Mining And Application of User Behavior Pattern Based on Operation And Maintenance Data

Wenji Zhang
Beijing University of Post and Telecom, Beijing, P.R. China
1351540186@qq.com

Wenan Zhou
Beijing University of Post and Telecom, Beijing, P.R. China
zhouwa@bupt.edu.cn

Jun Luo
Beijing University of Post and Telecom, Beijing, P.R. China
jennyluo@bupt.edu.cn

***Abstract*—In order to provide users with personalized services, or to implement user-centric management, operators usually need to collect various information of users. However, in an actual network scenario, such information is often difficult to obtain accurately due to reasons such as protecting user privacy and avoiding interference with the user experience process. Common methods such as setting up a laboratory network environment, implementing user research, etc., are difficult to implement in the live network. Through the service log of the network device interface, this paper extracts a variety of data related to the user's network behavior, and proposes a comprehensive multi-dimensional data user behavior expression method, which is transformed into the expression of user behavior, including time, space and behavior semantics. For the user's daily network behavior, we propose a user behavior symbolization method for different application scenarios, and propose a user behavior pattern mining method based on PrefixSpan to mine the user behavior sequence pattern after symbolization. This method can mine the patterns of user behavior and behavior from the data, and provide the basis for personalized service and management. We used the http connection log of all mobile users of a carrier in a city for one day to conduct our experiments.**

Keywords—User Behavior, Pattern Mining, Symbolic

I. INTRODUCTION

With the development of communication technology and market, the competitive pressure between operators is increasing. How to retain and attract customers is an issue that they must consider. Traditionally, operators expect to achieve this by improving service quality. However, simply facing the technical level will inevitably ignore the subjective influence factors of users. Therefore, finding and solving problems from users is the current mainstream solution. Many studies on QoE or CEM are based on this. Since user is a subjective participating entity with unique usage habits, business preferences, behavioral characteristics, etc., it's difficult to obtain the information accurately. In related researches, building experimental networks and implementing user research are widely used to solve the problem. However, they usually face many problems, such as being far from the actual network environment, affecting or even interfering with user business processes, and being difficult to apply widely.

User behavior is directly generated by the interaction process between the user and the business, not only related to the objective business service process, but also the direct performance of the user's subjective participation. Therefore, many researches for users start with user behavior. One of the most common research contents is the user's behavior pattern. In summary, current researches on user behavior patterns have many aspects, which can be roughly divided into: the data flow pattern discovery of user groups, the content-oriented pattern discovery, and pattern discovery for individual users.

The data flow pattern discovery ignores the user individual and focuses on the data flow in the network. The main purpose is to monitor, manage and optimize the network. In prior work[1], automated network management is supported by discovering interesting patterns from telecommunication network monitoring data, to accommodate the need for automated monitoring and management of heterogeneous wireless communication networks. In [2][3], methods for finding interesting patterns from the database were introduced.

The main research content of the content-oriented pattern discovery is to discover the combination of content frequently accessed by the user, the navigation sequence, and the like. It is mostly used in recommendation system, website design and management, etc. Works[4][5][6] analyzed users' web browsing mode through web logs.

The user-oriented pattern discovery is to describe the user's personal preferences, habits, context perception, etc., and to establish user portraits, which are mainly used for personalized management of users. Work [7] collects rich context and mobile device usage records through device logs to mine mobile users' personal preferences. Work [8] used Naive Bayes and logistic regression methods to predict the user's traffic patterns.

Since the pattern mining method is for a single variable problem, most of the related research focuses on only one aspect of user behavior, while ignoring most of the semantic content. In [9][10][11], pattern mining is carried out based on spatio-temporal data with semantics. Work [9] adopts the method of user moving hot regions and semantic tags to realize the prediction of user's location, time and semantic. For the multi-semantic content of spatio-temporal data, [9] proposes a Sequence Symbolization (SS) and Advanced Sequence Symbolization (ASS) method, symbolizing the combination of spatio-temporal and semantic information to apply the pattern mining method. However, its symbolization algorithm is only applicable to the scenarios with single user, small traffic, and few business types, so it cannot be directly generalized to a large and complex scene such as a communication network.

In order to study user behavior in communication networks, we propose a user behavior semantic expression for data services that based on http. The sequence symbolization algorithm is designed for single user scene and multi-user scene, and the user behavior pattern mining method based on PrefixSpan[12] is implemented to explore the user's behavior pattern. During the research process, the two main challenges we faced are: how to define and express a user's behavior, and how to symbolize multiple semantic content in single-user and multi-user scenarios. For the first

978-3-903176-15-7

problem, we need to divide the boundaries of the user's subjective perception of the business process, while retain as much user behavior-related data as possible. For the second question, we need to consider the accuracy and usability of spatio-temporal data, as well as the problem of the excessive symbol sample space.

The following contents of this paper are as follows. The second section describes the definition of user behavior and the extraction of user behavior semantic. The third section carries out the symbolization of multi-semantic user behavior in different scenarios, and sequence processing. Section IV will illustrate the application scenarios that this study applies to. Section V summarizes this article and looks forward to the future work.

II. USER BEHAVIOR EXTRACTION

A. Definition of User Behavior

The data used in this study comes from the http connection logs of all mobile users in a city within one day. The main equipment is the mobile smart device, and the connections come from various Apps. Since each network connection is not equal to the user's actual business process, we firstly need to aggregate these connections to the same level that the user perceives once. Because smart devices have a wide range of background applications and one application may call other ones, the application type cannot be directly used as the basis for dividing user behavior. Considering that a user's perception must be accompanied by a period of time, we mainly define the user behavior based on time, supplemented by the switching of business types. The algorithm for combining network connections to user behaviors is as follows.

Algorithm 1: Combine To User Behavior

Input: *L: links of one user that sorted by time;*
t_l: lower boundary of time interval
t_u: upper boundary of time interval
m: magnification of time interval increasing

Output: *S: a user behavior sequence*

```
1   timeInterval = t_l;
2   startPage = 0;
3   behavior = {};
4   for i=1 to |L|-1  do
5      curInterval = L[i].time-L[i-1].time;
6      if curInterval > timeInterval then
7         S.push_back(behavior);
8         behavior = {};
9         timeInterval = t_l;
10     else if curInterval * m > t_u then
11        timeInterval = t_u;
12     else if curInterval * m >  timeInterval then
13        timeInterval = curInterval * m
14     end
15     behavior. push_back(L[i]);
16  end
17  S.push_back(behavior)
18  return S;
```

B. Extraction of User Behavior Semantic

Since the business process comes from the user's subjective use process, we express the semantics of the user's behavior as the detailed information of the business process. The information contained in each user behavior includes spatial information, time information, and semantic information.

In our case, the spatial information is the cell's ID. The time information is the access start time of the business. Semantic information contains numerical semantics and enumeration semantics. Numerical semantics contain upstream traffic, downstream traffic, and duration. Enumeration semantics contain App type, App subtype, access host, client, and request content type. In particular, the content of an enumerated semantics may be a collection of multiple values.

III. MULTI-SEMANTIC USER BEHAVIOR SYMBOLIZATION

User behavior contains a variety of information, as described earlier, this study uses a variety of data to describe the semantic information of user behavior. In order to mine the user's behavior patterns, we used a sequential pattern mining algorithm. The pattern mining algorithm is basically oriented to single semantic content. Therefore, before the sequential pattern mining, we need to symbolize the user behavior with multiple semantic information. That is, multi-semantic information is converted into single semantic information while retaining the semantics of key user behavior information.

In [9], considering the time, space and semantics, the sequential symbolization (SS) method and the advanced sequence symbolization (ASS) method are proposed. The Jaccard distance is used to calculate the spatial semantic distance of the user, and the cosine similarity is used to calculate the temporal semantics and behavioral semantic distances. Then make whether the three distances be less than a certain threshold as the standard of symbolization. In this paper, since the time granularity of user behavior is too different from that of time and location, the symbolization is only oriented to user behavior semantics, while the spatial and temporal features can be analyzed when applied. For different application scenarios, this paper proposes different symbolization methods.

A. Symbolization in single-user scenarios

For a single-user scenario, since the value space of each user's behavioral semantics is limited, a symbolization method based on distance comparison can be adopted. Since the expression of user behavior includes both numeric types (duration, upstream traffic, downstream traffic) and enumeration types (App type, App subtype, access host, client, and request content type), we calculate the semantic distance separately for different types of semantic content. Then we use the sum of their distances to represent the distance of the user's behaviors, and the distance within the threshold is assigned the same symbol. Specifically, we use the cumulative-probability-difference to represent the distance of the numeric type semantics, and use the Jaccard-distance to calculate the distance of the enumerated type semantics. The distance between symbols is the weighted sum of the semantic distances. The specific algorithm process is as follows.

Algorithm 2: Symbolization Based on Distance Comparison

Input: *B: user behavior sequence*
t_x: threshold of each semantic item

Output: *S: symbol sequence*

```
SYM = 1
S.push_back(SYM);
for i=1 to |B|-1 do
    minDis = t_x;
    minIndex = i;
    found = false;
    for j=0 to i-1 do
        CALCULATE distance[i][j];
        if distance[i][j] < minDis then
            minDis = distance[i][j];
            minIndex = j;
            found = true;
        end
    end
    if found then
        S.push_back(S[minIndex]);
    else
        SYM ++;
        S.push_back(SYM);
    end
end
return S;
```

B. Symbolization in multi-user scenarios

For multi-user scenarios, especially for scenes with a large number of users, the symbolization principle of user behavior needs to be uniform. However, the value space of each semantic content will become huge at this time, and the algorithm based on distance comparison requires at least $O(n^2)$ time complexity, which is often unacceptable. To solve this problem, we have designed an algorithm for O(n) complexity. That is to say, the value space of each semantic content is firstly counted, and the numerical types are divided into equal intervals according to the cumulative probability distribution, and different symbols are used between different intervals. For enumerated types, the collection is first replaced with the most frequent elements, and then the parts with a ratio less than 0.1% are uniformly labeled. User behaviors of different enumeration values use different symbols. The specific algorithm process is as follows.

Algorithm 3: Symbolization Based on Value Space Statistics

Input: *B: user behavior sequence*
B_x: attribute of user behavior semantics

Output: *S: symbol sequence*

```
STATISTICS to space_x{} for each attribute;
for i=0 to |B|-1 do
    foreach B_x in B do
        if B_x[i].is_numerical() then
            sym_num = space_x(B_x[i]);
        else
            sym_enum = space_x(B_x[i]);
        end
    end
    sym = space(sym_num1, …, sym_enum1, …);
    S.push_back(sym);
end
return S;
```

When performing the STATISTICS, the semantics that rarely occur can be ignored. Because most of the semantic value distribution is concentrated, the error caused by sampling is also small and completely acceptable. We sampled 10 groups of users, each group of 2000 people, and the semantic unions in each group were then intersected as the available value space, which could cover 99.2% to 100% of any semantics.

C. Result of Symbolization

Comparing the two methods, due to the data filtering and statistical processing of the multi-user scene, the details of the data will inevitably be lost. However, the time complexity of the algorithm is greatly reduced, and this trade-off is still very valuable in a scene with a large number of users. For some special application scenarios, such as the "VIP User Group" scenario, although it is a multi-user, since the number of users is not too large, the symbolization algorithm based on distance comparison is still acceptable.

For the result of symbolization, the meaning of each symbol represents the overall description of the semantics of all user behaviors to which the symbol is assigned. For numerical items of semantic terms, they are described by means of an average. For semantic items of the enumerated type, the enumerated value or the most frequent of the multiple enumerated values is selected.

D. Sequence Processing

After symbolizing user behavior, we can get a long sequence of user behaviors. For the data in our study, it is a sequence of users' online behavior within one day. To mine the user's behavioral sequence pattern, we need to divide this long sequence into multiple subsequences first. Due to the orderliness and mutual exclusion of user behavior on time, we use time as the basis for subsequences. For each user's behaviors, the duration is firstly counted and sorted. According to the distribution characteristics of the duration, a scale value is selected and the duration at the scale value is token as the threshold. If the threshold is exceeded, a new subsequence is divided.

Now we get multiple behavior sequences for each user, and we can use the PrefixSpan method to perform pattern mining. The result of the mining is just the user's behavior pattern.

IV. APPLICATION CASES

The problem solved in this paper is the mining of user behavior patterns that with multiple semantic content under mobile networks. It can be widely used in a variety of areas related to user behavior. This section will give some examples in common scenarios.

A. User Dehavior Prediction

User behavior prediction is a single-user scenario and is one of the most common application scenarios of user behavior patterns. In the past research, usually only one of the user's semantic behavior characteristics can be considered., such as the websites of web browsing, the check-in behaviors of social Apps, etc. The multi-semantic

symbolization method proposed in this paper can support multiple semantics that comprehensively consider user behavior. For example, for the data used in this paper, the predicted results will include the user's the most likely location, the App used, the site visited, the type of resource, duration, and so on. These can provide better help for users' personalized services.

We implemented the work of user behavior prediction and used its results to verify the efficiency and accuracy of the symbolization algorithm. The data set we used was from an operator's http connection logs in a city for one day. 1000 users with enough business volumes are selected randomly. 80% of their behavioral symbol sequences were used for pattern mining and 20% for predicting. When the user behavior sequence matches the first N-1 of the pattern of length N, predict that the Nth behavior will also appear.

We analyzed the possible effects of different time interval when user behaviors are fused, and the different weights when symbolizing. The time intervals and weights are shown in Table I and II.

TABLE I. Time Interval When Combine User Behavior

Label	lower boundary	upper boundary	increase
S1	2	2	——
S2	2	10	2
S3	2	10	4

TABLE II. Weight of Semantics

Label	Numeric Semantics	Enumeration Semantics
Q1	0.7	0.3
Q2	0.5	0.5
Q3	0.3	0.7

For the predicted result, if the distance between the predicted symbol and the actual symbol is less than 90% of other symbols, it is considered correct. The accuracy under different support levels is shown in Figure 1.

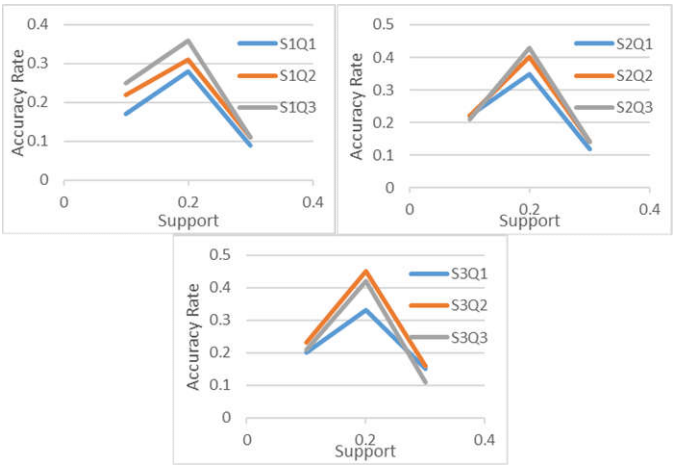

Figure 1. Accuracy Rate of Prediction Results

Based on the results, we can see that the fusion method of dynamically changing time thresholds is more efficient. The weight of the enumerated type semantics should be slightly higher than the numeric type semantics. When the degree of support is large, the number of models mined will be drastically reduced, resulting in a decrease in accuracy.

B. *Similar User Discovery*

In a multi-user scenario, many studies are dedicated to discovering similar users. Through the methods in this paper, after we get the result of symbolization, we can make an overall average description of the symbols, so that each symbol has its own semantics. Next we can calculate the semantic distance between symbols, and the cosine similarity equidistance calculation formula can be used. Since the value space of the symbol is controllable and difficult to change, the cost of the calculation is completely acceptable. After calculating the distance between the symbols, it is easy to evaluate the similarity between the sequences, which can be calculated based on the example between the symbols contained in the sequence, the length of the sequence, the position of the symbol, and the like. Based on this, we can evaluate the similarity of behavior patterns among multiple users. Through a series of distance and similarity measures, users with similar behavior patterns can be easily found, and business recommendations and user portraits can be improved based on similar users.

C. *QoE And CEM*

For many service providers, especially network operators, QoE (Quality of Experience) and CEM (Customer Experience Management) are important research contents to evaluate service levels and reflect user experience. But in the researches of QoE and CEM, the description of user characteristics has always been a recognized problem. Through the method of this paper, the characteristics of users can be described from the perspective of user behavior patterns, and then the user's usage habits and hobbies can be explored to improve the user domain content of QoE and CEM researches. It has a certain guiding role for the operator's network optimization, user management, resource allocation, and energy management.

V. Conclusion

The goal of this paper is to extract and express user behavior in a mobile network environment and to mine user behavior patterns. In order to achieve this goal, the problems we must solve include: (1) extracting and expressing the semantics of user behavior, and (2) mining frequent patterns of user behavior sequences. In order to solve the problem 1, we first divide the connections that are close in time into the same group according to the change of the duration of the connection, which belong to the same user behavior. Then, from the specific content of these connections, the business process information that can express the semantics of the user behavior is calculated and used to express the user behavior semantics. For problem 2, since sequential pattern mining requires single-semantic data, we propose a symbolic method for single-user scenarios and a symbolic method multi-user scenarios to apply the PrefixSpan method. The purpose of symbolization is that user behaviors with similar semantics will be assigned the same symbol. In addition, each symbol has an overall averaging semantic description, and can calculate the semantic distance between symbols, which is convenient for user personalized service and user group management in the actual application scenario. We also introduced some available application cases and verified the efficiency of the method by predicting cases of user behavior in a single-user scenario. In future work, we expect to be able to obtain sufficient user subjective feedback data to optimize the methodology of this study.

REFERENCES

[1] Qu Z, Keeney J, Robitzsch S, et al. Multilevel pattern mining architecture for automatic network monitoring in heterogeneous wireless communication networks[J]. China Communications, 2016, 13(7): 108-116.

[2] Fournier-Viger P, Lin J C W, Kiran R U, et al. A survey of sequential pattern mining[J]. Data Science and Pattern Recognition, 2017, 1(1): 54-77.

[3] Thakkar J, Parikh M. A Survey on Efficient Frequent Pattern Mining Techniques[J]. 2018.

[4] Dharmarajan K, Dorairangaswamy M A. Discovering User Pattern Analysis from Web Log Data using Weblog Expert[J]. Indian Journal of Science and Technology, 2016, 9(42).

[5] Rao R S, Arora J. A Survey on Methods used in Web Usage Mining[J]. 2017.

[6] Prakash P O, Jaya A. Analyzing and predicting user behavior pattern from weblogs[J]. International Journal of Applied Engineering Research, 2016, 11(9): 6278-6283.

[7] Zhu H, Chen E, Xiong H, et al. Mining mobile user preferences for personalized context-aware recommendation[J]. ACM Transactions on Intelligent Systems and Technology (TIST), 2015, 5(4): 58.

[8] Singh R, Srinivasan M, Murthy C S R. A learning based mobile user traffic characterization for efficient resource management in cellular networks[C]//Consumer Communications and Networking Conference (CCNC), 2015 12th Annual IEEE. IEEE, 2015: 304-309.

[9] Chen C C, Kuo C H, Peng W C. Mining spatial-temporal semantic trajectory patterns from raw trajectories[C]//Data Mining Workshop (ICDMW), 2015 IEEE International Conference on. IEEE, 2015: 1019-1024.

[10] Zhang C, Han J, Shou L, et al. Splitter: Mining fine-grained sequential patterns in semantic trajectories[J]. Proceedings of the VLDB Endowment, 2014, 7(9): 769-780.

[11] Su H, Zheng K, Zeng K, et al. STMaker: a system to make sense of trajectory data[J]. Proceedings of the VLDB Endowment, 2014, 7(13): 1701-1704.

[12] Pei J, Han J, Mortazavi-Asl B, et al. Prefixspan: Mining sequential patterns efficiently by prefix-projected pattern growth[C]//icccn. IEEE, 2001: 0215.