pSPADE : Mining Sequential Pattern using Personalized Support Threshold value

Suraya Alias[1], Norita Md Norwawi[2]
College of Arts and Sciences, Universiti Utara Malaysia
[1]s86929@ss.uum.edu.my, [2]nmn@uum.edu.my

Abstract

As the web log data is considered as complex and temporal, applying Sequential Pattern Mining technique becomes a challenging task. The min sup threshold issue is highlighted - as a pattern is considered as frequent if it meets the specified min sup. If the min sup is high, few patterns are discovered else the mining process will be longer if too many patterns generated using low min sup. The format of web log data that creates consecutive occurring pages has made it difficult to generate frequent sequences. Also, as each user' behaviour is unique; using one min sup value for all users may affect the pattern generation. This research introduced a ***personalized minimum support*** *threshold for each web users using their* ***Median*** *item access (support) value to curb this problem. The* ***pSPADE*** *performance was the highest on the discovery of user's origin and also interesting pattern discovery attribute.*

1. Introduction

Web Usage Mining manipulates the web access log data in order to identify the hidden pattern of the visitors that accessed the website [1-4]. There are 3 major steps in Web Usage Mining process that includes; Pre-processing, Pattern Discovery and Pattern Analysis. This research highlighted the goal of Sequential Pattern Mining in web usage (log) data that is to discover users' frequent sequences pattern while navigating a website.

Referring to the definition: A sequence is considered as *frequent* if it occurs more than *min sup* times [5], there exist an inter-relationship between pattern generation and the *min sup* threshold specified in the algorithm. The *min sup* variable determines the numbers of frequent sequences patterns generated in the Sequential pattern mining process. Problem arises when the *min sup* is set ***too high***, that will lead into uninteresting or few pattern discovery due to short frequent sequences patterns generated. However, if the threshold is set ***too low,*** the response time are affected due to huge frequent sequences pattern generated (all the patterns may or may not be interesting). The other challenge in this research is the format of web log data that logs every user' access that creates consecutive occurring pages or duplicate pages that has made it harder to generate frequent sequences. Also, as each user' behaviour is unique; by using one min sup value for all users may affect the overall results or pattern generation.

The goal of this research is to introduce a ***personalized*** minimum support threshold for each web users using their ***Median*** item access (support) value to curb this problem. The proposed technique (*pSPADE)* was implemented in the SPADE algorithm and even though the personalized min sup value for each user is different; their range (half or 50%) still remains the same. Thus, the use of the personalized min sup will increase the efficiency in frequent sequence access pattern generation, maintain the number of frequent pattern generated at every sequence; at the same time avoiding lengthy or uninteresting pattern discovery.

2. Related Works

Earlier works in Sequential Pattern mining until current are mainly focused on the algorithm improvement and efficiency [5-11], whilst this research is more interested on the usage of variable ***minimum support*** (denoted *min sup*) threshold which is specified by the analyst in the Sequential Pattern Mining Algorithm. A research by [2], discussed in comparing which pattern discovery algorithm is suitable for mining web server logs. The algorithm comparison consists of four frequent pattern mining algorithm; SPADE (Sequential Pattern Discovery using Equivalence classes), GSP (Generalized Sequential Pattern), Breadth First Search (BFS) and Depth First Search (DFS).

Their result has showed that SPADE algorithm performs better compared to others in terms of omitting irrelevant pattern due to its accurate calculation function in identifying frequent sequences.

[12] manipulated the GSP algorithm by [13] in order to solve the min sup issue by allowing the user

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

to specify multiple minimum supports in the database in the algorithm named **MSApriori**. Even though the result shows that the technique is quite effective, the user still has to have some knowledge regarding the data to avoid recurring mining process if the multiple min sup applied is not suitable – thus the issue of unsuitable min sup applied is still open.

Moving to the use of web log data as the data source, [6] proposed the additional use of Regular Expressions (REs) constraint in the mining process together with the min sup value in the **SPIRIT** Algorithm. This technique is considered useful as it focus on users' interest and requirement, however the min sup issues still occurs since the user's behaviour is dynamic and the WWW has evolved since the last algorithms was introduced.

Also related to this research~~,~~ [14] proposed of mining the Top-*k* frequent Closed pattern, where *k* is the number of pattern to be generated without using the min sup in the **TFP** Algorithm. The same algorithm was then improved to **TSP** by [15] with the same method of removing the min sup value and replacing it with the *min_l* or minimal length of each pattern. As for this research, the closed sequential pattern mining approach is not applicable due to the complex data source used (web log); also as stated earlier by assuming the same support for each user can increase the risk of losing some interesting pattern.

Later, mining the updated database becomes a challenge in the Sequential Pattern Mining area. **ISM** by [16] was the first extension from the SPADE algorithm to solve these issues by updating the support (incrementally) and then generating the frequent sequences. As this approach solved the updated data issue, when the support becomes higher, problems will still occurs as it will not be relevant to mine the web log data since some of the frequent sequences may not be available anymore.

Another interesting approach by [17] is **KISP**, whereby the solution is to change the support by manipulating the knowledge discovered from previous mining. A knowledge base consists of queries from various min sup was generated to avoid the re-mining process from scratch. The proposed technique is faster than previous GSP algorithm, nevertheless can be used as reference for this technique. However, this study focused on user's behaviour that is unique, and the website content, design and structure will undergo such changes that even a knowledge based was created today may not be relevant for the next mining process.

By taking into accounts previous initiative in solving the min sup issues, the research has found that there should be a technique that can simplify or assist the process of determining the most "appropriate" min sup value when handling web log data that is considered as temporal, complex and incomplete. Even though in real-world application, there are many Web Usage Mining products available in the market, i.e. eNuggets, Clementine and 123LogAnalyzer, the research has to agree that not all is designed and customized up to our requirement and classification. Furthermore, by manipulating Sequential Pattern Mining algorithm (SPADE) with the personalized min sup, this research aims to address the potential of Web Usage Mining in Academic Website.

3. Significance and Contribution

The results from this research will provide assistance in discovering any significant information that can support UUM in forecasting prospective student locally or internationally by identifying the web user's country of origin. By understanding the visitors' access behaviour pattern in terms of the frequent pages that they accessed and how they navigated the Academic Website can lead to better target marketing strategy that will help to increase profit and revenue. Also, it can improve the current marketing strategy that still uses 'mouth-to-mouth' technique in order to spread information especially for international student.

Significant from that, more business opportunity can be explored since we had identified potential prospective student from the pattern derived and analyze the relationship against Postgraduate Admission data whereby there exist similarity between both entities. Thus, we are being pro-actively step ahead in preparing reliable information for the prospective students that will make us competitive in this education world.

Furthermore, this research also aims to contribute in Sequential Pattern Mining research area related to web log data. By implementing users' ***personalized*** minimum support or ***pSPADE*** technique in the Sequential Mining algorithm (SPADE) has increased the effectiveness of frequent sequence pattern generation by avoiding too little or too many pattern generated in every sequence. This technique has eased up the process of determining the most "appropriate" min sup that should be used to mine the web log data rather than manually selecting the min sup value.

Thus, the proposed technique has indirectly improves the performance of the Sequential Mining algorithm; simplified the Pattern Discovery and Analysis process in the Web Usage Mining task.

4. Web Usage Mining

Web Usage Mining high level process involves 3 major steps that consists of Pre-processing, Pattern Discovery and Pattern Analysis [3].

4.1. Pre-processing

For this research, the web server access log is retrieved from the PPS server from 1st November 2005 until 16th May 2006 (http://www.pps.uum.edu.my). The content of the PPS web server log file consists of 7 fields that are separated by spaces (see Figure 1). The common log file format is as follows; 1) remote IP address 2) remote login name 3) user name 4) date 5) file requested (URL) 6) request status, and 7) total of bytes transferred. The fields 1,4,5,6 and 7 will be used respectively in this research as the two other fields are not available.

```
-----------------------------------------------------------------
--
66.249.64.52 - - [24/Feb/2005:03:45:40 -0500] "GET
/admission.htm HTTP/1.0" 200 3116
66.249.64.54 - - [24/Feb/2005:03:45:43 -0500] "GET
/handbook/rules/couse_load.htm HTTP/1.0" 200
6363
66.249.71.55 - - [24/Feb/2005:03:45:45 -0500] "GET
/handbook/rules/period_study.htm HTTP/1.0" 200
6222
66.249.71.53 - - [24/Feb/2005:03:45:47 -0500] "GET
/uum.htm HTTP/1.0" 200 7169
66.249.64.47 - - [24/Feb/2005:03:45:47 -0500] "GET
/fee.htm HTTP/1.0" 200 4361
66.249.64.16 - - [24/Feb/2005:03:45:49 -0500] "GET
/handbook/rules/grading_sys.htm HTTP/1.0" 200
12706
66.249.71.55 - - [24/Feb/2005:03:45:59 -0500] "GET
/handbook/rules/mode_study.htm HTTP/1.0" 200
6210
-----------------------------------------------------------------
--
```

Figure 1: Raw PPS Web Server Log Data

The first task in Pre-processing is Data Cleaning whereby the log file is be filtered for any access errors, search engines and robots.txt. Image files accessed i.e. *.jpg, *.gif and other symbols are also removed.

Next, User Identification step based on the unique IP Address field is performed. Due to incomplete information available in the PPS web log data, there are known issues in this process that has been discussed by [3, 18] that should be highlighted such as;

1) Incorrect inference of IP Address, there exist some rare cases whereby the IP Address of some visitors may not belong to the local machine; instead it belongs to the Internet Service Protocol (ISP) proxy IP Address.
2) Existence of multiple sessions, IP Address and agent has made it more difficult to trace the users.
3) Incomplete log data i.e. no referrer, login name and browser version.

However, few solution has been considered and summarized by [19] to reduce this problem, that is either using login data in the common log file or by using cookies and site topology.

After the user's origin has been identified, the next step is to divide the user's navigation into sessions. A user session is a delimited numbers of pages that was requested by user to a server during a specific time period. There are few techniques available to identify the user's session such as using a time gap or referrer as detailed by [19] in their journal. The most accepted technique is by using a time gap between the entries and is applied in this research as follows.

1) Select minimum access *Datetime* for each user
2) Set duration threshold as 20 minutes
3) For each user record, calculate duration between entries
4) If the duration >= 20 minutes, create new session, else it's a same session

The steps above also can be generally written as below whereby user session is denoted by *s* and timestamp is *t*;

If $s.t_{n+1} - s.t_{n} \geq time_{threshold}$ **then *new session***

The research session threshold time is set to 20 minutes, as the structure of PPS website is very straightforward and the content is not very heavy. The threshold value can varies through different application and the standard cut-off time is determined as 25.5 minutes according to previous research by [20]. However, related research using academic website log by Ciesielski and Lalani (2003) has defined that any transaction or request made to the server that is less than 30 minutes is considered as part of the same session. After the user's session has been constructed, the next step in Web Usage Mining process is the Pattern Discovery that involves the development of pSPADE technique.

5. *pSPADE* Development

The main features of *pSPADE* technique are summarized as below;

1) Generation of *session id timestamp list*, where the web log data are transformed or formatted into $\text{sid}_{i} = (\text{PAGES}_{i})$ format, *sid* is the number of session and *PAGES* refers to the web pages that belongs to the specific user session. Then, the list was improved by removing consecutive occurring pages using Microsoft.Net RegEx (regular expression) approach.

2) Generation of *personalized min sup* according to the median item support value for each user.

3) Implementing the personalized min sup in the SPADE Algorithm and discovering frequent sequences pattern for each user.

5.1. Session Id Timestamp List

The web access log data need to be formatted since the purpose of Sequential Pattern Mining technique previously is to analyze market basket transactions, that makes the web log data is not suitable for direct mining process [21]. Therefore, for each user's session, the series of web pages accessed by the user is transformed into below format using the proposed transformation technique by [22].

$$\underset{\text{Session id}}{sid_i} = \overbrace{1^{st}\ Item}^{\text{Timestamp 1}} \rightarrow \underbrace{2^{nd}\ Item}_{\text{Timestamp 2}} \rightarrow \overbrace{3^{rd}\ Item}^{\text{Timestamp 3}} \rightarrow \underbrace{4^{th}\ Item}_{\text{Timestamp 4}}$$

The item is the web pages accessed in the user's session that will be used as the *timestamp* in generating the session id-timestamp list; meanwhile *i* refer to number of session for each user (session id). In this research, the item or pages is denoted using alphabet in lexicographic or dictionary order sorted by the Top 20 highest access hits (webpage) to the PPS website. Thus, "*/academic%20handbook.pdf*" page is written as Page $_A$ or *A* in this paper in order to simplify the term.

In the process of converting all users' session data into the session id timestamp list format above, the issue of consecutive occurring pages in every session is solved using RegEx (regular expression) approach in the Microsoft.Net framework. The improved list is more relevant and efficient compared to the original generated list despite the fact that the nature of the web log that is not '*user friendly'* – it logs every user' movement on the website. Example of the improved session format after removing all consecutive occurrences of pages is illustrated as below Figure 2, whereby there are five web pages *(B,D,L,H,N*) being accessed in four different sessions by the user.

sid $_1$ = *BDDDBDDDDDDDDDDDDDDBLHHHHHH*
sid $_2$ = *BLHHHHHHHHHHHNNNNNNNLBL*
sid $_3$ = *DDDDDDDDDDDDD*
sid $_4$ = *BDDDDDDDDDDD*

⇩

sid $_1$ = *BDBDBLH*
sid $_2$ = *BLHNLBL*
sid $_3$ = *D*
sid $_4$ = *BD*

Figure 2: Web log data before and after re-formatting

Then, the *session id timestamp list* for the user is generated as listed in below Table 1.

Table 1: Example of session id timestamp list

Page $_B$ or (*B*)		Page $_D$ or (*D*)		Page $_H$ or (*H*)		Page $_L$ or (*L*)		Page $_N$ or (*N*)	
sid	ts	sid	ts	sid	ts	sid	ts	sid	ts
1	1	1	2	1	7	1	6	2	4
1	3	1	4	2	3	2	2		
1	5	3	1			2	5		
2	1	4	2			2	7		
2	6								
4	1								

5.2. Personalized Min Sup

In order to select the most "appropriate" min sup value for each user; the process of generating personalized min sup is done by manipulating the mathematical median formula. The support for each item or pages is counted, sorted and then the median or middle item value is selected based on the formula presented.

Due to that, even if the minimum support value for each user is different; their range (half or 50%) still remains the same. Next are the steps in order to generate personalized min sup for each user.

Step 1: Support counting for each item

Each item or pages is counted as one if it occurs in each session (regardless how many times it was accessed in each session). The support for each item or pages is counted using below formula (1);

Support (P) = (item Count / number of session) (1)

Step 2: Sort the item according to highest support

Step 3: Apply mathematical median formula to get the median item

Median Item = ([Total number of item] + 1) / 2 (2)

Step 4: Personalized minimum support value

Personalized min sup = median item value / total of session (3)

Example of the personalized min sup process is illustrated in below Table 2. Since there are five items or web pages that were accessed in four sessions, by applying the median formula in Step 3 (refer equation 2), (5+1)/2 = 3. Thus, the 3rd item is Median, with the item count value of 2. The personalized min sup formula (refer equation 3), 2/4 = 0.5. Thus, from here an item sequence is considered as *frequent* if it occurs >= 0.5.

Table 2: Example of personalized min sup generation

#	Item or Page	Item Count	Support (P)
1	*B*	3	0.75
2	*D*	3	0.75
3	***H***	**2**	**0.5**
4	*L*	2	0.5
5	*N*	1	0.25

5.3. *pSPADE* Implementation

After generating the personalized min sup, the technique is then implemented in the SPADE Algorithm. The steps are given as follows;

Step 1: Pattern Lattice Construction

By using the personalized min sup, any item or page that has the support below it is removed. Each single item or page refers to length-1 prefix equivalence class. The frequency of upper elements or items that has the length of *n* can be calculated using two *n*-1 length patterns from the class. Figure 3 below is lattice for three items or pages denoted as *A*, *B* and *C* with length-2 patterns (Frequent 2-Sequences) and together with some example of length-3 pattern (Frequent 3-Sequences).

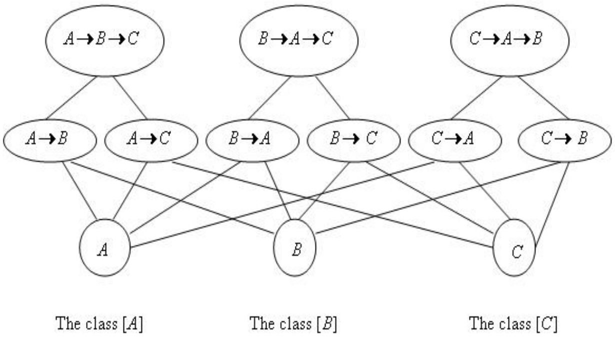

Figure 3: Example of Prefix based classification of pattern lattice

Step 2: Breadth First Search (BFS)

BFS approach is chosen since it has the benefit of keeping only current length-*k* pattern in the memory. Here, the lattice of equivalence classes is generated by the recursive application and the whole lattice is explored from bottom to top manner. The next frequent pattern (Frequent 2-Sequences) are generated using union operation $\vee$ from the single item (Frequent 1-Sequences); where all child length-n patterns are generated before moving into parent patterns with length n+1.

Using the Figure 3 above, for example the patterns $A \rightarrow B$ and $A \rightarrow C$ are generated before $A \rightarrow B \rightarrow C$ and $A \rightarrow C \rightarrow B$. For each sequence generation, again the personalized min sup is checked to see if the pattern generated is frequent or not. In this pruning process, if the child or current item is not frequent, then the parents cannot be frequent or generated.

Step 3: SPADE Algorithm extension

The first step in the *pSPADE* technique is creating the Session id-timestamp list as explained earlier. The list was then sorted according to their support (highest to lowest). The generation of personalized min sup as highlighted is the extension or major enhancement for this algorithm. After the personalized min sup has been determined, the item or pages that have the support lower from the personalized min sup is eliminated. Then, frequent patterns from a single item or page are enumerated or constructed using the union operation $\vee$ based on prefix-based pattern lattice approach as illustrated in Figure 3. Following after, the BFS approach is implemented in order to search for the frequent pattern in each k-sequence based on personalized min sup.

The extension of SPADE Algorithm with personalized min sup or *pSPADE* technique is given as below;

P_minsup = personalized min sup
S = support

```
SPADE (P_minsup, S)
P1 = all frequent atoms with their Session id-timestamplist
P2 = all frequent length-2 patterns with their Sessionid-timestamp list
E = All equivalence classes of Atoms $\in$ P1

    For all [X] $\in$ E do
        Construct pattern lattice of [X]
        Explore frequent patterns in [X] by
    Using BFS.
```

6. Results

The results form this research is analyzed using two approaches; 1) Statistical Analysis and 2) Pattern Analysis Comparison Attributes.

6.1. Statistical Analysis on web log and admission data

By using *correlation coefficient r* formula to measure the linear correlation (4), whereby r is the real value $r \in [-1,1]$, will determine either this two variables *X* and *Y* have positive or negative correlation.

$$Correl(X,Y) = \frac{\sum(x-\bar{x})(y-\bar{y})}{\sqrt{\sum(x-\bar{x})^2\sum(y-\bar{y})^2}} \quad (4)$$

Given variable *X* as the total of student that registered in each faculty according to region (admission data), while variable *Y* represents the total of visitors that accessed to each faculty website link according to region (web log data). Thus, the purpose of this experiment is to discover the relationship between the Admission data in term of faculty that the students enrolled *X* and the website' faculty link that the users' accessed *Y*.

The result from experiment done on South East Asia (SEA) region is detailed in table 3. The Admission data column or variable *X* indicates the total of student that registered in each faculty with the total of (856) students from SEA region, while the web log data column or variable *Y* represents the total of visitors that accessed each faculty website with the total of (15260) users from SEA region.

Table 3: Correlation Results from SEA region

Faculty	**South East Asia Region (SEA)**	
	Admission data (X)	***Web log data (Y)***
FE	9	378
FKBM	29	802
FPA	58	605
FPAU	29	648
FPH	11	475
FPK	36	1127
FPP	312	4927
FPSM	118	1640
FPT	3	612
FSK	6	444
FSKP	109	1638
FTM	88	1124
FWB	48	840
Total	**856**	**15260**
Correlation	**+ 0.976476**	
Indicator	**Strong Positive**	

Thus the correlation findings can be summarized as;

- Variable *X* and *Y* from SEA region have a ***strong*** correlation of + 0.97. This indicates that users' navigation behaviour in accessing information from the respective faculty link is similar or dependent with the students' admission to each faculty. For example FPP has the highest student admission of (312) and (4927) visitor access This means that if user navigates to FPP link, there are strong dependency that the student are interested to enter the respective faculty. This result is predictable since most of the students and users from SEA region is from Malaysia.

From the Statistical Analysis done on the preliminary results of web log data and admission data has proven that there exist positive relationships between both entities. The reason being is that all the information regarding the postgraduate admission to UUM is available online that had made it easier for prospective user or student to retrieve it.

This also indicates that from the user's pattern on navigating the web pages (i.e. faculty link) can lead in predicting which faculty the users are interested in. From here also, many proactive measures and action can be taken to upgrade the level of the information in the website in order to attract more prospective student to further their higher education in UUM.

6.2. Pattern Analysis Comparison Attributes

The experiment that was conducted uses three (3) different minimum supports for each user and was implemented in the SPADE algorithm. For *High* minimum support the value was set to be *0.7*, for *Low* minimum support the value is *0.2,* and the last min sup value is personalized for each user. The High and Low min sup value was chosen randomly in this research as the min sup value is usually determined by the analyst. Thus, each user has three different *min sup* threshold value (0.7, personalized and 0.2) used throughout this experiment in order to extract the set of all frequent sequences that was generated from Frequent 1-Sequences until Frequent 3-Sequences in each users' access session.

The Pattern Analysis attributes are chosen based on the limitation information from the web data, as only the URL, IP Address and access date time value are used throughout the whole web usage mining process. The comparison attributes that were evaluated for each min sup are as listed;

1) Number of patterns discovered,
2) Web Users' Origin,
3) Number of users processed and
4) Interesting Pattern Discovery.

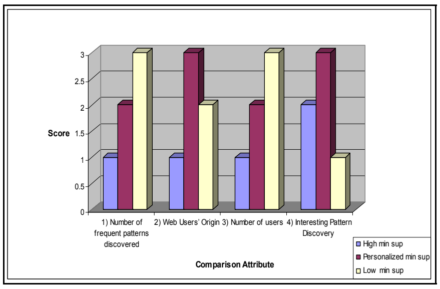

Figure 4: Summary of Personalized min sup evaluation results

Figure 4 is the illustration of the results from the four comparison attribute that was used to evaluate the effectiveness of the proposed *pSPADE* technique. The personalized min sup performance was the highest on the discovery of user's origin and interesting pattern discovery attribute, while moderate in other attributes. The numbers of pattern discovered using the personalized min sup is sufficient for further analysis – not to huge or too little patterns generated. Thus has lower down the mining process time and simplify the Pattern Analysis process.

From the web user' origin analysis results has indicates that using the *personalized* min sup for each user has the effect in determining the relationship between web log data and postgraduate admission data in terms of user's origin whereby the results shows the number of countries that was discovered is closer to the postgraduate admission data. As for the number of users processed, by using this approach the risk of omitting prospective users that their support value is too high or too low can be lowered down.

Furthermore, the essential of this research that is to unleash the "interesting" behaviour or pattern from the web user has been fulfilled by using this proposed *pSPADE* technique.

7. Summary and Future Works

From the preliminary findings that have been carried out in this research, it should be highlighted that user's behaviour in navigating the Academic website does reflects the Higher Education admission process. For example if the user's navigates to the respective faculty link, there exist probabilities that the prospective student is interested to register in that faculty based on the correlation analysis that was carried out. The min sup issue that was raised in this research is reduced or eased by implementing the *pSPADE* technique in the Sequential Pattern Mining algorithm. As each user has their own personalized min sup, even if their value differs, the range still remains the same since the user's median item value is manipulated. The result has shown that this technique is viable since the pattern extracted is closer to the admission data pattern in terms of countries that was discovered; also the performance is the highest in findings of interesting pattern.

As this research only focus on Academic website, for future work the research aims to apply the proposed technique to other e-commerce or any type of website, thus can widen up the challenge in the Web Usage Mining research area.

8. References

[1] P. Desikan and J. Srivastava, "Mining Information from Temporal Behavior of Web Usage," AHPCRC Technical Report, Department of Computer Science, University of Minnesota. TR-2003-121, 2003.

[2] M. A. Bayir, I. H. Toroslu, and A. Cosar, "A Performance Comparison of Pattern Discovery Methods on Web Log Data," in *4th ACS/IEEE International Conference on Computer Systems and Applications (AICCSA-06)*. Dubai/Sharjah, UAE, 2006.

[3] J. Srivastava, R. Cooley, M. Deshpande, and P.-N. Tan, "Web Usage Mining : Discovery and Applications of Usage Patterns from Web Data," *ACM SIGKDD Explorations Newsletter*, vol. 1, pp. 12-23, 2000.

[4] J. X. Yu, Y. Ou, C. Zhang, and S. Zhang, "Identifying interesting visitors through Web log classification," *Intelligent Systems, IEEE [see also IEEE Intelligent Systems and Their Applications]*, vol. 20, pp. 55-59, 2005.

978-1-4244-2328-6/08/$25.00 © 2008 IEEE

[5] R. Agrawal and R. Srikant, "Mining Sequential Patterns," in *11th International Conference on Data Engineering (ICDE'95)*. Taipei, Taiwan, 1995.

[6] M. N. Garofalakis, R. Rastogi, and K. Shim, "SPIRIT: Sequential pattern mining with regular expression constraints," in *Proceedings of the 25th VLDB Conference*. Edinburgh, Scotland, 1999.

[7] M. J. Zaki, "SPADE: An Efficient Algorithm for Mining Frequent Sequences," *Machine Learning Journal*, vol. 42, pp. 31-60, 2001.

[8] F. Masseglia, P. Poncelet, and R. Cicchetti, "An efficient algorithm for web usage mining," *Networking and Information Systems Journal (NIS)*, vol. 2, pp. 571-603, 1999.

[9] A. Demiriz, "webSPADE : A Parallel Sequence Mining Algorithm to Analyze Web Log Data," in *Proceedings of The Second IEEE International Conference on Data Mining (ICDM 2002)*. Maebashi City, Japan: IEEE Computer Society (2002), 2002, pp. 755-758.

[10] M. Leleu, C. Rigotti, J.-F. Boulicaut, and G. Euvrard, "GO-SPADE: Mining Sequential Patterns over Datasets with Consecutive Repetitions," in *Machine Learning and Data Mining Conference(MLDM'03)*. Leipsig, Germany, 2003, pp. 293-306.

[11] S. Aseervatham, A. Osmani, and E. Viennet, "bitSPADE: A Lattice-based Sequential Pattern Mining Algorithm Using Bitmap Representation," in *Proceedings of the Sixth International Conference on Data Mining*. Boston, MA, USA, 2006.

[12] Bing Liu, W. Hsu, and Y. Ma, "Mining Association Rules with Multiple Minimum Supports," in *ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD-99)*. San Diego, USA, 1999.

[13] R. Srikant and R. Agrawal, "Mining Sequential Patterns: Generalizations and Performance Improvements," in *Proc. of the Fifth Int'l Conference on Extending Database Technology*. Avignon, France, 1996.

[14] J. Han, J. Wang, Y. Lu, and P. Tzvetkov, "Mining Top.K Frequent Closed Patterns without Minimum Support," in *ICDM '02: Proceedings of the 2002 IEEE International Conference on Data Mining (ICDM'02)*. Washington, DC, USA, 2002.

[15] P. Tzvetkov, X. Yan, and J. Han, "TSP: Mining top-k closed sequential patterns," *Knowl. Inf. Syst.*, vol. 7, pp. 438-457, 2005.

[16] S. Parthasarathy, M. J. Zaki, M. Ogihara, and S. Dwarkadas, "Incremental and interactive sequence mining," in *Proceedings of the eighth international conference on Information and knowledge management*. Kansas City, Missouri, United States, 1999.

[17] M.-Y. Lin and S.-Y. Lee, "Improving the Efficiency of Interactive Sequential Pattern Mining by Incremental Pattern Discovery," in *Proceedings of the 36th Hawaii International Conference on System Sciences (HICSS'03)*, 2003.

[18] M. Rosenstein, "What is Actually taking Place on Web Sites: E-Commerce Lessons From the Web Server Log," in *Proceedings of the 2nd ACM conference on Electronic commerce*. Minneapolis, MN USA, 2000, pp. 38-43.

[19] Z. Pabarskaite and A. Raudys, "A process of knowledge discovery from web log data: Systematization and critical review," *Journal of Intelligent Information Systems*, vol. 28, pp. 79-104, 2007.

[20] L. D. Catledge and J. E. Pitkow, "Characterizing Browsing Strategies in the World-Wide Web," *Computer Networks and ISDN Systems*, vol. 27, pp. 1065-1073, 1995.

[21] D. Tanasa, "Web Usage Mining : Contributions to Intersites Logs Preprocessing and Sequential Pattern Extraction with Low Support." Nice, France: University of Nice Sophia Antipolis, 2005, pp. 168.

[22] E. Frias-Martinez and V. Karamcheti, "A Prediction Model for User Access Sequences," in *Proceedings of the WEBKDD Workshop: Web Mining for Usage Patterns and User Profiles*, 2002.