks Workshops.

[23] E.S. Nan Niu, M. El-Ramly, Understanding web usage for dynamic web-site adaptation: A case study, in: Proceedings of the Fourth International Workshop on Web Site Evolution (WSE_02), IEEE, 2002, pp. 53–64.

[24] Configuration file of W3C httpd, http://www.w3.org/Daemon/User/Config/ (1995).

[25] W3C Extended Log File Format, http://www.w3.org/TR/WD-logfile.html (1996).

[26] D.M. Kristol, Http cookies: standards, privacy, and politics, ACM Transactions on Internet Technology (TOIT) 1 (2) (2001) 151–198.

[27] Pilot Software, Web site analysis, Going Beyond Traffic Analysis http://www.marketwave.com/productssolutions/hitlist.html (2002).