PROCEEDINGS OF SCIENCE

Frequent Sequence Mining from Massive Access Log for User's Behaviour Investigation

Wei Chen[1]
Wuhan Digital Engineering Institute
Wuhan, 430074, China
*E-mail :*772382203@qq.com

Yan Tong[2]
Wuhan Digital Engineering Institute
Wuhan, 430074, China
*E-mail :*tongyan.cherish@139.com

Jian Zhang[3]
Wuhan Digital Engineering Institute
Wuhan, 430074, China
*E-mail :*richardxx@126.com

Tao Qin
Xi'an Jiaotong University
Xi'an, 710049, China
*E-mail :*qin.tao@mail.xjtu.edu.cn

With the fast development of Web 2.0, users can obtain everything that they want from the Web. and their access behaviours are recorded by the access log. Based on mining the frequent access sequence, we can deeply understand their access interests. In turn, it can improve the efficiency of network management. In this paper, we firstly present the methods for log pre-processing and extract the features. Secondly, we employ the PrefixSpan algorithm to achieve the goal of frequent sequences mining. In order to process the massive log data in network today, we also combined the proposed methods with Spark. Finally, experimental results based on the log data collected from the campus network of Xi'an Jiaotong University verify the efficiency of the developed methods, which are useful for the understanding and management of the user's behaviour.

CENet2017
22-23 July, 2017
Shanghai, China

[1] Speaker
[2] Correspongding Author
[3] The research presented in this paper is supported in part by the Natural Science Foundation of China (61502438,61672026), Natural Science Foundation of Shaanxi Province (2016JM6040),and Chinese Defense Advance Research Program (B0820132036)

1. Introduction

The fast development of Internet has greatly benefited our daily life. We can get everything that we want by different Web services, for example, we can buy whatever we want from online business sites, e.g. amazon. We can also make friends through online social networks. In the course of using those Web services, our behaviors are recorded by the access log,. Based on mining the frequent access patterns, the administrator can understand the user's behavior and infer their interests. Those results are very important for management of their behavior management. In this paper, we proposed a method for massive access log analysis based on Spark and PrefixSpan.

In the past several years, study on the characteristic of access log has attracted great attentions. In Ref.[1], the authors discussed the possibilities of identifying the documents in users' navigation paths and optimized the web structure. In Ref.[2], the authors presented a personalization recommendation model to recommend potentially interesting resources to users based on the web access log of users. In Ref.[3], the authors proposed a recommendation approach that recommended a listing of pages based mostly upon the client's historic pattern. In Ref. [4], an algorithm was developed for mining user interesting/preferred access patterns from Frequent Link and Access Tree. In Ref.[5], the authors combined tree projection and prefix growth features together to reduce execution time. Based on the above researches, we can find that as the web access log records the user access behavior, many researchers focus on mining the frequent access sequence of different users from web access log, which is useful to optimize the topology of the website, and the website manager can also better understand the user's access interests so as to provide personalized services to them. But with the fastdevelopment of Internet, how to deal with the massive access log generated by the huge number of users is still an open question.

In order to realize the mining of user interest and deal with the massive log data, in this paper, we proposed a method for mining frequent sequence by using PrefixSpan. The analysis methods are also combined with Spark to increase the analysis efficiency. The analysis results based on the web logs collected by the real servers in Xi'an Jiaotong University show that the proposed method can accurately obtain the user's frequent access pattern, in turn understand the user's behaviors and mine the interest.

2.Data Processing

2.1 Introduction to Web Log

Web access log records the user access behavior detailedly. We selected the example from our data set and it is listed as follows:

204.108.127.28 - - [13/Oct/2015:07:45:13 +0800] "GET /people/zhwu/files/2010/12/deepWeb1.jpg HTTP/1.1" 200 64948 "https://www.google.com/" "Mozilla/5.0 (iPhone; CPU iPhone OS 9_0_1 like Mac OS X) AppleWebKit/601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13A404 Safari/601.1"

The above log record indicated that at 12:15:13 on October 13, 2015, the user 204.108.127.28 successfully requested the resource /people/zhwu/files/2010/deepWeb1.jpg at https://www.goog le.com/, and the server returned 64948 bytes. Browser used Mozilla / 5.0

(iPhone; CPU iPhone OS 9_0_1 like Mac OS X) AppleWebKit /601.1.46 (KHTML, like Gecko) Version/9.0 Mobile/13A404 Safari/601.1.

The detailed meaning of each field in a log record from Apache server access log is shown in Table 1:

Field	Meaning
IP	The user IP accessing the server
Client	The client issuing the HTTP request
UserID	User identification
Time	The timestamp of the Web request being completed by server
Method/ Resource /Protocol	The request method/requested URL/protocol of client
Status	The status code of server response
Size	The size of data returning to client
Referrer	The source address of request
Agent	User agent

Table 1: Apache Access Log Format

2.2Data Collection

The log collection and analysis platform is shown in Figure 1. The platform collects system log from more than 50 Web servers and 1 manual target server in the campus network of Xi'an Jiaotong University. The web servers are selected from the campus network and the manual target server is used to generate labeled logs. The raw web log preprocessing includes data cleaning, user identification, session recognition and path supplementation [6]. As the data set in this paper are collected from the school's server and are not suitable for studying the behavior of user accessing a specific website, and the path supplementation needs to know the specific topology of the site, the data preprocessing in this paper does not involve the path supplementation.

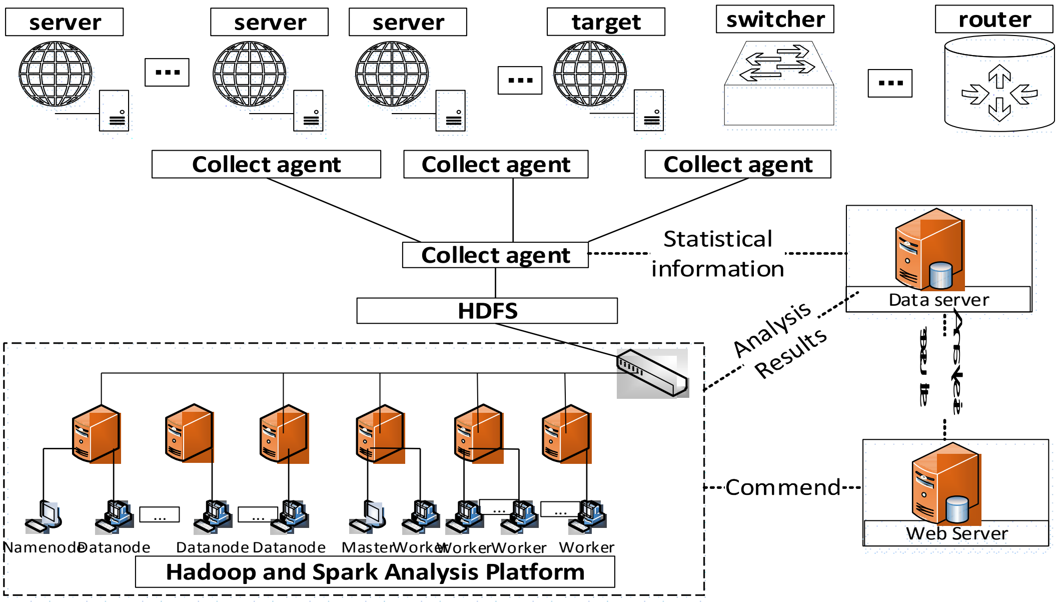


Figure 1: The Analysis Platform of Massive Logs

2.3Data Preprocessing

2.3.1Data Cleaning

Data cleaning mainly merges some of the relevant records and deletes the references between objects not important or unrelated to mining analysis (including text files, graphics, and sound files). The data cleaning process is given as follows:

Step 1: the filter out the access log of which the request method is GET and the status code of server response is 200. The GET method retrieves the data from the specified server. The status code 200 indicates that the request is successfully accepted by server.

Step 2: delete the record not required. When a user requests a page, the information about the page is requested and downloaded, such as audio files, pictures and scripts. These requests are also recorded in the access log. Butthis algorithm is to find out the user's frequent access pattern, so these unrelated records should be deleted.

Step 3: delete the records where source URL of request is "-". The source URL "-" means the user enters the requested page directly instead of clicking the link to get the page.

Step 4: delete the useless attribute. Usually it is necessary to extract useful information such as the user IP, the source address of the request and the timestamp of the request from the log records. Other information is deleted.

2.3.2User Identification

To improve the efficiency of the frequent sequence mining, it is necessary to identify that the access log records are generated by which user. There may be several reasons for making this task very difficult. Firstly, different users may access the web server through the same proxy server at the same time. Secondly, the same user may access the web server on different machines with different IPs. Thirdly, the same user may use different browsers to access the web server on a machine; and the same machine may also be used by different users to access web server. Thus we identify the users by using the following methods:

Step 1: if the IP addresses of the access log records are different, it is considered that different users access the web server.

Step 2: if the IP addresses of the access log records are the same, the users can be distinguished by different browsers or operating system.

Step 3: if the IP addresses of the access log records are the same, and the browser and the operating system are the same, the topology of the website will be used to identify the user.

2.3.3Session Recognition

After the user identification, the access sequence of each user IP can be obtained. In order to mine the practical meaning of the access sequence, the session recognition is needed, that is to say, the entire sequence should be split into different sessions.

In the access log, if the page is accessed by different users, it is considered as a different session. However, if the same user accesses a page on a large time window, it is considered that the user has visited the page for many times and the user's visit sequence can be divided into multiple sessions. The user session is defined as: the user session *S* is a two-tuple which consists of a user identity (*uid*) and a page collection of user access in a period of time (*RS*), namely <*uid, RS*>. *RS* is composed of the page identifier (*pid*) and the requested time. Therefore, *S* can be expressed as Formula (2-1):

$$S = \langle uid, (pid1, time1), ..., (pidn, timen) \rangle \quad (2.1)$$

Where *uid* is the user IP and *pid* is the source URL of request. Usually the time-out method is used for identification of the user session. There are two time-out methods:

Step 1: given the time threshold for the entire session, a session in Formula (2-1) needs to satisfy the following formula:

$$time(k) - time(1) \leqslant T \quad (2.2)$$

Where T is the threshold of the time difference between two adjacent records, the threshold is often given as 30 minutes according to the experience. If the subsequent access happens within T from the first access of the session, it is considered that the subsequent visits also belong to this session.

Step2: given the time threshold for accessing adjacent pages. The session recognition method used in this paper is giving a time-out value for two adjacent access records:

$$time(k) - time(k-1) \leqslant T \quad (2.3)$$

Where T is the threshold of the time difference between the two adjacent records, and the threshold is often given as 10 minutes according to experience. If the time difference between the two adjacent access records is less than the threshold T, the two access records can be considered belonging to the same session. According to the temporal characteristics of the collected data and the experiment, the algorithm sets the threshold T to 5 minutes.

3.Mining Frequent Sequence of User Access based on PrefixSpan

3.1PrefixSpan Algorithm

The concept of sequence pattern was firstly proposed by Agrawal and Srikant, and the sequence pattern analysis is designed to find the orderly correlation between events. The PrefixSpan algorithm is an algorithm for sequence pattern mining in 2004. It uses the idea of dividision and conquer, constantly generating multiple smaller sequence, and then carries out sequence pattern mining on each projection database [7]. PrefixSpan can greatly reduce the generation of candidate sub-sequences, but also greatly reduce the space occupied by the projection database [8]. The concepts involved in the algorithm are as follows:

Step 1: Item set: an item set is a non-empty set of items, and can be expressed as $X = (x_1, x_2 ... x_m)$. x_i $(1 \leq i \leq m)$ is an element of the set.

Step 2: Sequence: a sequence is the ordered arrangement of item sets, which can be expressed as $S = <s_1\ s_2 ... s_n>$, where s_i $(1 \leq i \leq n)$ is an item set and an element of the sequence.

Step 3: Prefix: Prefix: suppose that all items in each element of a sequence be sorted by lexicographical order. Given the sequence $\alpha = <e_1\ e_2 ... e_n>$, $\beta = <e_1'\ e_2' ... e_m'>$ $(m \leq n)$, if $e_i' = e_i$ $(i \leq m-1)$, $e_m' \subseteq e_m$, and all the items in $(e_m - e_m')$ are after the items in e_m' in order, then β is defined as the prefix of α.

Step 4: Suffix: the projection of sequence α on its subsequence β is $\alpha' = <e_1 e_2 ... e_n>$ $(n \geq m)$, then the suffix of α on β is $<e_m\ e_{m+1} ... e_n>$, where $e_m'' = (e_m - e_m')$.

Step 5: Projection: given the sequence α and β, if β is a subsequence of α, then the projection α' of α on β must satisfy: β is the prefix of α', α' is the largest subsequence satisfying the above condition of α.

Step 6: the projection database: let α be a sequence pattern in the sequence database S, then the α-projection database be expressed as: S | α, which is the set of projections for the prefix α in S

Step7: support the projection database: let α be a sequence pattern in the sequence database S, and β is a sequence with a prefix α. The support of β in the α-projection database S | α is expressed as S | α (β), that is, the frequency of β in S | α.

3.2Combined with Spark

Spark is a generic and open source parallel framework designed by UC Berkeley AMP lab, which is similar to Hadoop MapReduce and features the advantages of Hadoop MapReduce. But its job output can be saved in memory, superior to MapReduce which can only read and write HDFS. In this paper, the data preprocessing is implemented by Spark, and the PrefixSpan algorithm is also selected from the Spark machine learning library MLlib [9]. Based on the network management experience, the minimum support for PrefixSpan is selected as 0.5.

4.Result Analysis of Mining Frequent Sequence of User Access

4.1Data Set

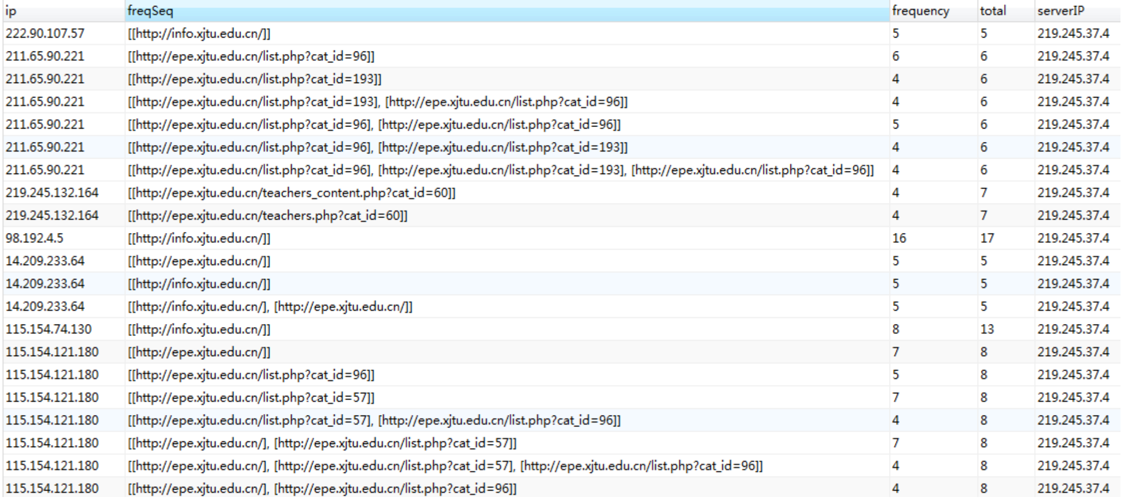
The test data were collected from access logs stored in more than 40 web servers of Xi'an Jiaotong University campus network. The data were stored on the HDFS platform of Xi'an Jiaotong University Network Center. The access logs were stored in the server with IP 219.245.37.4 from October 29 to November 29 were selected, the size of the test data is 120M.

4.2Experiment Result

Part of the result is shown in Figure 2. The result surrounded by a red rectangle is selected as example, and some useful web access patterns of the user with IP 115.154.121.180 can be found. This IP generated a total of eight sessions, and there are seven frequent sequences. The most frequent URL access sequences are [[http://epe.xjtu.edu.cn/list.php?cat_id=57], [[http://epe.xjtu.edu.cn/]] and [http://epe.xjtu.edu.cn/list.php?cat_id=57]]. The support count of these sequences is 7. Through the practically visit of the three URL sequences, it can be found that the user with IP 115.154.121.180 frequently accessed the home page (http://epe.xjtu.edu.cn/) and the bulletin page (http://epe.xjtu.edu.cn/list.php?cat_id=57) of School of Energy and Power of Xi'an Jiaotong University. The access of the two pages belong to the same session. The users always visit the home page at first, and then click the link of the bulletin page and visit that page, indicating the users are always concerned about the announcement of School of Energy and Power. The whole mining process uses only 11 seconds, which is very efficiency.

5.Conclusion

In this paper, we firstly introduce a method for access preprocessing, including the data cleaning and session congestion to obtain the data for frequent sequence mining. Secondly, the frequent sequence pattern mining algorithm based on PrefixSpan is designed, and HDFS and Spark are used to improve the calculation efficiency. Finally, the proposed methods are tested based on the access log collected from Xi'an Jiaotong University, and the mining results verify the efficiency and correctness of the developed method. In the future work, we will mainly focus on designing a better way to identify the session and users.



ip	freqSeq	frequency	total	serverIP
222.90.107.57	[[http://info.xjtu.edu.cn/]]	5	5	219.245.37.4
211.65.90.221	[[http://epe.xjtu.edu.cn/list.php?cat_id=96]]	6	6	219.245.37.4
211.65.90.221	[[http://epe.xjtu.edu.cn/list.php?cat_id=193]]	4	6	219.245.37.4
211.65.90.221	[[http://epe.xjtu.edu.cn/list.php?cat_id=193], [http://epe.xjtu.edu.cn/list.php?cat_id=96]]	4	6	219.245.37.4
211.65.90.221	[[http://epe.xjtu.edu.cn/list.php?cat_id=96], [http://epe.xjtu.edu.cn/list.php?cat_id=96]]	5	6	219.245.37.4
211.65.90.221	[[http://epe.xjtu.edu.cn/list.php?cat_id=96], [http://epe.xjtu.edu.cn/list.php?cat_id=193]]	4	6	219.245.37.4
211.65.90.221	[[http://epe.xjtu.edu.cn/list.php?cat_id=96], [http://epe.xjtu.edu.cn/list.php?cat_id=193], [http://epe.xjtu.edu.cn/list.php?cat_id=96]]	4	6	219.245.37.4
219.245.132.164	[[http://epe.xjtu.edu.cn/teachers_content.php?cat_id=60]]	4	7	219.245.37.4
219.245.132.164	[[http://epe.xjtu.edu.cn/teachers.php?cat_id=60]]	4	7	219.245.37.4
98.192.4.5	[[http://info.xjtu.edu.cn/]]	16	17	219.245.37.4
14.209.233.64	[[http://epe.xjtu.edu.cn/]]	5	5	219.245.37.4
14.209.233.64	[[http://info.xjtu.edu.cn/]]	5	5	219.245.37.4
14.209.233.64	[[http://info.xjtu.edu.cn/], [http://epe.xjtu.edu.cn/]]	5	5	219.245.37.4
115.154.74.130	[[http://info.xjtu.edu.cn/]]	8	13	219.245.37.4
115.154.121.180	[[http://epe.xjtu.edu.cn/]]	7	8	219.245.37.4
115.154.121.180	[[http://epe.xjtu.edu.cn/list.php?cat_id=96]]	5	8	219.245.37.4
115.154.121.180	[[http://epe.xjtu.edu.cn/list.php?cat_id=57]]	7	8	219.245.37.4
115.154.121.180	[[http://epe.xjtu.edu.cn/list.php?cat_id=57], [http://epe.xjtu.edu.cn/list.php?cat_id=96]]	4	8	219.245.37.4
115.154.121.180	[[http://epe.xjtu.edu.cn/], [http://epe.xjtu.edu.cn/list.php?cat_id=57]]	7	8	219.245.37.4
115.154.121.180	[[http://epe.xjtu.edu.cn/], [http://epe.xjtu.edu.cn/list.php?cat_id=57], [http://epe.xjtu.edu.cn/list.php?cat_id=96]]	4	8	219.245.37.4
115.154.121.180	[[http://epe.xjtu.edu.cn/], [http://epe.xjtu.edu.cn/list.php?cat_id=96]]	4	8	219.245.37.4

Figure 2 : Part of the Result of Mining Frequent Sequence of User Access

References

[1] Z. Eremic, D. Radosav, B. Markoski.*Mining user access logs to optimize navigational structure of adaptive web sites*[C]. 11th International Symposium on Computational Intelligence and Informatics.18-20(2010)

[2] X. Peng, Y. Cao, Z. Niu. *Mining Web Access Log for the Personalization Recommendation*[C]. International Conference on MultiMedia and Information Technology.30-31(2008)

[3] D. Sahu, R. Soni, *A New Method for Detecting Users Behavior from Web Access Logs*[C]. International Conference on Computational Intelligence and Communication Networks.12-14(2015)

[4] C. Chen, *Discovery of user preferred access patterns from web logs*[C].Eighth International Conference on Fuzzy Systems and Knowledge Discovery.26-28(2011)

[5] S. Vijayalakshmi, V. Mohan, M. S. Sassirekha, O. R. Deepika. *Extracting Sequential Access Pattern from Pre-Processed Web Logs*[C]. International Conference on Process Automation, Control and Computing.20-22(2011)

[6] L. Lu, Y. Yang, X. Guan, H. Wei, *Research on data preprocessing in web log mining* [J].in proceedings of Computer Engineering. 26(4),66-67(2000)

[7] J Pei, J Han, B Mortazavi-Asl, *Mining Sequential Patterns by Pattern-Growth: The PrefixSpan Approach* [J]. IEEE Transactions on Knowledge and Data Engineering (Impact Factor: 2.07). 16(11),1424-1440(2004)

[8] J. He, *Sequential pattern mining based on web log and its application in E-commerce* [D]. Tianjin: Tianjin University(2008)

[9] *Frequent Pattern Mining - RDD-based API*[EB/OL]. http://spark.apache.org/docs/latest/mllib-frequent-pattern-mining.html#prefix-span.