Different Aspects of Web Log Mining

Renáta Iváncsy, István Vajk

Department of Automation and Applied Informatics,
and HAS-BUTE Control Research Group
Budapest University of Technology and Economics
Goldmann Gy. tér 3, H-1111 Budapest, Hungary
e-mail: {renata.ivancsy, vajk}@aut.bme.hu

Abstract: The expansion of the World Wide Web has resulted in a large amount of data that is now freely available for user access. The data have to be managed and organized in such a way that the user can access them efficiently. For this reason the application of data mining techniques on the Web is now the focus of an increasing number of researchers. One key issue is the investigation of user navigational behavior from different aspects. For this reason different types of data mining techniques can be applied on the log file collected on the servers. In this paper three of the most important approaches are introduced for web log mining. All the three methods are based on the frequent pattern mining approach. The three types of patterns that can be used for obtain useful information about the navigational behavior of the users are page set, page sequence and page graph mining.

Keywords: Pattern mining, Sequence mining, Graph Mining, Web log mining

1 Introduction

The expansion of the World Wide Web (Web for short) has resulted in a large amount of data that is now in general freely available for user access. The different types of data have to be managed and organized such that they can be accessed by different users efficiently. Therefore, the application of data mining techniques on the Web is now the focus of an increasing number of researchers. Several data mining methods are used to discover the hidden information in the Web. However, Web mining does not only mean applying data mining techniques to the data stored in the Web. The algorithms have to be modified such that they better suit the demands of the Web. New approaches should be used which better fit the properties of Web data. Furthermore, not only data mining algorithms, but also artificial intelligence, information retrieval and natural language processing techniques can be used efficiently. Thus, Web mining has been developed into an autonomous research area.

The focus of this paper is to provide an overview how to use frequent pattern mining techniques for discovering different types of patterns in a Web log database. The three patterns to be searched are frequent itemsets, sequences and tree patterns. For each of the problem an algorithm was developed in order to discover the patterns efficiently. The frequent itemsets (frequent page sets) are discovered using the ItemsetCode algorithm presented in [1]. The main advantage of the ItemsetCode algorithm is that it discovers the small frequent itemsets in a very quick way, thus the task of discovering the longer ones is enhanced as well. The algorithm that discovers the frequent pase sequences is called SM-Tree [2] and the algorithm that discovers the tree-like patters is called PD-Tree [3]. Both of the algorithms exploit the benefit of using automata theory approach for discovering the frequent patterns. The SM-Tree algorithm uses state machines for discovering the sequences, and the PD-Tree algorithm uses pushdown automatons for determining the support of the tree patterns in a tree database.

The organization of the paper is as follows. Section 2 introduces the basic tasks of Web mining. In Section 3 the Web usage mining is described in detail. The different tasks in the process of Web usage mining is depicted as well. Related Work can be found in Section 4 and the preprocessing steps are described in Section 5. The results of the mining process can be found in Section 6.

2 Web Mining Approaches

Web mining involves a wide range of applications that aims at discovering and extracting hidden information in data stored on the Web. Another important purpose of Web mining is to provide a mechanism to make the data access more efficiently and adequately. The third interesting approach is to discover the information which can be derived from the activities of users, which are stored in log files for example for predictive Web caching [4]. Thus, Web mining can be categorized into three different classes based on which part of the Web is to be mined [5,6,7]. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining. For detailed surveys of Web mining please refer to [5,6,8,9].

Web content mining [10,9] is the task of discovering useful information available on-line. There are different kinds of Web content which can provide useful information to users, for example multimedia data, structured (i.e. XML documents), semi-structured (i.e. HTML documents) and unstructured data (i.e. plain text). The aim of Web content mining is to provide an efficient mechanism to help the users to find the information they seek. Web content mining includes the task of organizing and clustering the documents and providing search engines for accessing the different documents by keywords, categories, contents etc.

Web structure mining [11,12,13,14] is the process of discovering the structure of hyperlinks within the Web. Practically, while Web content mining focuses on the inner-document information, Web structure mining discovers the link structures at the inter-document level. The aim is to identify the authoritative and the hub pages for a given subject. Authoritative pages contain useful information, and are supported by several links pointing to it, which means that these pages are highly-referenced. A page having a lot of referencing hyperlinks means that the content of the page is useful, preferable and maybe reliable. Hubs are Web pages containing many links to authoritative pages, thus they help in clustering the authorities. Web structure mining can be achieved only in a single portal or also on the whole Web. Mining the structure of the Web supports the task of Web content mining. Using the information about the structure of the Web, the document retrieval can be made more efficiently, and the reliability and relevance of the found documents can be greater. The graph structure of the web can be exploited by Web structure mining in order to improve the performance of the information retrieval and to improve classification of the documents.

Web usage mining is the task of discovering the activities of the users while they are browsing and navigating through the Web. The aim of understanding the navigation preferences of the visitors is to enhance the quality of electronic commerce services (e-commerce), to personalize the Web portals [15] or to improve the Web structure and Web server performance [16]. In this case, the mined data are the log files which can be seen as the secondary data on the web where the documents accessible through the Web are understood as primary data.

There are three types of log files that can be used for Web usage mining. Log files are stored on the server side, on the client side and on the proxy servers. By having more than one place for storing the information of navigation patterns of the users makes the mining process more difficult. Really reliable results could be obtained only if one has data from all three types of log file. The reason for this is that the server side does not contain records of those Web page accesses that are cached on the proxy servers or on the client side. Besides the log file on the server, that on the proxy server provides additional information. However, the page requests stored in the client side are missing. Yet, it is problematic to collect all the information from the client side. Thus, most of the algorithms work based only the server side data. Some commonly used data mining algorithms for Web usage mining are association rule mining, sequence mining and clustering [17].

3 Web Usage Mining

Web usage mining, from the data mining aspect, is the task of applying data mining techniques to discover usage patterns from Web data in order to understand and better serve the needs of users navigating on the Web [18]. As

every data mining task, the process of Web usage mining also consists of three main steps: (i) preprocessing, (ii) pattern discovery and (iii) pattern analysis.

In this work pattern discovery means applying the introduced frequent pattern discovery methods to the log data. For this reason the data have to be converted in the preprocessing phase such that the output of the conversion can be used as the input of the algorithms. Pattern analysis means understanding the results obtained by the algorithms and drawing conclusions.

Figure 1 shows the process of Web usage mining realized as a case study in this work. As can be seen, the input of the process is the log data. The data has to be preprocessed in order to have the appropriate input for the mining algorithms. The different methods need different input formats, thus the preprocessing phase can provide three types of output data.

The frequent patterns discovery phase needs only the Web pages visited by a given user. In this case the sequences of the pages are irrelevant. Also the duplicates of the same pages are omitted, and the pages are ordered in a predefined order.

In the case of sequence mining, however, the original ordering of the pages is also important, and if a page was visited more than once by a given user in a user-defined time interval, then it is relevant as well. For this reason the preprocessing module of the whole system provides the sequences of Web pages by users or user sessions.

For subtree mining not only the sequences are needed but also the structure of the web pages visited by a given user. In this case the backward navigations are omitted, only the forward navigations are relevant, which form a tree for each user. After the discovery has been achieved, the analysis of the patterns follows. The whole mining process is an iterative task which is depicted by the feedback in Figure 1. Depending on the results of the analysis either the parameters of the preprocessing step can be tuned (i.e. by choosing another time interval to determine the sessions of the users) or only the parameters of the mining algorithms. (In this case that means the minimum support threshold.)

In the case study presented in this work the aim of Web usage mining is to discover the frequent pages visited at the same time, and to discover the page sequences visited by users. The results obtained by the application can be used to form the structure of a portal satisfactorily for advertising reasons and to provide a more personalized Web portal.

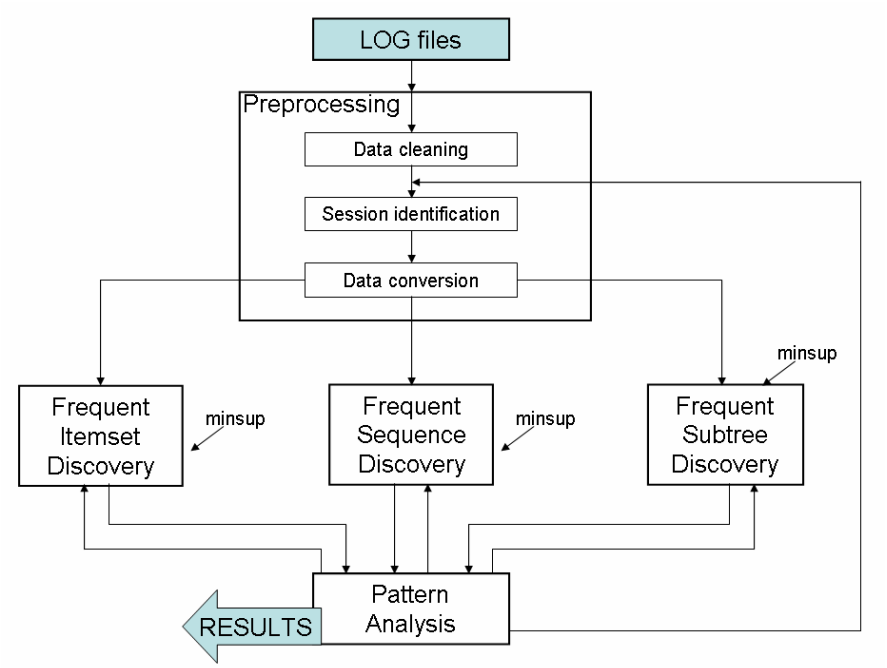

Figure 1
Process of Web usage mining

4 Related Work

In Web usage mining several data mining techniques can be used. Association rules are used in order to discover the pages which are visited together even if they are not directly connected, which can reveal associations between group of users with specific interest [15]. This information can be used for example for restructuring Web sites by adding links between those pages which are visited together. Association rules in Web logs are discovered in [19,20,21,22,23]. Sequence mining can be used for discover the Web pages which are accessed immediately after another. Using this knowledge the trends of the activity of the users can be determined and predictions to the next visited pages can be calculated. Sequence mining is accomplished in [16], where a so-called WAP-tree is used for storing the patterns efficiently. Tree-like topology patterns and frequent path traversals are searched by [19,24,25,26].

Web usage mining is elaborated in many aspects. Besides applying data mining techniques also other approaches are used for discovering information. For

example [7] uses probabilistic grammar-based approach, namely an Ngram model for capturing the user navigation behavior patterns. The Ngram model assumes that the last *N* pages browsed affect the probability of identifying the next page to be visited. [27] uses Probabilistic Latent Semantic Analysis (PLSA) to discover the navigation patterns. Using PLSA the hidden semantic relationships among users and between users and Web pages can be detected. In [28] Markov assumptions are used as the basis to mine the structure of browsing patterns. For Web prefetching [29] uses Web log mining techniques and [30] uses a Markov predictor.

5 Data Preprocessing

The data in the log files of the server about the actions of the users can not be used for mining purposes in the form as it is stored. For this reason a preprocessing step must be performed before the pattern discovering phase.

The preprocessing step contains three separate phases. Firstly, the collected data must be cleaned, which means that graphic and multimedia entries are removed. Secondly, the different sessions belonging to different users should be identified. A session is understood as a group of activities performed by a user when he is navigating through a given site. To identify the sessions from the raw data is a complex step, because the server logs do not always contain all the information needed. There are Web server logs that do not contain enough information to reconstruct the user sessions, in this case for example time-oriented heuristics can be used as described in [31]. After identifying the sessions, the Web page sequences are generated which task belongs to the first step of the preprocessing. The third step is to convert the data into the format needed by the mining algorithms. If the sessions and the sequences are identified, this step can be accomplished more easily.

In our experiments we used two web server log files, the first one was the *msnbc.com anonymous data*[1] and the second one was a Click Stream data downloaded from the *ECML/PKDD 2005 Discovery Challenge*[2]. Both of the log files are in different formats, thus different preprocessing steps were needed.

The msnbc log data describes the page visits of users who visited msnbc.com on September 28, 1999. Visits are recorded at the level of URL category and are recorded in time order. This means that in this case the first phase of the preprocessing step can be omitted. The data comes from Internet Information Server (IIS) logs for msnbc.com. Each row in the dataset corresponds to the page

1 http://kdd.ics.uci.edu/databases/msnbc/msnbc.html
2 http://lisp.vse.cz/challenge/CURRENT/

visits of a user within a twenty-four hour period. Each item of a row corresponds to a request of a user for a page. The pages are coded as shown in Table 1. The client-side cached data is not recorded, thus this data contains only the server-side log.

Table 1

Codes for the msnbc.com page categories

category	**code**	**category**	**code**	**category**	**code**
frontpage	1	misc	7	summary	13
news	2	weather	8	bbs	14
tech	3	health	9	travel	15
local	4	living	10	msn-news	16
opinion	5	business	11	msn-sport	17
On-air	6	sports	12		

In the case of the msnbc data only the rows have to be converted into itemsets, sequences and trees. The other preprocessing steps are done already. A row is converted into an itemset by omitting the duplicates of the pages, and sorting them regarding their codes. In this way the ItemsetCode algorithm can be executed easily on the dataset.

In order to have sequence patterns the row has to be converted such that they represent sequences. A row corresponds practically to a sequence having only one item in each itemset. Thus converting a row into the sequence format needed by the SM-Tree algorithm means to insert a -1 between each code.

In order to have the opportunity mining tree-like patterns the database has to be converted such that the transactions represent trees. For this reason each row is processed in the following way. The root of the tree is the first item of the row. From the subsequent items a branch is created until an item is reached which was already inserted into the tree. In this case the algorithm inserts as many -1 item into the string representation of the tree as the number of the items is between the new item and the previous occurrence of the same item. The further items form another branch in the tree. For example given the row: “1 2 3 4 2 5” then the tree representation of the row is the following: “1 2 3 4 -1 -1 5”.

In case of the Click Stream data, the preprocessing phase needs more work. It contains 546 files where each file contains the information collected during one hour from the activities of the users in a Web store. Each row of the log contains the following parts:

- a shop identifier
- time
- IP address

- automatic created unique session identifier
- visited page
- referrer

In Figure 2 a part of the raw log file can be observed. Because in this case the sessions have already been identified in the log file, the Web page sequences for the same sessions have to be collected only in the preprocessing step. This can be done in the different files separately, or through all the log files. After the sequences are discovered, the different web pages are coded, and similarly to the msnbc data, the log file has to be converted into itemsets and sequences.

Figure 2
An example of raw log file

6 Data Mining and Pattern Analysis

As it is depicted in Figure 1, the Web usage mining system is able to use all three frequent pattern discovery tasks described in this work. For the mining process, besides the input data, the minimum support threshold value is needed. It is one of the key issues, to which value the support threshold should be set. The right answer can be given only with the user interactions and many iterations until the appropriate values have been found. For this reason, namely, that the interaction of the users is needed in this phase of the mining process, it is advisable executing the frequent pattern discovery algorithm iteratively on a relatively small part of the whole dataset only. Choosing the right size of the sample data, the response time of the application remains small, while the sample data represents the whole data accurately. Setting the minimum support threshold parameter is not a trivial task, and it requires a lot of practice and attention on the part of the user.

The frequent itemset discovery and the association rule mining was accomplished using the ItemsetCode algorithm with different minimum support and minimum confidence threshold values. Figure 3 (a) depicts the association rules generated from msnbc.com data at a minimum support threshold of 0.1% and at a minimum

confidence threshold of 85% (which is depicted in the figure). Analyzing the results, one can make the advertising process more successful and the structure of the portal can be changed such that the pages contained by the rules are accessible from each other.

Another type of decision can be made based on the information gained from a sequence mining algorithm. Figure 3 (b) shows a part of the discovered sequences of the SM-Tree algorithm from the msnbc.com data. The percentage values depicted in Figure 3 (b) are the support of the sequences.

The frequent tree mining task was accomplished using the PD-Tree algorithm. A part of the result of the tree mining algorithm is depicted in Figure 4 (a). The patterns contain beside the trees (represented in string format), also the support values. The graphical representations of the patterns are depicted in Figure 4 (b) without the support values.

opinion & misc & travel	-->	on-air	90.26%
news & misc & business & bbs	-->	frontpage	90.24%
living & business & sports & bbs	-->	frontpage	90.00%
news & misc & business & sports	-->	frontpage	89.68%
news & tech & living & business & sports	-->	frontpage	89.00%
news & living & business & bbs	-->	frontpage	88.01%
frontpage & tech & living & business & sports	-->	news	87.87%
frontpage & opinion & living & sports	-->	news	87.81%
frontpage & tech & opinion & living	-->	news	87.60%
frontpage & tech & on-air & business & sports	-->	news	87.59%
news & misc & sports & bbs	-->	frontpage	87.55%
news & tech & on-air & business & sports	-->	frontpage	87.43%
news & living & business & sports	-->	frontpage	87.18%
news & business & sports & bbs	-->	frontpage	86.70%
misc & living & travel	-->	on-air	86.55%
tech & living & sports & bbs	-->	frontpage	86.52%
tech & business & sports & bbs	-->	frontpage	86.40%
news & misc & living & business	-->	frontpage	86.22%
on-air & business & sports & bbs	-->	frontpage	86.22%
news & tech & misc & bbs	-->	frontpage	86.18%
on-air & misc & business & sports	-->	frontpage	86.16%
tech & misc & travel	-->	on-air	86.09%
tech & living & business & sports	-->	frontpage	86.08%
news & living & sports & bbs	-->	frontpage	85.99%
misc & business & sports	-->	frontpage	85.79%
frontpage & tech & opinion & sports	-->	news	85.78%
news & opinion & living & sports	-->	frontpage	85.69%
misc & business & travel	-->	on-air	85.65%
news & tech & misc & business	-->	frontpage	85.63%
misc & business & bbs	-->	frontpage	85.57%
tech & living & sports & bbs	-->	news	85.49%
local & misc & business & sports	-->	frontpage	85.43%
news & opinion & business & bbs	-->	frontpage	85.32%
news & misc & living & sports	-->	frontpage	85.19%
news & on-air & business & sports	-->	frontpage	85.01%

(a)

misc → local	2.07%
frontpage → frontpage → sports	2.02%
local → frontpage	1.83%
on-air → misc → on-air	1.72%
on-air → frontpage	1.69%
on-air → news	1.51%
news → frontpage → news	1.49%
local → news	1.46%
frontpage → frontpage → business	1.35%
news → sports	1.33%
news → bbs	1.23%
health → local	1.16%
misc → frontpage → frontpage	1.16%
on-air → local	1.15%
misc → on-air → misc	1.15%
frontpage → frontpage → living	1.14%
local → frontpage → frontpage	1.13%
health → misc	1.12%
misc → on-air → on-air	1.10%
local → misc → local	1.09%
misc → news	1.06%
news → living	1.06%
on-air → misc → on-air → misc	1.00%

(b)

Figure 3

(a) Association rules and (b) sequential rules based on the msnbc.data

summary frontpage bbs	0.15%
health news misc	0.15%
on-air misc – business	0.15%
news sports living	0.15%
frontpage travel – local	0.15%
health local – on-air	0.15%
frontpage tech – news – tech	0.15%
frontpage tech – sports – news	0.15%
frontpage misc – news misc	0.15%
frontpage sports – news – business	0.10%
news travel – on-air misc	0.03%
tech living opinion weather	0.03%
frontpage news – misc – travel – news	0.03%
frontpage news – living – tech – local	0.03%
frontpage tech – living – news – living	0.03%
frontpage sports – news – sports bbs	0.03%

(a) (b)

Figure 4

Frequent tree patterns based on msnbc.data in (a) string and (b) graphical representation

Conclusions

This paper deals with the problem of discovering hidden information from large amount of Web log data collected by web servers. The contribution of the paper is to introduce the process of web log mining, and to show how frequent pattern discovery tasks can be applied on the web log data in order to obtain useful information about the user's navigation behavior.

Acknowledgement

This work has been supported by the Mobile Innovation Center, Hungary, by the fund of the Hungarian Academy of Sciences for control research and the Hungarian National Research Fund (grant number: T042741).

References

[1] R. Iváncsy and I. Vajk, "Time- and Memory-Efficient Frequent Itemset Discovering Algorithm for Association Rule Mining." *International Journal of Computer Applications in Techology, Special Issue on Data Mining Applications (in press)*

[2] R. Iváncsy and I. Vajk, "Efficient Sequential Pattern Mining Algorithms." WSEAS Transactions on Computers, vol. 4, num. 2, 2005, pp. 96-101

[3] R. Iváncsy and I. Vajk, "PD-Tree: A New Approach to Subtree Discovery.", WSEAS TRansactions on Information Science and Applications, vol. 2, num. 11, 2005, pp. 1772-1779

[4] Q. Yang and H. H. Zhang, "Web-log mining for predictive web caching." *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 1050-1053, 2003

[5] Kosala and Blockeel, "Web mining research: A survey," *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM*, vol. 2, 2000

[6] S. K. Madria, S. S. Bhowmick, W. K. Ng, and E.-P. Lim, "Research issues in web data mining," in *Data Warehousing and Knowledge Discovery*, 1999, pp. 303-312

[7] J. Borges and M. Levene, "Data mining of user navigation patterns," in WEBKDD, 1999, pp. 92-111

[8] M. N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, "Data mining and the web: Past, present and future," in *ACM CIKM'99 2nd Workshop on Web* Information *and Data Management (WIDM'99), Kansas City, Missouri, USA, November 5-6, 1999*, C. Shahabi, Ed. ACM, 1999, pp. 43-47

[9] S. Chakrabarti, "Data mining for hypertext: A tutorial survey." *SIGKDD: SIGKDD Explorations: Newsletter of the Special Interest Group (SIG) on* Knowledge *Discovery and Data Mining, ACM*, vol. 1, no. 2, pp. 1-11, 2000

[10] M. Balabanovic and Y. Shoham, "Learning information retrieval agents: Experiments with automated web browsing," in *Proceedings of the AAAI Spring Symposium on Information Gathering from Heterogenous, Distributed Resources*, 1995, pp. 13-18

[11] S. Chakrabarti, B. Dom, and P. Indyk, "Enhanced hypertext categorization using hyperlinks," in *SIGMOD '98: Proceedings of the 1998 ACM SIGMOD international conference on Management of data*. New York, NY, USA: ACM Press, 1998, pp. 307-318

[12] J. M. Kleinberg, R. Kumar, P. Raghavan, S. Rajagopalan, and A. S. Tomkins, "The Web as a graph: Measurements, models and methods," *Lecture Notes in Computer Science*, vol. 1627, pp. 1-18, 1999

[13] J. Hou and Y. Zhang, "Effectively finding relevant web pages from linkage information." *IEEE Trans. Knowl. Data Eng.*, vol. 15, no. 4, pp. 940-951, 2003

[14] H. Han and R. Elmasri, "Learning rules for conceptual structure on the web," *J. Intell. Inf. Syst.*, vol. 22, no. 3, pp. 237-256, 2004

[15] M. Eirinaki and M. Vazirgiannis, "Web mining for web personalization," *ACM* Trans. *Inter. Tech.*, vol. 3, no. 1, pp. 1-27, 2003

[16] J. Pei, J. Han, B. Mortazavi-Asl, and H. Zhu, "Mining access patterns efficiently from web logs," in *PADKK '00: Proceedings of the 4th Pacific-Asia Conference on Knowledge Discovery and Data Mining, Current Issues and New Applications*. London, UK: Springer-Verlag, 2000, pp. 396-407

[17] R. Cooley, B. Mobasher, and J. Srivastava, "Data preparation for mining world wide web browsing patterns," *Knowledge and Information Systems*, vol. 1, no. 1, pp. 5-32, 1999

[18] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web usage mining: Discovery and applications of usage patterns from web data," *SIGKDD Explorations*, vol. 1, no. 2, pp. 12-23, 2000

[19] M. S. Chen, J. S. Park, and P. S. Yu, "Data mining for path traversal patterns in a web environment," in *Sixteenth International Conference on* Distributed *Computing Systems*, 1996, pp. 385-392

[20] J. Punin, M. Krishnamoorthy, and M. Zaki, "Web usage mining: Languages and algorithms," in *Studies in Classification, Data Analysis, and* Knowledge *Organization*. Springer-Verlag, 2001

[21] P. Batista, M. ario, and J. Silva, "Mining web access logs of an on-line newspaper," 2002

[22] O. R. Zaiane, M. Xin, and J. Han, "Discovering web access patterns and trends by applying olap and data mining technology on web logs," in *ADL '98: Proceedings of the Advances in Digital Libraries Conference*. Washington, DC, USA: IEEE Computer Society, 1998, pp. 1-19

[23] J. F. F. M. V. M. Li Shen, Ling Cheng and T. Steinberg, "Mining the most interesting web access associations," in *WebNet 2000-World Conference on the WWW and Internet*, 2000, pp. 489-494

[24] X. Lin, C. Liu, Y. Zhang, and X. Zhou, "Efficiently computing frequent tree-like topology patterns in a web environment," in TOOLS '99: Proceedings of the 31st International Conference on Technology of Object-Oriented Language and Systems. Washington, DC, USA: IEEE Computer Society, 1999, p. 440

[25] A. Nanopoulos, Y. Manolopoulos, "Finding generalized path patterns for web log data mining," in ADBIS-DASFAA '00: Proceedings of the East-European Conference on Advances in Databases and Information Systems Held Jointly with International Conference on Database Systems for Advanced Applications. London, UK: Springer-Verlag, 2000, pp. 215-228

[26] A. Nanopoulos and Y. Manolopoulos, "Mining patterns from graph traversals," Data and Knowledge Engineering, vol. 37, no. 3, pp. 243-266, 2001

[27] X. Jin, Y. Zhou, and B. Mobasher, "Web usage mining based on probabilistic latent semantic analysis," in KDD '04: Proceedings of the 2004 ACM SIGKDD international conference on Knowledge discovery and data mining. New York, NY, USA: ACM Press, 2004, pp. 197-205

[28] S. Jespersen, T. B. Pedersen, and J. Thorhauge, "Evaluating the markov assumption for web usage mining," in WIDM '03: Proceedings of the 5th ACM international workshop on Web information and data management. New York, NY, USA: ACM Press, 2003, pp. 82-89

[29] A. Nanopoulos, D. Katsaros, and Y. Manolopoulos, "Exploiting web log mining for web cache enhancement," in WEBKDD '01: Revised Papers from the Third International Workshop on Mining Web Log Data Across All Customers Touch Points. London, UK: Springer-Verlag, 2002, pp. 68-87

[30] A. Nanopoulos, D. Katsaros and Y. Manolopoulos, "A data mining algorithm for generalized web prefetching," IEEE Transactions on Knowledge and Data Engineering, vol. 15, no. 5, pp. 1155-1169, 2003

[31] J. Zhang and A. A. Ghorbani, "The reconstruction of user sessions from a server log using improved timeoriented heuristics." in CNSR. IEEE Computer Society, 2004, pp. 315-322