EAI Endorsed Transactions
on Security and Safety
Research Article EAI.EU

User Identity Linkage Method Based on User Online Habit

Yan Liu[1,*]

[1]Information Engineering University

Abstract

User Identity Linkage (UIL) across social networks refers to the recognition of the accounts belonging to the same individual among multiple social network platforms. Due to the network user's identities have the characteristics of various sources and real identity cannot be confirmed, it is very easy to become the main means of malicious user to carry out network attacks and spread rumors. User Identity Linkage not only can make the service provider to understand the user and thus to provide better service to the user, but also plays a significant role in improving the ability to find and track malicious users. For the credibility problem on the associated clues of user identification resulted from dynamic IP, shared Internet access and other factors, a user identity linkage method based on user online habit is proposed. This method assumes that the people use multiple network services crosswise when using the internet, converts the association analysis problem of user identification to the frequent pattern mining problem, and performs the optimization from three respective aspects: the online transaction database construction, the fast algorithm for mining frequent patterns and frequent co-occurrence identities consolidation. In order to improve the efficiency of frequent pattern mining, a parallelization of FPGrowth algorithm called MRFP-Growth algorithm is proposed to mine the user identifications of frequent co-occurrence quickly and efficiently. Experiments show that this method can associate multiple accounts of a user in network traffic with more than 85% accuracy in the scenario of dynamic variable IP address with only IP address and online time.

Keywords: Network Traffic Analysis, User Identity Linkage, Frequent Pattern Mining, FP-Growth.

Received on 20 August 2020, accepted on 21 October 2020, published on 30 October 2020

doi: 10.4108/eai.22-6-2021.170240

* Corresponding author. Email: ms.liuyan@foxmail.com

1 Introduction

With the rapid development of the Internet and the gradual popularity of mobile devices, online social networks have become more and more popular, bringing great convenience to the communication between people. Different social networks provide different types of services. People usually join into different social networks according to the needs of work and life. Social networks have become a bridge connecting virtual cyberspace and the real physical world [1][2]. Network has anonymity, most network service providers protect the privacy of users, allowing users to freely choose the user name to replace their own real identity mark. In addition, because of the single sign-on between different network services, the way users use network services is diverse, the same user may register multiple user names, and the same user name may also be used by multiple users. Therefore, the phenomenon of multi-representation of users leads to the fragmentation of user identity information.

User identity linkage technology is to discriminate whether the network user information from different social media refers to the same social person, to find the association clue by mining the hidden knowledge in the user attributes, the relationships, the behavior data, and to build the data bridge for the user portrait of the crosssocial media, also known as anchor linking prediction problem [3], network alignment problem [4]. The research of user identity linkage problem plays an important role in multiple internet applications, such as cross-platform friend recommendation [5][6], user behavior prediction [7], cross-network information dissemination [8], etc.

Some existing ways of user identity linkage are mainly by calculating the similarity of attributes and behaviors

between these accounts obtained from the Web[9-12]. However, more often than not, this approach relies heavily on account information in Web data resources, and its essence is still based on the characteristic similarity between accounts. It has so far been a difficult task to distinguish between the accounts of different users in large-scale, sparse Web data. In fact, a lot of user account behavior information is not directly available through the Web. There is rich information contained in the network traffic data. For example, when a user only logs in to browse information on a social network platform without doing other explicit actions (such as posting), this login, browsing behavior is not reflected in Web data. However, from the network traffic data, each operation of the user, will form one or more packets, these packets of cookies and other information carry a large number of user account correlation. Moreover, the user's operation on different network service platforms in a short time will be carried into the network traffic data in this form, which makes those apparently unrelated data have a certain correlation in this period of time. Therefore, unlike the data carried by the Web, the network traffic data carries a lot of valuable hidden association information.

In recent years, more and more researchers have been focusing on the number of traffic on the Internet as the issue of privacy leaks has intensified. On the one hand, many researchers in the field of security have studied the privacy disclosure of user cookies in network traffic data, trying to restore user behavior and host tracing by restoring cleared cookies records[13][14]. On the other hand, some researchers try to build a fingerprint model based on browser features, starting with browser information in web traffic data[15][16], so as to achieve the tracking and determination of the host. From these studies, these network traffic data not only carry a lot of hidden information that cannot be obtained in the Web, but also have natural timing, association. However, the existence of dynamic IP, NAT, proxy server and other technologies in the network have a great impact on these studies, which makes the traditional spatiotemporal association mining methods[17-19]It is difficult to effectively divide appropriate space-time fragments for users to carry out association analysis.

In this paper, the relationship between user account and time and space (IP) is analyzed based on the data of network traffic, and the linkage problem of user identity is transformed into frequent pattern mining problem. A network user identity linkage method based on the user online habits is proposed.

2 Problem Description

2.1 Analysis of User's Online Habits

Before the problem description, we first assume that the user's online behavior is universal, that is, the user cross-uses a variety of network services when engaged in network activities. This behavior habit is a long-term and frequent behavior pattern of the user in the whole process of Internet access.

Based on the above assumption, we select the Internet traffic of campus network for statistics. The traffic spans five months, including 3980 users, 33 social applications, 12952 IP addresses, a total of 12289 accounts.

During the statistical period, the following table shows the statistical results of users using multiple accounts.

Table 1. Number of users using accounts

Number of accounts	1	2	3	4	5	6	7	8	9	10
Number of users	868	530	425	380	333	306	309	292	266	269

From the statistical results, the traffic of users using multiple platforms is very common in the statistical time, accounting for 78.2% of our test data. The statistical results verify that users use multiple user IDs is common.

One of the known users is further analyzed who use three commonly used social apps in nearly three hours, namely the shopping site (www.taobao.com), the instant messaging software (qq) and the e-mail (@126.com), is presented below.

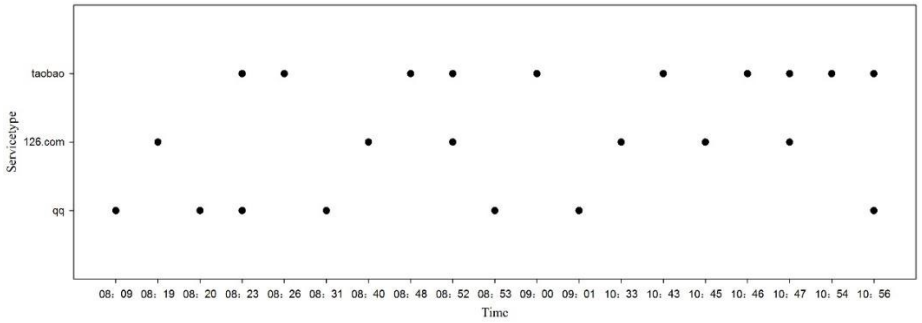


Fig. 1. Online activity diagram of a user for a certain period of time

The "black dot" in Fig.1 indicates that at a certain point in time, the corresponding network application accessed by the user is found in the traffic. As can be seen from Fig.1, the user cross-uses three network services when engaging in network activities. If the online activity of the user during that period is divided by a certain length of time slice (such as 1 hour), the partitioning results shown in Fig.2 can be obtained.

As shown in Fig.2, the phenomenon of cross-using multiple network services is still reflected on each time slice.

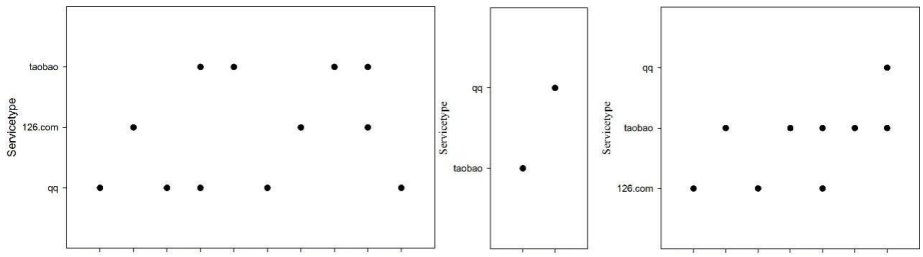


Fig. 2. Result chart of user online activities by time slice

Further statistics on the user's cross-access to the network can reach the co-occurrence results shown in the table below:

Table 2. Cross-use statistics for multiple network identifiers

Serial number	Network identity collection	Number of appearances
1	{qq,126.com}	2
2	{qq,taobao}	3
3	{126.com,taobao}	2
4	{qq,126.com,taobao}	2

It can be seen from Table 2 that the correlation of multiple user identities in network traffic data is the phenomenon of frequent co-occurrence of network accounts.

2.2 Related Concepts

The formal definition of the problem is given below.

Internet Access Record: Internet access record refers to the network traffic data parsed out of the network address, time stamp, as well as the user identification and service type of information carried by the record, recorded as Z = (IP, time,userid,servicetype).

Time Slice: Time slice refers to a time slice, also known as a time region, recorded as *TIME*. The traffic record within a certain time piece is recorded as Z' = (*IP,TIME, userid,servicetype*)

Internet Transaction: Online transactions record the online records of a period of time on a single IP after de duplication. It is recorded as

$T_{(IP+TIME)}=\{(userid+servicetype)_1,...,(userid+servicetype)_m\}$

Internet transaction database: Internet transaction database is a collection of all Internet transactions which is recorded as $D=\{T_1, T_2,...,T_N\}$.

Itemset: An itemset is a collection of one or more items in an association analysis, denoted as X. If an item set contains k items, it is called k -item set.

Support: The support of an item set X is the number of transactions in an online transaction database that contain the set X. It is denoted as sup(X), $\sup(X)=|\{T_i \mid X \subset T_i, T_i \in D\}|$

Minimum Support: Minimum Support is a threshold for support during frequent pattern mining. It is denoted as min_sup .

Frequent Itemset: If a set X satisfies sup(X) ≥ min_sup in an online transaction database D , it is called a frequent itemset.

Problem Description: Under the premise of user's online habits, given the online record set Z , build a collection D of user IDs belong to a single user, and use some association rule mining algorithm to analyze the association, so as to get the different network identification of a group of target users.

This paper uses frequent pattern mining method to solve the problem of user identity linkage. We give a research framework of user identity linkage method based on the user's online habits. The structure of the framework is shown in Fig 3:

2.3 Research Framework

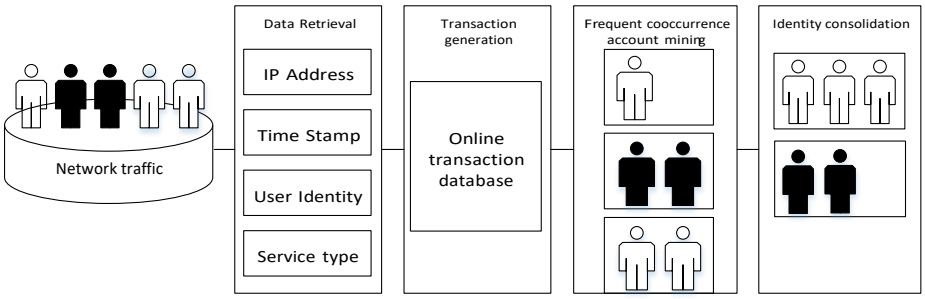


Fig. 3. Research framework of user identity linkage method based on the user's online habits

It can be seen from Fig. 3 that the framework is mainly divided into four parts: data retrieval, transaction generation, frequent co-occurrence account mining and account consolidation.

(i) Data retrieval. Through network programming, the information of source IP, time stamp, user identification and service type carried in the packet is extracted from the original network traffic data.
(ii) Transaction generation. According to the assumption in this paper, select the appropriate time slice as relatively stable period, pool the user identity hosted by the same IP address in that time period to generate the transaction set required for frequent pattern mining.
(iii) Frequent co-occurrence account mining. The appropriate frequent pattern mining algorithm is selected to correlate the frequently co-occurrence user identification. the comparison of commonly used frequent pattern mining algorithms is given in reference [20].
(iv) Account consolidation. The purpose of consolidation is to merge the identities of the same user scattered in multiple frequent itemsets.

2.4 Key Issues

In order to find out multiple identities of one user, the following key problems need to be solved:

(i) How to construct an online transaction database. The online transaction database constructed in this paper should have the characteristic of simulating the frequent cross-use of multiple network services on dynamically changing network addresses.
(ii) How to quickly and effectively mine frequent models from massive amounts of data. In the analysis and processing of big data, the traditional frequent pattern mining methods have some limitations, and often have problems such as insufficient memory space and excessive running time. How to quickly and effectively mine frequent patterns from massive data is one of the key problems to be solved in this paper.
(iii) How to merge the frequently mined itemsets correctly. Not all frequent itemsets can fully express the relationship between multiple identities belonging to the same user. How to merge the frequent itemsets and aggregate the user identities scattered in different frequent itemsets to generate the final association results is one of the key problems to be solved in this paper.

3 Online Transaction Database Construction

The online transaction database required for frequent pattern mining is generally composed of a group of records, where each record represents an online transaction, and each transaction corresponds to a unique transaction identification number and column of items. Compared with the construction of shopping cart transaction database, the construction process of online transaction database in this paper is more complicated.

Table 3 shows the schematic diagram of online transaction database.

Table 3. Online transaction database presentation

TID	Itemsets
$TIME_1+IP_1$	$\{(userid+servicetype)_1, \ldots, (userid+servicetype)_m\}$
...	...
$TIME_1+IP_i$	...
...	...
$TIME_k+IP_1$	...
...	...
$TIME_k+IP_j$	$\{(userid+servicetype)_1, \ldots, (userid+servicetype)_n\}$

There are k time slices, $TIME_1$~ $TIME_k$. The entire network data is divided into k sections according to the time stamp information. There are i network addresses that appear in the time slice $TIME_1$, IP_1~IP_j, and there are j network addresses that appear in the time slice $TIME_k$;

In a transaction $TIME_k+IP_j$, $\{(userid+servicetype)_1,\ldots,(userid+servicetype)_m\}$ is the collection of all user identities and service types hosted on IP_j in the time slice $TIME_k$;

Reasonable time slice partition has great influence on correlation results. If the time slice partition is too long, then the items contained in a transaction may belong to multiple users, resulting in items in the frequent item set may still belong to multiple users; if the time slice partition is too short, the number of items contained in a transaction will be few, resulting in the loss of association clues. How to determine the reasonable size of the time slice, through a large number of experiments, according to the specific analysis of different scenes.

4 User Identity Linkage Based on Parallel Frequent Pattern Mining

In the research of this paper, how to find the frequent co-occurrence user identities from the online transaction database is a key question to be solved. Therefore, by analyzing the traditional FP-Growth algorithm, this paper presents a parallel FPGrowth algorithm based on MapReduce, MRFP-Growth, which is suitable for this research.

The traditional FP-Growth algorithm, because of its frequent process of building trees, has some problems such as low efficiency and insufficient memory when processing large-scale data. In this paper, a parallel FP-Growth algorithm based on MapReduce is presented, which is based on traditional FP-Growth algorithm.

MapReduce is a computing model, framework and platform for parallel processing of big data. MapReduce consists of two parts:

(1) Map is responsible for the task decomposition.
(2) Reduce is responsible for the summary of the results.

The following combines the parallelization principle of MapReduce and the basic idea of the traditional FP-Growth algorithm to give the basic idea and algorithm flow of the MRFP-Growth algorithm to solve the problem in this paper. The MRFPGrowth algorithm mainly realizes the hierarchical mining of the traditional FPGrowth algorithm, and then realizes parallelization and improves the operation efficiency. The MRFP-Growth algorithm flow chart is shown in Fig. 4:

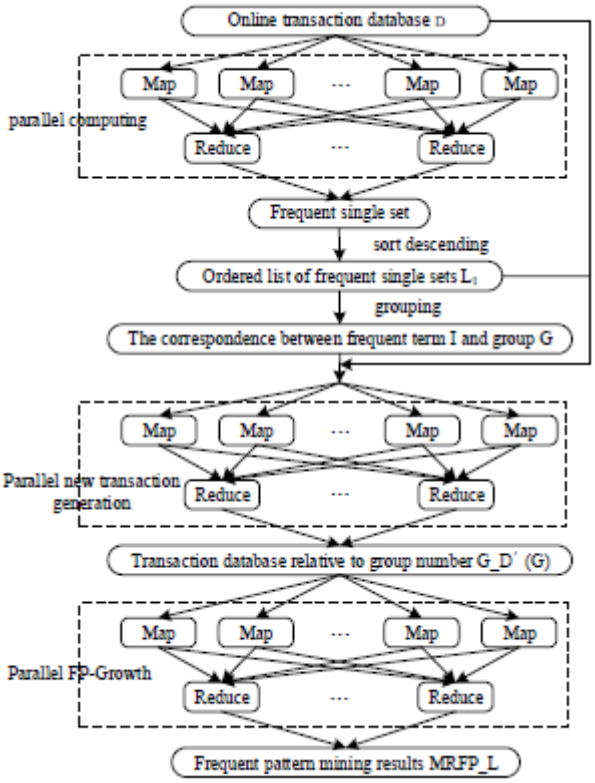


Fig. 4. MRFP-Growth algorithm flowchart

As can be seen from figure 4, the MRFP-Growth algorithm realizes the function of frequent pattern mining through three Maps processes and a Reduce process.

(1) Generate Orderly List of Frequent Single Itemsets L_1

According to the number of Map nodes in the cluster, the online transaction database is divided into equal parts, and then the corresponding Map node is given for parallel counting, and then all the counting results are given to the corresponding Reduce node for statistics. According to the pre-set threshold, the frequent single set is generated, and then the frequent single set is arranged in descending

order according to its support value, and the orderly single set list L_1 is obtained.

(2) Construct the Correspondence between Frequent Term I and Group G

According to the number of Map nodes in the cluster, the frequent items in the ordered frequent single item list are divided into groups. According to the order of the items in the list L_1 and the number of items in each group, the frequent items are divided into N groups. Then we get a correspondence between the frequent item I and the group G . The process is shown in Fig. 5.

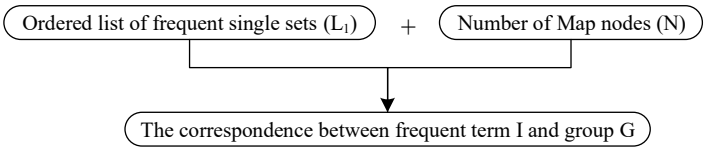


Fig. 5. Flowchart of the correspondence between frequent items and groups

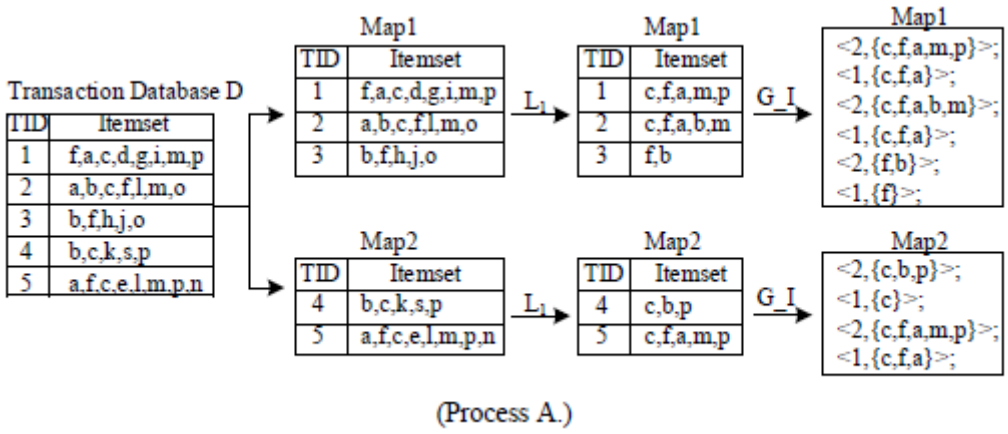
(3) Generate Transaction Databases $G\_D'(G)$ Corresponding to Group G

Based on the number of Map nodes in the cluster, the transaction database is divided equally and given to the corresponding Map node, and then, according to the ordered list of frequent single items L_1 , the items in each transaction T in the Map node are sorted, and the infrequent items are deleted to generate the transaction T'.

Then, according to the corresponding relation between frequent items I and groups G , local transactions within a group G_T'' are generated.

Finally, all the local transactions G_T'' are delivered to the corresponding node reduce for summary and $G\_D'(G)$ is outputted.

The process of obtaining the transaction database corresponding to the group G is shown in Fig.6:



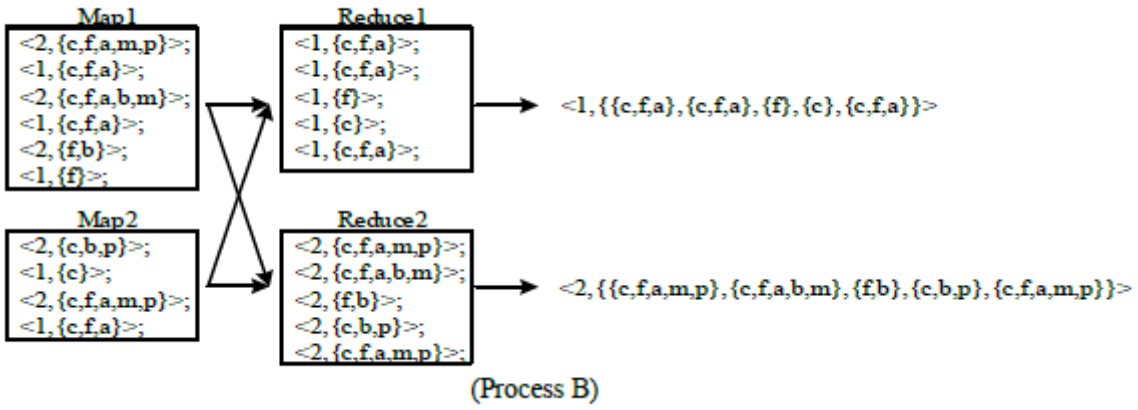
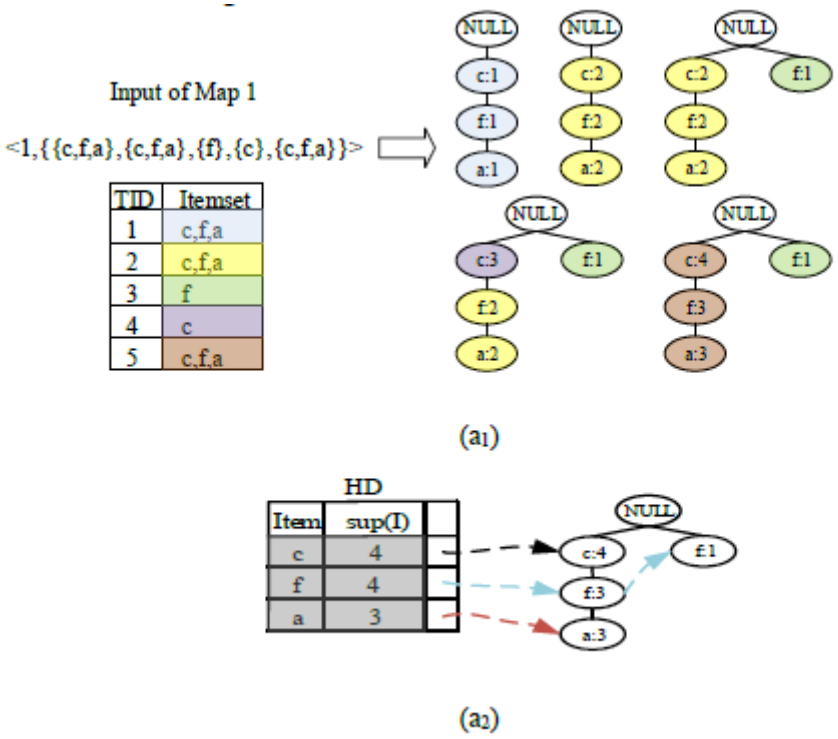


Fig. 6. Process of generating transaction databases $G\_D'(G)$ corresponding to group G

(4) Generate a collection of all frequent patterns MRFP_L

The key value pair $<G,D'(G)>$ of the key N is assigned to the $N-th$ Map node for frequent pattern mining. Each Map node only excavates the frequent patterns corresponding to the frequent items grouped to that Map node, and finally gives all the frequent patterns to the Reduce node for summary, outputting the collection MRFP_L of all the frequent patterns in the online transaction database. The partial process is shown in Fig.7:



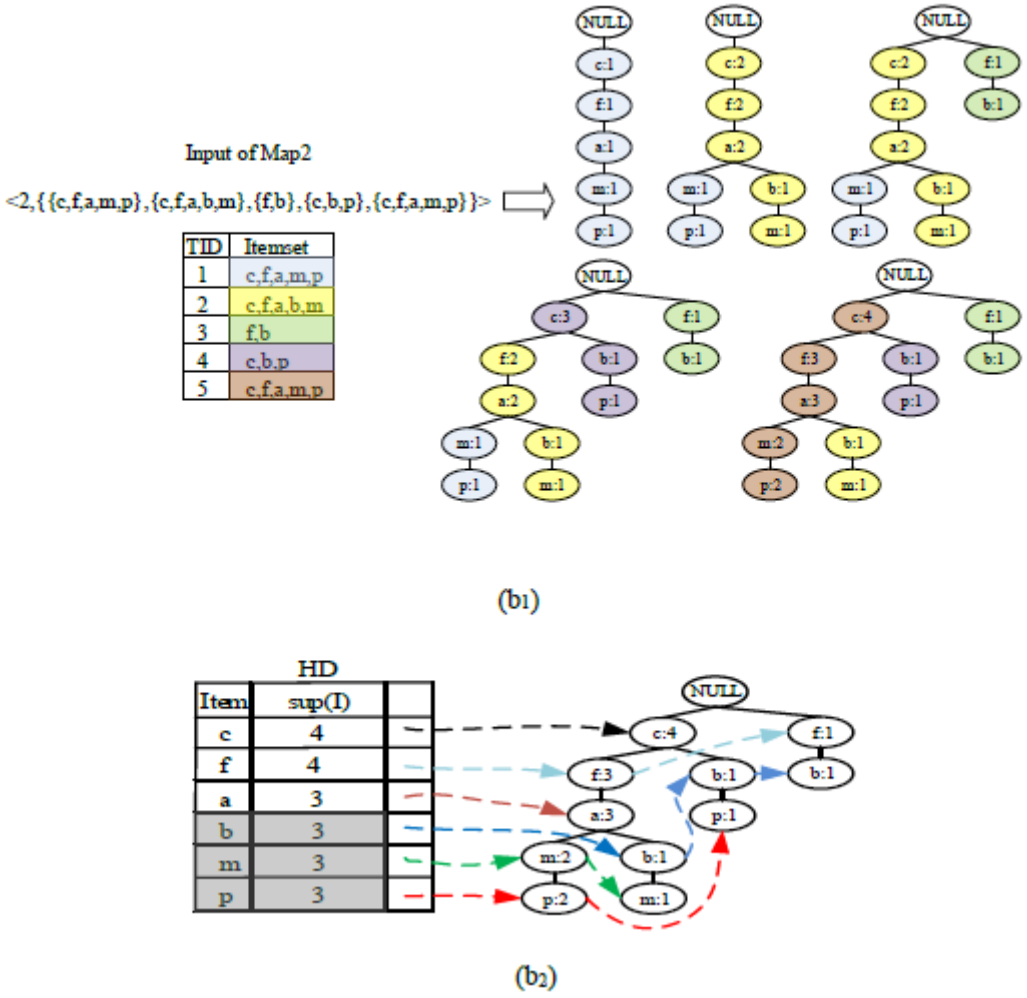


Fig. 7. The process of generating a collection of all frequent patterns

5 Frequent Co-occurrence Identities Consolidation

Looking at the results of frequent pattern mining, it can be found that multiple identities of the same user are scattered in different frequent item sets. To solve this problem, we introduce the concept of complete matching ratio based on the idea of Jaccard similarity calculation.

Complete Match Ratio: The complete match ratio is the ratio of the size of the intersection to the itemset X_1 and X_2 to the size of the itemset with fewer elements in it, recorded as

$$CM = \frac{|X_1 \cap X_2|}{\min\{|X_1|,|X_2|\}} \times 100\%$$

in which, $0 \le CM \le 100\%$, $|X_1 \cap X_2|$ is the number of elements at the intersection of X_1 and X_2; and $\min\{|X_1|,|X_2|\}$ is the number of elements in an item set with fewer elements. The minimum Complete Match Ratio can be used as a threshold to measure whether two item sets are combined.

The minimum complete matching ratio is min_CM . When the frequent itemsets X_i and X_j satisfy $CM(X_i, X_j) \ge$ min_CM , if the two itemsets are merged, the merged itemset can more fully express the identity set of a user.

6 Experiments

6.1 Data retrieval

Data Sources. The experimental data used in this paper is obtained through the following two methods:

(i) Campus network traffic log. User audit system records the traffic data of the main application flow as log files. The log files contain information of each user.

(ii) Data obtained from raw data flow analysis. We use CAIDA's open source

software Corsaro to identify and parse the underlying network traffic, sets up the network application identification features and user identification features through the secondary development of plug-ins, and develop the function to extract the user information.

The data obtained by the two methods are stored uniformly as shown in Fig. 8. The content includes four parts: user identities, hosting IP address, timestamp, and social media type coding. Some user names have been masked for privacy purposes.

user_name	local_ip	time	media_type
13[redacted]96	10.104.149.231	2015-03-01 00:00:00	1179652
90[redacted]70	10.101.154.94	2015-03-01 00:00:15	262145
19[redacted]279	10.104.147.210	2015-03-01 00:00:19	1179652
cntaobaotb[redacted]09_2012	10.104.174.25	2015-03-01 00:00:22	262150
13[redacted]96	10.104.149.231	2015-03-01 00:00:29	1179652
28[redacted]577	10.101.153.233	2015-03-01 00:00:30	1179652
69[redacted]27	10.104.177.198	2015-03-01 00:00:32	262145
cntaobao[redacted]g入云	10.104.173.216	2015-03-01 00:00:41	262150
10[redacted]042	10.101.152.39	2015-03-01 00:00:50	262145
57[redacted]86	10.104.122.206	2015-03-01 00:00:59	262145
21[redacted]321	10.101.152.37	2015-03-01 00:01:03	262145
28[redacted]16@qq.com	10.104.126.214	2015-03-01 00:01:05	262147

Fig. 8. Screenshots of Internet records

Experiment Dataset. Our experiment uses the campus network traffic log in March 2015 as the original data, which contains 581832 online records.

For the needs of campus network user audit, students use unified online authentication for network access, and their access numbers are used as the only user identification of students using the Internet. Even if students visit the Internet at different times, they are assigned different IP addresses, their unique user access number will not change, which can be used as the basis for this paper's experiment to verify the accuracy of user association. In the experiment, 3980 students is selected, involving 12952 IP addresses and 12289 accounts in total.

In the process of generating online transaction database, the traffic records are divided into time slices, and then duplicates are remove from data. After this step, only one online record with the same user ID and service type carried on the same network address in the same time slice is retained. Finally, the data are extracted to get the online transaction database, as shown in Fig 9. Some users' names have been masked for privacy purposes.

tid	items	count	local_ip	time
1	63[redacted]43--262145,cntaobao唯[redacted]猪--262150	2	10.100.21.232	99117
2	52[redacted]69--262145,o04IBALIDcRNx6[redacted]D3zgUd4Ss--1179652,zhu[redacted]ang49@163.com--65537	3	10.100.21.232	99119
3	13[redacted]917--262145,52[redacted]69--262145,5[redacted]69--1179652,o04IBALIDcRNx6U[redacted]3zgUd4Ss--1	4	10.100.21.232	99124
4	13[redacted]917--262145,52[redacted]69--262145,o04IBALIDcRNx6[redacted]D3zgUd4Ss--1179652	3	10.100.21.232	99125
5	31[redacted]97--1179652,cntaobaotb6[redacted]9--262150,o04IBAEIOQOwK[redacted]KZ9EuJzaxc--1179652,o04I	4	10.100.21.251	98979
6	cntaobaotb[redacted]9--262150,cntaobao[redacted]日安--262150	2	10.100.21.251	98980
7	9[redacted]8--262145,9[redacted]8--1179652,cntaobaotb6[redacted]9--262150	3	10.100.21.251	98981
8	9[redacted]8--262145,9[redacted]8@qq.com--65543,cntaobac[redacted]日安--262150	3	10.100.21.251	98984
9	31[redacted]97--1179652,cntaobao[redacted]kk--262150,cntaobaotb6[redacted]9--262150,o04IBAN381Dk[redacted]jKI	4	10.100.21.251	98985
10	41[redacted]4--1179652,cntaobaotb6[redacted]9--262150,o04IBAEIOQOwK[redacted]KZ9EuJzaxc--1179652	3	10.100.21.251	98986

Fig. 9. Screenshot of Internet transaction database

6.2 Algorithm Performance Assessment

To ensure fast and efficient mining of frequent co-occurrence user identities from massive data, the MRFP-Growth algorithm is developed with Java and is tested on the 64-core processor, 64G memory, 1T hard disk, Windows Server 2008 R2 operating system server. We evaluate its performance and compares it with the traditional FPGrowth algorithm. the algorithm execution time is shown in Fig. 10.

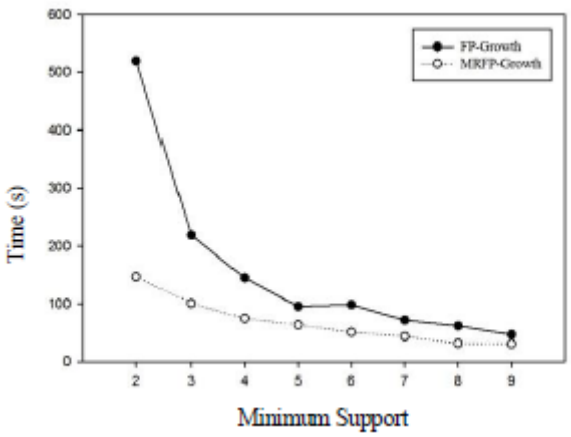


Fig. 10. Comparison of algorithm execution time

It can be seen from Fig.10 that the operation efficiency of MRFP-Growth algorithm is obviously higher than that of traditional FP-Growth algorithm on the online transaction database constructed by this experiment.

6.3 Parameter Setting

Time Slice Length Setting. The length of time slice has a great influence on the experimental results. Too long time slice will reduce the frequency of co-occurrence of user's identities, too short time slice will cut off the association between user's identities. By comparing the changes of the related content in the online transaction database and the related results under different time slice lengths, we finally select a time slice length suitable for the experimental data in this paper.

The time slice starts at 00:00:00 every day. The number of itemsets in the association results is changed under the minimum support of 3 to 4. As shown in Fig.11.

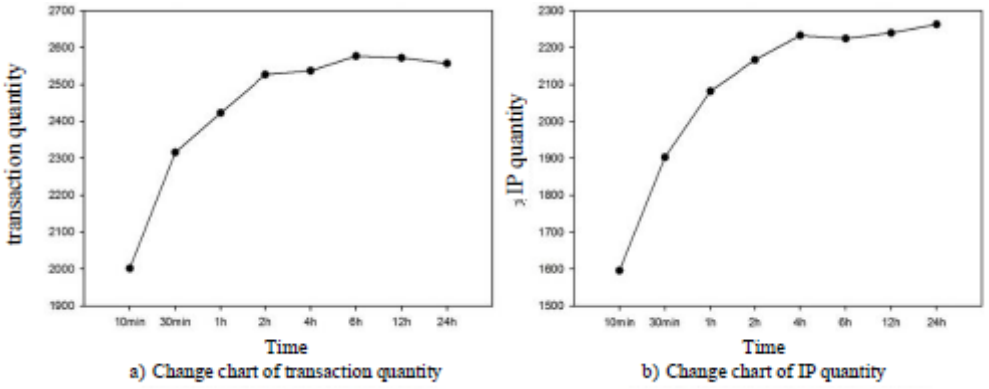


Fig. 11. Change chart of data related to online transaction database

It can be seen from Fig.11 that on the data set selected in this experiment, the time slice length is the most reasonable choice of 4 hours, which is also more consistent with the rest time rule of the users engaged in network activities in the campus network.

The Setting of Minimum Support. The support threshold has a great influence on the experimental results. Too large a threshold will lose a small correlation clue, and too small a threshold will aggravate the negative effect of the noise point on the associated results.

It can be seen from Fig.12 that the number of frequent itemsets and association results decreases with the increase of minimum support, and the number of elements of 2~5 in association results increases with the increase of minimum support. considering the above two factors synthetically, the experimental data is relatively reasonable when the minimum support is set to 4.

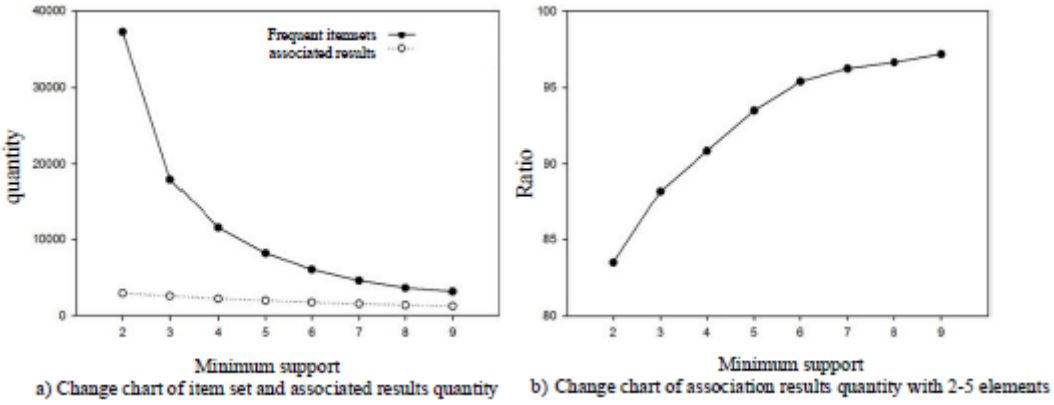


Fig. 12. Variation of association results under different minimum support

Setting of Minimum Complete Match Ratio. When the frequent itemsets are merged, the minimum complete match ratio will directly affect the scale of the later itemsets. Through the continuous analysis of the data, we set it to 50%, that is merge two itemsets when more than half of them are the same.

6.4 Analysis of Results

Analysis on the law of users' access to the Internet. We make statistics on the IP switching time interval of each user in the test data set, as shown in the table 4.

Table 4. Time interval statistics of user switching IP

Interval time(min)	no	(0,20]	(20,40]	(40,60]	(60,80]	(80,100]	(100,120]	(120,140]	(140,160]
Number of users	980	1885	1386	997	847	766	608	579	546

The results show that DHCP is widely used by campus network users to register online, and the average time interval between users to switch IP address is 53.9 minutes, which is related to users' living habits.

Within the statistical time range, we further make statistics on the time interval between individual users switching accounts on the same network address, and the results are as follows:

Table 5. Time interval statistics of single user switching account on the same network address

Interval time(min)	<5	5-10	11-20	21-30	31-40	41-50	51-60
Switching times	23712	4496	4683	2862	2024	1480	1178

The results show that the switching time of multiple accounts varies from less than 5 minutes to more than 60 minutes, with an average of 11.2 minutes, which is far less than the online period of a user using the network normally. The statistical results further verify the hypothesis that users cross use multiple network services in the Internet.

Analysis of association results. Based on the traffic data of the March 2015 log data in the campus network, the distribution of the number of each group of data elements in the association result is shown in Table 6 in two cases with minimum support of 3 and 4:

Table 6. Distribution of the number of elements per group of data in the association results

Number of elements in each data	2~5	6 to 10	More than 10
min_sup=3	88.2%	8.7%	3.1%
min_sup=4	90.9%	7.1%	2.0%

In the association results, if the number of elements in each group of data is less than 10, the grouping results are closer to the real situation, which are recorded as valid results, accounting for about 97% to 98% of the total association results. Extracting some of these association results, as shown in Fig.13 for further verification. It can be seen that group 566 is a group of QQ numbers and QQ mailboxes, which are correctly divided into one group; group 253 is the user identification using QQ account as another network service; group 435 is the two user identification using the same prefix into one group; and group 1452 is the association of the two user identifications that are not related to the appearance in one group (verified).

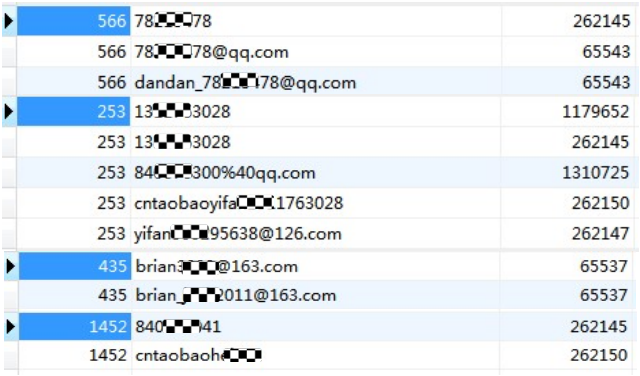


Fig. 13. Screenshot of some analysis results

For the 6~10 association results of the number of elements in the group, we think it is the data generated by sharing the Internet connection in a small area; for the association result of the number of elements in the group being more than 10, we think it may be the data generated by the collective use of the same wireless hot spot connection.

The evaluation results are shown in Table 7.

Table 7. Evaluation of the accuracy of association results

Data sets	Number of users (groups)
Number of groups to be verified	3980
Number of groups with correct associations	3394
Accuracy of association results	85.3%

7 Conclusion

In this paper, the association problem of user identities is transformed into the problem of frequent pattern mining by analyzing the user's online habit, and the research framework of user identity linkage method based on the user's online habit is given. The performance of the MRFP-Growth algorithm is analyzed in the experiment, and the effectiveness of the user identity linkage method proposed is verified. The experimental results show that the proposed method can quickly and efficiently extract frequent co-occurrence user identities from massive data, and the accuracy is relatively high.

The association analysis of account in the network traffic discussed in this method is carried out under the assumption that the account is plaintext visible. This kind of traffic is still common in some social applications at present, but considering the demand of user traffic privacy protection, some network application traffic is encrypted, in this case, this method is not applicable.

In the future research, we will further study how to carry out user identity linkage in network traffic data in special environments such as WiFi sharing, NAT and so on.

References

[1] Zhang, J., Yu, P. S., & Zhou, Z. H.: Meta-path based multi-network collective link prediction. In Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1286-1295 (2014).

[2] Wang, C., Zhao, Z., Wang, Y., Qin, D., Luo, X., & Qin, T.: Deepmatching: A structural seed identification framework for social network alignment. In 2018 IEEE 38th International Conference on Distributed Computing Systems (ICDCS), pp. 600-610. IEEE. (2018).

[3] Zhang, J., & Philip, S. Y.: Integrated anchor and social link predictions across social networks. In Twenty-Fourth International Joint Conference on Artificial Intelligence. (2015).

[4] Shu, K., Wang, S., Tang, J., Zafarani, R., & Liu, H.: User identity linkage across online social networks: A review. Acm Sigkdd Explorations Newsletter, 18(2), 5-17. (2017).

[5] Huang, S., Zhang, J., Schonfeld, D., Wang, L., & Hua, X. S.: Two-stage friend recommendation based on network alignment and series expansion of probabilistic topic model. IEEE Transactions on Multimedia, 19(6), 1314-1326. (2017).

[6] Huang, S., Zhang, J., Wang, L., & Hua, X. S.: Social friend recommendation based on multiple network correlation. IEEE transactions on multimedia, 18(2), 287-299. (2015).

[7] Jiang, M., Cui, P., Yuan, N. J., Xie, X., & Yang, S.: Little is much: Bridging crossplatform behaviors through overlapped crowds. In Thirtieth AAAI Conference on Artificial Intelligence. (2016).

[8] Zhan, Q., Zhang, J., Wang, S., Philip, S. Y., & Xie, J.: Influence maximization across partially aligned heterogenous social networks. In Pacific-Asia Conference on Knowledge Discovery and Data Mining, pp. 58-69. Springer, Cham. (2015).

[9] Zafarani, R., & Liu, H.: Connecting corresponding identities across communities. In Third International AAAI Conference on Weblogs and Social Media. (2009).

[10] Vosecky, J., Hong, D., & Shen, V. Y.: User identification across multiple social networks. In 2009 first international conference on networked digital technologies, pp. 360-365. IEEE. (2009).

[11] Iqbal, F., Binsalleeh, H., Fung, B. C., & Debbabi, M.: Mining writeprints from anonymous e-mails for forensic investigation. digital investigation, 7(1-2), 56-64. (2010).

[12] Yuan, N. J., Zhang, F., Lian, D., Zheng, K., Yu, S., & Xie, X.: We know how you live: exploring the spectrum of urban lifestyles. In Proceedings of the first ACM conference on Online social networks, pp. 3-14. (2013).

[13] Soltani, A., Canty, S., Mayo, Q., Thomas, L., & Hoofnagle, C. J.: Flash cookies and privacy. In 2010 AAAI Spring Symposium Series. (2010).

[14] Acar, G., Juarez, M., Nikiforakis, N., Diaz, C., Gürses, S., Piessens, F., & Preneel, B.: FPDetective: dusting the web for fingerprinters. In Proceedings of the 2013 ACM SIGSAC conference on Computer & communications security, pp. 1129-1140. (2013).

[15] Mayer, J. R. Any person... a pamphleteer": Internet Anonymity in the Age of Web 2.0. Undergraduate Senior Thesis, Princeton University, 85. (2009).

[16] Eckersley, P. How unique is your web browser?. In International Symposium on Privacy Enhancing Technologies Symposium, pp. 1-18. Springer, Berlin, Heidelberg. (2010).

[17] Huang, Y., Zhang, L., & Zhang, P.: A framework for mining sequential patterns from spatio-temporal event data sets. IEEE Transactions on Knowledge and data engineering, 20(4), 433-448. (2008).

[18] Karli, S., & Saygin, Y.: Mining periodic patterns in spatio-temporal sequences at different time granularities. Intelligent Data Analysis, 13(2), 301-335. (2009).

[19] Li, Z., Ding, B., Han, J., Kays, R., & Nye, P. Mining periodic behaviors for moving objects. In Proceedings of the 16th ACM SIGKDD international conference on

Knowledge discovery and data mining, pp. 1099-1108. (2010).
[20] Garg, R., & Gulia, P.: Comparative Study of Frequent Itemset Mining Algorithms Apriori and FP Growth. International Journal of Computer Applications, 126(4). (2015).