Sequential Pattern Discovery from Web Log Data

Rajashree Shettar
R.V College of Engineering
Mysore Road, Bangalore
Karnataka, INDIA

ABSTRACT
Pattern mining from the web log data leads to discovery of usage patterns of the user who navigate the web. Patterns which appear frequently in the web log data are item-sets and sequences. In this paper, a novel algorithm Intelligent Generalized Sequential pattern (IGSP) is designed which shows better results than the Generalized Sequential Pattern (GSP) algorithm. Experiment is conducted with respect to running time and number of patterns discovered from the log data and results has shown that IGSP outperforms the well-known algorithms (GSP) algorithm.

Keywords
Web usage mining, sequential patterns, web log.

1. INTRODUCTION
The semi-structured data of the Web in the form of web pages is available in abundance. It has become increasingly difficult to extract useful and meaningful information from the web pages. Web mining provides the mechanism to extract useful and meaningful information. Web mining can be categorized into three different classes based on which part of the Web is to be mined [1, 2, 3]. These three categories are (i) Web content mining, (ii) Web structure mining and (iii) Web usage mining. Web content mining deals with finding useful information from the web pages based on contents of the web pages [14]. Web content mining helps to find, discovery and extract relevant information from the web pages. The structure of the hyperlinks of the Web can be found by Web structure mining. Web structure mining discovers the link structures at the inter document level. The information available can be put in the form of a graph and information can be extracted by performing classification and clustering of relevant information from the graph. Web usage mining helps in identifying the behavior pattern of the users such as browsing and navigation of the web. This helps in identifying the navigation preferences of the visitors which in turn helps to enhance the quality of electronic commerce services, to personalize the web portion, to improve the web structure and server performances. 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. To extract and analyze useful information such as traversal behavior of the users who navigate the web, Web Usage Mining is required [4]. Many different techniques for mining sequential patterns have been proposed in the recent past. The authors in [5] discuss that the candidate-generation-and-test approach outperforms the pattern-growth approach on mining short patterns, while pattern-growth approach is better on mining long patterns. The authors in [6] discuss the process of web log mining. Sequence mining is accomplished in [7], where a so-called WAP-tree is used for storing the patterns efficiently. Tree-like topology patterns and frequent path traversals are searched by [8, 9, 10, 11]. Mining traversal patterns is one of the techniques of web usage mining. The web log data which is considered as secondary data of the web has been taken up in this work for the discovery of frequent patterns. The web usage mining domain has several types of information available that can be used as surrogates for domain knowledge. The various applications in which it is used are advertisements, creation of dynamic user profiles and modification of the structure of portal in such a way that the pages are accessible from each other. A web server usually registers a log entry for every access of a web page. First, raw web log data need to be cleaned, condensed and transformed in order to retrieve and analyze significant and useful information. Second, pattern mining can be performed on log records to find association patterns, sequential patterns and trends of web accessing. With the use of such patterns, studies have been conducted on analyzing system performance, improving system design by web caching, web page pre-fetching, understanding the nature of web traffic, etc. A frequent pattern is the set of web pages visited together in a session, whose support is above the minimum support threshold. In this regard, repeated visited web pages are ignored and web pages are not necessarily ordered and consecutive. Frequent patterns are very helpful to identify pages accessed together in one session. The process of web usage mining consists of three major steps [12] (1) data preprocessing, (2) pattern discovery, and (3) pattern analysis stages. Log files are stored on the server side, on the client side and on the proxy servers. In this work, only the server side web log data is being used which is the Click stream data set [13]. The data preprocessing phase includes the data cleaning, user identification, session identification and data transformation which are indicated. The pattern discovery phase involves the discovery of frequent sequences. The pattern analysis phase involves the analysis of the frequent patterns generated by the pattern discovery phase.

2. DATA PREPROCESSING
Data preprocessing phase involves data cleaning, user identification and session identification and data transformation [12]. In data cleaning stage phase, the web log is examined and irrelevant or redundant items such as image, sound, video files, executable cgi files and HTML files which could be downloaded without an explicit user request are removed. Data cleaning stage also involves removal of HTTP errors, records created by crawlers, etc. Data cleaning is performed by checking for file extensions such as GIF, JPEG, jpg, mpg, avi, etc. These are removed from the log. Each

record of the log contains the following 6 fields. They are the shop-id, time the page was visited, IP address of the user visiting the page, a unique session identifier, the current page that is visited and the last field is the referrer which refers to the page from which the current page was visited.
The user identification phase involves identification of users from the log data. The following procedure is adopted for identifying the users. (1) A new IP identifies a new user. (2) If the same IP is used, but different web browsers or different operating system in terms of type and version is being used, this is considered as new user.

The session identification stage involves identifying sessions according to different users. A session is a group of activities performed by the user while navigating through a given site within a given time period. In this work, a set of pages visited by a specific user is considered as a single user session if the pages are requested at a time interval not larger than a specified time period of 60 milliseconds.

In the data transformation phase, the web log data is converted into a format needed by the mining algorithms. For every session that is identified, the log data is converted into two databases – transactional database and sequential database. The transactional database consists of two fields-transaction id and the set of pages visited in a session. A sequential database consists of sequence of pages visited in a session and the data set used is Click Stream data which is server side log data. The records are sorted in ascending order based on IP address and the number of users is identified. The boundaries of each user are marked and stored in a two dimensional array. The boundaries specify the starting index and the ending index of a particular user in a sorted log file. In the session identification phase of the pre-processing stage, a session timestamp value is set which specifies the total time of a particular session. The value of the session timestamp has been set as 60 milliseconds. The record obtained is checked to see if it belongs to a particular session. If it belongs to the same session, it is added to the session list and the next record is processed. If it does not belong to the same session, then the previous session records are put onto the file and then the session list is cleared. A new session is created and the current record is added to the new session and its timestamp value is initialized as the session start time. This process is repeated for every user and all sessions are identified. The records in every session are then transformed into item sets and sequences. In the data transformation phase, all the fields of the records are transformed into appropriate format.

3. FREQUENT SEQUENCES

Intelligent Generalized Sequential Pattern (IGSP) algorithm is used to find the frequent sequences that occur in the log file for a given threshold. This algorithm is an extension of GSP algorithm [14] and efficient than GSP as the running time is lesser when compared to GSP. Hash tables are used which improves the efficiency. The pre-processed data, in the form of sequences is taken as input to the algorithm. Every row in the sequential database corresponds to a sequence of web pages visited in a particular session. This sequential database also contains the subsequences of the sequences found. A file containing the various frequent sequences that have been identified from the log file is obtained as output. The pseudo code for the IGSP algorithm is listed in steps from (1) to (6) below.

3.1 Algorithm IGSP

1. Read the sequential database and store the sequences in a list.
2. Store the sequences read from the sequential database into a hash table and initialize the key values of sequences to '0'.
3. Initialize the hash table entries with key value '0'.
4. If the sequence is present in the hash table, increment the count value corresponding to the string by one. If the sequence is not present in the hash table, insert the sequence into the hash table and set count=1. The hash table now contains the sequences along with their counts.
5. Check the sequence count and compare it with minimum support threshold value. If the sequence count is less than the minimum support threshold, then ignore the sequence. Otherwise add the sequence to the file as frequent sequence.
6. If remaining sequences are not exhausted go to step (4) else exit.

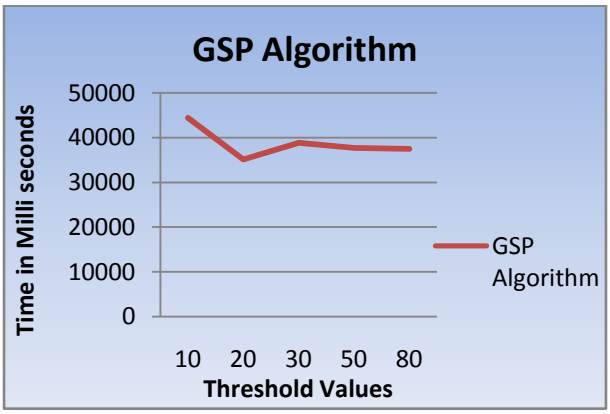
4. EXPERIMENTAL RESULTS

The experimental results and analysis of frequent pattern discovery algorithms from web log data is described in this section of the paper. The frequent patterns generated from the web log data are compared using well-known algorithms and also performance of pattern discovery algorithms on the web log data based on parameters such as running time and number of patters generated is analyzed. The experimental analysis is conducted on the Click Stream Data set [13] which is a server side web log data containing 12,000 records. Each record contains the following fields – a shop identifier, time stamp, IP address, unique session identifier, page visited, and referrer. The frequent sequences generated from click stream data set with minimum support threshold count of 10, 20, 30, 50 and 80. The performance is tested on a computer with a 1.41GHz processor. The program is developed using JDK 1.6. Table 1 shows the comparison between the running times of GSP and IGSP algorithms. From graph 1 and graph 2, it is observed that the running time taken by the IGSP is lesser compared to the running time taken by GSP algorithm. Table 2 shows the comparisons between number of patterns generated by GSP and IGSP algorithms. It is observed from graph 3 and graph 4, the total number of patterns generated by the IGSP is higher than that of the GSP.

Table 1: Comparison of the running time of GSP and IGSP algorithms

Threshold Values	GSP ALGORITHM (in mS)	IGSP ALGORITHM (in mS)
10	44444	374
20	35146	361
30	38860	363
50	37690	359
80	37501	359

Graph 1: Comparison of the running time of GSP with respect to the threshold values



Graph 2: Comparison of the running time of IGSP with respect to the threshold values

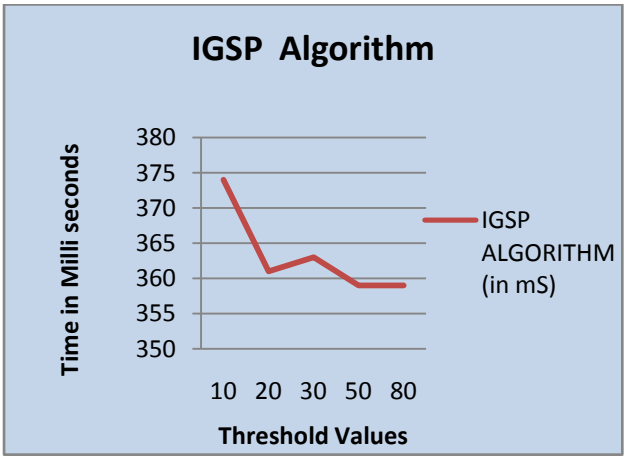
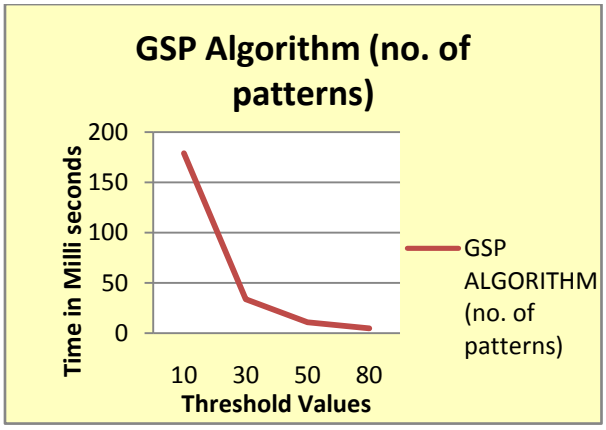


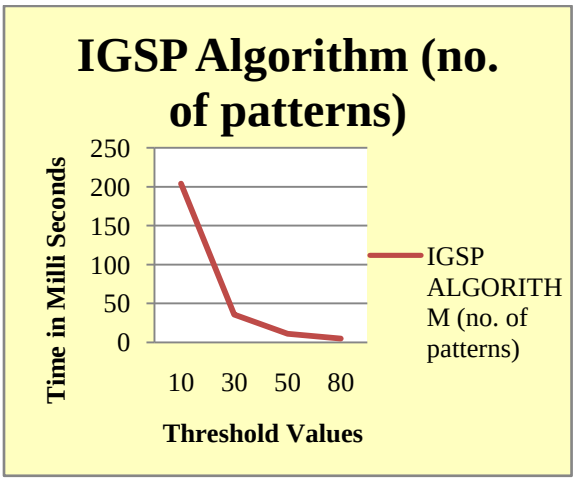
Table 2: Comparisons with respect to the number of patterns generated by GSP and IGSP algorithms

Threshold Values	GSP ALGORITHM (no. of patterns)	IGSP ALGORITHM (no. of patterns)
10	179	204
30	34	36
50	11	11
80	5	5

Graph 3: Comparison of GSP algorithm with respect to the number of patterns generated and threshold values.



Graph 4: Comparison of IGSP algorithm with respect to the number of patterns generated and threshold values.



5. CONCLUSIONS

Web usage mining techniques has been applied to large web repositories to extract usage patterns. In this paper analysis of pattern mining algorithms is done based on running time and number of patterns generated from the web log data. It is found from the experimental analysis that IGSP algorithm performs better in terms of running time and number of patterns discovered compared to GSP algorithm.

The work carried out in this paper with respect to finding sequential patterns in web log data involves statistical analysis of web usage mining of server side log data. This can be further enhanced to find frequent patterns considering client side log data applying graph mining methods to discover web usage patterns.

6. REFERENCES

[1] 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

[2] 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, pp. 303-312, 1999.

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

[4] R. Kosala and H.Blockeel. Web mining research: a survey. In ACM SIGKDD Explorations, 2000.

[5] Sun, L and Zhang, X 2004, 'efficient frequent pattern mining on web logs', in JX Yu et al, "Advanced Web Technologies and Applications:6th Asia-Pacific Web Conference, APWeb 2004, Berlin, March 2004.

[6] Renáta Iváncsy, István Vajk , "Frequent Pattern Mining in Web Log Data", pp.77-70, Vol. 3, No. 1, 2006.

[7] 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

[8] 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

[9] 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

[10] A. Nanopoulos and 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

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

[12] Cooley R., Mobasher B., and Srivastava J., "Data Preparation for Mining World Wide Web Browsing Patterns," In *J. Knowledge and Information Systems*, pp. 5.32, vol. 1, no. 1, 1999. 15. R. Kosala, H. Blockeel, "Web Mining Research: A Survey," In *SIGKDD Explorations*, ACM.

[13] Stream data downloaded from the *ECML/PKDD 2005 DiscoveryChallenge2*. http://lisp.vse.cz/challenge/CURRENT

[14] S. Chakrabarti, "Data Mining for hypertext: A tutorial survey", SIGKDD Explorations; Newsletter of the special Interest Group (SIG) on Knowledge Discovery and Data Mining, ACM, Vol.1, No.2, pp.1-11, 2000.