An Improved Algorithm Research on the PrefixSpan Based on the Server Session Constraint

Hong-Guo CAI [1, a], Chang-An YUAN [2, b,*]

[1]*Department of Mathematics and Computer Science, the Guangxi College of Education, Nanning, 530023, China*
[2]*computer and information engineering college, the Guangxi Teachers Education University, Nanning, 530023, China*
[a]*webminning@163.com,* [b]*yca@gxtc.edu.cn*
**Corresponding author: yca@gxtc.edu.cn*

Abstract: When we mine long sequential pattern and discover knowledge by the PrefixSpan algorithm in Web Usage Mining (WUM).The elements and the suffix sequences are much more may cause the problem of the calculation, such as the space explosion. To further solve the problem a more effective way is that. Firstly, a server session-based server log file format is proposed. Then the improved algorithm on the PrefixSpan based on server session constraint is discussed for mining frequent Sequential patterns on the website. Finally, the validity and superiority of the method are presented by the experiment in the paper.

1 Introduction

Sequential pattern mining is one of the important areas of data mining which would discover frequent ordered events and sub sequences. There have three sequential pattern mining algorithms are mostly discussed [1]. Among them, the GSP (sequence Patten General) algorithm and the SPADE (Sequential Patten Discovery usage Equivalent classes) algorithm are directly or indirectly by the way of the Apriori method to mine Frequent Sequential pattern .The PrefixSpan (prefix projection mode) algorithm is different from the GSP algorithm and the SPADE algorithm that need to generate all the candidate frequent patterns. The pattern growth method of the PrefixSpan algorithm is by the way for separately constructing prefix and suffix sequence conditions. Some analysis of the three algorithms had shown that the PrefixSpan algorithm has the best overall performance [2][3]. More and more improved algorithms are studied on the basis of these [4][5]. Because sequential pattern mining is mainly a computational challenge, the sequential pattern mining algorithms that based some constraint are particularly efficient and become the focus of research.

The Web Usage Mining (WUM) is an important branch of the Web mining [6-8]. The sequential pattern mining of the WUM can service such as the intelligent website etc. The data source of WUM is mainly access records that the users had browsed Web server. The Web services technology has also become more and more perfect with the rapid development of dynamic web site. The traditional technology to record server log could not keep up with the technological innovation. In this paper, a server log record format based on session server is studied. This log format can better record users' access behaviors under the traditional Web technology and The Web services technology. That can greatly improve the efficiency of sequence pattern discovery and the log data preprocessing comparing with traditional login WUM. In general, the data preprocessing in WUM accounted for 70% of the entire work by traditional logs. In order to mine frequent sequential patterns based new server logs format, the concept of server session constraint was defined. Then the improved algorithm on the PrefixSpan based on the server session constraint (ss-PrefixSpan) is discussed. Finally, the validity and superiority of the ss-PrefixSpan algorithm are presented by the two experiments.

2 Web Server Logs Based on the Server Session

There are five commonly used formats for logging user's access logs on a Web server: the W3C extended log file format, the Microsoft IIS format, the NCSA common log file format, the Intensive binary log format, the ODBC log format. In order to further analyze the log records and ensure the consistency of terms in log format. W3C WCA (Web Characterization Activity) had issued a draft of the Web terminology standardization associated with the WUM. The draft provided a series of definitions of the data abstraction by WCA W3C and included users, page view, click-stream, user session, server session and episode. These definitions could be looked up in the W3C website [9].

With the development of the technology of the web2.0 and web services the commonly used formats for logs

© The Authors, published by EDP Sciences. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (http://creativecommons.org/licenses/by/4.0/).

record had been unable to meet the need of Web services in cloud computing. In this paper, the standardized log file format based on server session had been defined. It has the following advantages: The server sessions would be saved in the hard disk rather than in the memory that used to discard. Each user is uniquely identified through the cookies login or IP/ proxy / path etc. Each user's click stream is split into different session and the unique session ID marks the click stream. The efficiency of the standardized log file is better than traditional log file for the identification and session in WUM.

D**efinition1 the server session entity:** When a session begins, a session number that uniquely identifies the session is assigned by the server. In order to analyze user behavior, the server session entity makes up of the session identification and a series of operation entities.

Definition2 the session identification entity:[beginning | end] – session ID - user ID - time stamp. Each session has a beginning and end tag. The user ID can be the anyway to identify a user. The time stamp would record the time from the beginning to the end of the session.

Definition3 the operation entity: Session ID - Service - operation - time stamp. The Service is that the server' page had been requested (or application program called).The operation is the server' page had responded [10, 11].

Definition4 the server session constraint: the click-stream can be seen as the selection of the template nodes. A server session had been equal to the different themes of the division of the website can be seen as a different navigation template. The node is a meaningful sub sequence of the server session. There is a corresponding the server session constraint and the given node of the navigation template. is a choice for the node. The operation entity is a choice for the node [10].

For example, a server session of the purchase type would be presented. It needn't be in accordance with the order of the attraction, the resident, the purchase and the leave that is distributed in the server session sequence. What's more, the operation entity may also repeat. They can be expression by BNF (Backus-Naur Form) :< the session identification entity for begin><the pages for the purchase >< the session identification entity for end>|(< the session identification entity for begin > { < the pages for the attraction > | < the pages for the resident > | < the pages for the purchase > } < the session identification entity for end >).

3 The Ss-PrefixSpan Algorithm

The sequential pattern growth which does not need to generate candidate is one type of the frequent pattern mining methods. The PrefixSpan algorithm had first put forward the theory of the Prefix Projection sequential pattern growth. The algorithm is that. Firstly, find out the frequent items. Then, generate set of projection database .Each projection database would be related to a frequent item and be mined single-handed. Lastly, the Prefix pattern and the Suffix pattern are constantly constructed by iterated method. The definitions of Item, Prefix, Suffix and Projection had been presented in literature [9].

For example, the table1 is the input of the initial sequence and the table2 is the sequence patterns mined by the PrefixSpan algorithm.

Table 1 . the input of the initial sequence

<(b d) c b (a c)>
<(b f) (c e) b (f g)>
<(a h) (b f) a b f>
<(b e) (c e) d>
<a (b d) b c b (a d e)>

Table 2. the projected databases and sequential patterns mined by PrefixSpan

Prefix	Projected (postfix) database	Sequential patterns
<a>	<(_c)> <(bf)abf> <(bd)bcb(ade)>	<a>, <aa>, <ab>, <aba>, <abb>
<b>	<(_d)cb(ac)> <(_f)(ce)bf> <(_f)abf> <(_e)(ce)d> <(_d)bcb(ade)>	<b>, <ba>, <bb>, <bba>, <bbc>, <bbf>, <bc>, <bca>, <bcb>, <bcba>, <bcd>, <b(ce)>, <bd>, <(bd)>, <(bd)a>, <(bd)b>, <(bd)ba>, <(bd)bc>, <(bd)c>, <(bd)ca>, <(bd)cb>, <(bd)cba>, <be>, <bf>, <(bf)>, <(bf)b>, <(bf)bf>, <(bf)f>
<c>	<b(ac)> <(_e)bf> <(_e)d> <b(ade)>	<c>, <ca>, <cb>, <cd>, <(ce)>, <cba>
<d>	<cb(ac)> <bcb(ade)>	<d>, <da>, <db>, <dc>, <dba>, <dbc>, <dca>, <dcb>
<e>	<bf> <d>	<e>
<f>	<(ce)bf> <abf>	<f>, <fb>, <ff>, <fbf>

There are many sequence patterns that have been accessed by users in WUM. In that way, so many projection databases would be constructed which can cause a significant overhead. Then the initial sequence patterns and the size of candidate item sets can be reduced by the server session constraint method. The dimensions of the suffix databases that would be scanned are exponentially reduced [11-15].

Based on the above reasoning, we have the algorithm of ss-PrefixSpan as follows.

Algorithm (ss-PrefixSpan: the improved PrefixSpan algorithm based on the server session constraint)

Input: A sequence database, and the minimum support threshold *min_sup*

Output: The complete set of sequential patterns

Parameters: a: a sequential pattern; SD |a the a-projected database, if a≠<>, otherwise, the sequence database SD; ss: selection sequence by the specified constraint type.

Method:

1. Scan SD once, if meeting up ss, restructured SD'.
2. Scan SD' |a, find the set of frequent items b +such that **b** can be assembled to the last element of a to form a sequential pattern <b> can be appended to a to form a sequential pattern.
3. For each frequent item b, append it to a to form a

sequential pattern a', and output a'.

4. For each a', restructured a -projected database SD' |a'.
5. The loop executes to the step 2.

Analysis (constraint space efficiency): if *N: the length of the candidate sequences; A': the average candidate length; s: the size of the SD; s': the size of the SD'.* **Then,** the space size of the PrefixSpan : $\sum = \int(N\ A's)$, the space size of the ss-PrefixSpan : $\sum' = \int(N\ A's')$.Then the main factor of the space efficiency is$\frac{s'}{s}$. So the scale of the constraint size becomes obvious .The space efficiency is better and better.

4 Experimental Results and Performance Study

In this section, we report our experimental results on the performance of the ss-PrefixSpan algorithm. The data in experiment 1 was drawn from the Off-line Training manage Web that is developed by ourselves in our College and the programming of the log record based server session had been designed. The 600 thousand sequences had been caught. Figure 1shows the scalability of the ss-PrefixSpan algorithm with respect to the number of sequences when the support threshold is set to 0.1% and 0.2%. The experimental results show that the Sequential pattern mining based on the server session can successfully mine sequence patterns.

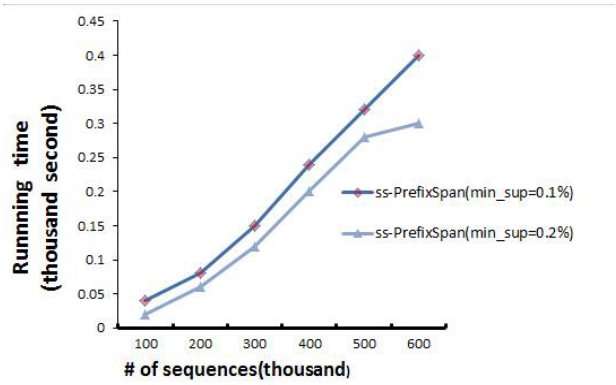

Figure 1. Scalability of ss-PrefixSpan

The experiment 2 is based on the e-commerce website mentioned in section 2.The sequence patterns of the purchase type would be mined based the server session from the simulation data. The service operations are represented by three bit integers which are not more than 4 at the highest level. The highest level of the integer represented the type of the operation entity. The For example, the number of 32 belongs to the type of the purchase. The number of 45 belongs to the type of the leave. The number of 123 belongs to the type of the attraction. The number of 2 belongs to the type of the resident. There had generated 100 thousands session identification entities and 80 thousands operation entities. The experiment had selected the execution time as the main means to compare the performance of the ss-PrefixSpan algorithm, the PrefixSpan algorithm, the GSP algorithm and the SPADE algorithm. The experimental results show the superiority of the ss-PrefixSpan algorithm.

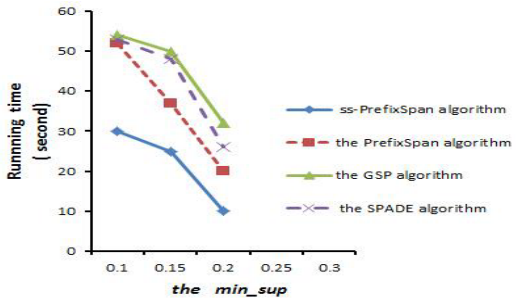

Figure 2. The execution time at the different support

Summary

In this paper, we have developed a novel, scalable and efficient record log in a renovation of web technologies that changed rapidly. This log format can better record users' access behaviors under the traditional Web technology and The Web services technology. Then the ss-PrefixSpan algorithm is discussed for the logs based on the server session. Finally, the validity and superiority of the ss-PrefixSpan algorithm are presented by the two experiments which provide a reliable basis.

Acknowledgements

This work is supported by the National Science Foundation of China Grant #61562008 and the scientific research project of the Guangxi Education Department Grant #ZD2014083 and #KY2015LX588. YUAN Chang-an is the corresponding author (e-mail: yca@gxtc.edu.cn).

References

1. Yin J,ZhengZ,CaoL.Uspan:an efficient algorithm for mining high utility sequential patterns. Proceedings of the 18th ACM SIGKDD .2012.
2. J.Yin,Z.Zheng,L.Cao,Y.Song,W.Wei. Efficiently mining Top-k high utility sequential patterns.2013 IEEE 13th International Conference on Data Mining (ICDM 2013) .2013.
3. Wang Le, Wang shui, Liu Sheng-lan, Wang Hui-bing. An algorithm of Mining Sequential pattern with wildcards based on Index-Tree[J]. Chinese Journal of Computers.2016.11, 39(178).
4. Wu Xindong, XieFei, Huang Yongming, Hu Xuegang, Gao Jun. Mining sequential patterns with wildcards and One-Off conditions[J]. Journal OfSoftware.2013, 24(8).
5. CHENG Si-Yuan, MA Chao, LI Cong-Cong. High Utility Sequential Pattern Mining Algorithm Based on MapReduce .Computer Systems & Applications.2015, 24(12).
6. Zaki M J . SPADE: an efficient algorithm for

mining frequent sequences[J] . Machine Learning,2001, 42 (1/ 2) : 31-60.
7. CeddiaJ ,Sheard J , Tibbey G. WAT :a tool for classifying learning activities from a log file[C]/ / Proceedings of the 9thAustralasian Conference on Computing Education . Darling Hurst: Australian Computer Society, 2007.
8. Liang Q A, Miller S, Chung J . Service mining for Web service composition[C]/ / IEEE International Conference on In Formation Reuse and Integration. Las Vegas,Nevada, 2005.
9. Han J W, Kamber M. Data Mining: Concept s and Techniques [M]. The 2nd editor.San Francisco: Morgan Kaufmann Publishers, 2006.
10. Asbagh M J,Abolhassani H. Web service usage mining: mining for executable sequences [C]/ / Proceedings of the 7th WSEAS International Conference on Applied Computer Science .Wisconsin : World Scientific and Engineering Academy and Society ,2007.
11. Lin M Y ,Hsueh S C , Chang C W. Fast discovery of sequential patterns in large databases using effective time-indexing[J] . Information Sciences , 2008 ,178 (22) :4228-4245.
12. SilvestriC, Orlando S. Approximate mining off frequent patterns on streams [J]. Intelligent Data Analysis, 2007 ,11 (1).
13. Mulvenna M D ,Anand S S ,Büchner A G. Personalization on the net using Web mining [J]. Communications of ACM, 2000, 43 (8):1222-1251.
14. CAI Hong-guo, YUAN Chang-an, etc. Server Session Constraint-based Serial Pattern Growth Mining Research .The journey of Zhengzhou University.2010, 42(1).
15. J Pei,J Han,B Mortazavi-Asl,H Pinto,Q Chen. PrefixSpan: Mining Sequential Patterns Efficiently by Prefix-Projected Pattern Growth. International Conference on Data Engineering , 2001:215-224.