Identifying Users Based on Time-Frequency Characteristics

Yaqiong Gao, Wanghu Chen, Bo Yang, Yuxiang Mu and Jing Li

Institute of Computer Science and Engineering, Northwest Normal University
Lanzhou, 730070, China.
Email: 1144652220@qq.com

Abstract. With the rapid development of information technology, whether it is with regard to the personalized recommendation of the business or network warning by the public security organ, the identification of online users is a very important research topic. The user's web browsing behavior data contains potential behavioral characteristics, in this paper, we utilize the time-frequency analysis method, which extracts the behavior characteristics when users accessing the network, so as to obtain the user behavior pattern and calculate the similarity between the behavior patterns for achieving the purpose of user identification. This method is verified by the weblog, and the experiment results show that the method extracts user network behavior features based on time-frequency analysis to ensure the accuracy of user identification.

1. Introduction
With the rapid development of computer technology and network technology, the Internet has played an increasingly important role in people's daily lives. Faced with the proliferation of Internet users and the explosive growth of information, how to better use the Internet to analyze the customer's demands is a valuable research issue [1]. At present, the majority of the user requirements analysis is mainly based on their specific behaviors, in the case of the definite identity of the user, such as the analysis of the user's shopping behaviour [2], in order to produce a recent demand list that makes a recommendation for the users. Accessing users' social networks is widely used in information dissemination [3], advertising [4], etc. Personalized recommendation of the user behavior is applied in news promotion [5], online education [6] and other fields. However, users may change their account ID or anonymous access when they use the network service in real life. Therefore, it is greatly significant to identify anonymous online users by mining their online behavior data to obtain periodic behavior characteristics, which is implied in their online behavior.

The user's network behavior is dynamic change over time. There are differences in life, work, and personality among users, resulting in different life rules or habits. The types and scales of internet services are becoming more and more diversified, and the network information is changing dynamically so that there are few rules to follow [7]. The above reasons have created many challenges for online users identification. If we can grasp the individual behavior characteristics of users, and further discover the periodic behavior patterns for achieving the user's accurate identification, which will have practical value in the personalized recommendation of business and the network warning of public security organs. Consequently, studying the regularity of users’ behavior is great theoretical and practical significance in time and space. In this paper, how to extract the characteristics of online behavior data for user identification, the discrete Fourier transform is applied to obtain the features, which is related to the user's browsing behavior in the time domain and frequency domain for achieving the purpose of user identification. Currently, there are few studies on the processing of network behavior characteristics by Fourier transform technology. As a consequence, our contribution
is adopting the time-frequency analysis method to extract the user's network behavior characteristics, so as to achieve the purpose of identifying users.

The rest of this paper is organized as follows: Section 2 describes the current related work. Section 3 describes the user behavior data. Section 4 extracts the user's browsing behavior features by the DFT. Section 5 describes the process of user identification and gets relevant results. The experiment results show that time-frequency analysis ensures the high accuracy of the user’s recognition according to the user's browsing behavior characteristics. This paper concludes with the research conclusion and the prospects for the next step of section 6.

2. Related Work

Nowadays, with the vigorous development of the information network, the normal life, working and study increasingly rely on the support of the Internet. As an important resource, the user's accessing network data can understand the users' requirements on the one hand, and integrate business models on the other hand. Taking the website as an example, after identifying each user, we can not only know more clearly how many users have visited the website and who they are (user ID, mailbox, gender, age, etc.), but also track these users better, find out their behavior characteristics, hobbies and personalized settings, so as to better grasp users requirements and enhance the users experience [8].

Li Y et al.[9] Formulated the user identification model based on tweet, in order to build better user's profiles and benefit many applications. To take advantage of the full range of services that online social networks (OSNs) offer, people commonly open several accounts on diverse OSNs where they leave lots of different types of profile information. The integration of these pieces of information from various sources can be achieved by identifying individuals across social networks [10]. To solve these problems, Feng S et al.[11] Proposed an iterative two-stage algorithm (GAUI) using Global view features with user Attribute features to solve User Identification. Literature [12] and [13] were to identify the same user by using different social media, the former user identification algorithm was based on friendship, while the latter was based on user-specific behavior patterns to find personal identity mapping, all of which in order to promote the creation of new online services across websites.

The related research works on user identification mainly focus on social networks, and the user's accessing network behavior, which is actually the browsing behavior along the time axis. We mainly focus on the temporal distribution of users' browsing behavior. To this end, the time-frequency analysis method is used to extract the characteristics of the temporal distribution of users' browsing behavior. At present, there are relatively few studies in this area. However, another kind of research uses the time-frequency analysis method to extract frequency-domain features from data in some temporal-spatial domains.

Wenxin LI et al. proposed a feature extraction method by converting a palmprint image from a spatial domain to a frequency domain using Fourier Transform. The features extracted in the frequency domain are used as indexes of the palmprint templates in the database to achieve the best match [14]. Liu Xiang et al. utilized clustering algorithm to discretize the space trajectory data, and designed an equally spaced sampling algorithm to regulate time series data, thus translated spatiotemporal trajectory data in continuous domain into discrete binary sequences, and found periodic behavior patterns with the discrete Fourier transform and circular autocorrelation methods [15]. Literature [16] collected digital records of Internet behavior from the Web server of the LAN gateway, and used clustering and time-frequency analysis (such as DFT) to process collected data, and obtained features in the time and frequency domain, which were closely related web users' mental status, and built up classification models to recognize depression.

In general, the current researches on identifying users and extracting features using time-frequency analysis provide a good basis for the work done in this paper. However, these studies were limited to data on one single website (such as Facebook and Twitter) or a single type of website (such as the social network), which only represented a small part of individual Internet behavior [16]. Nowadays, the Internet is a multi-site network environment, so the above methods have some limitations on the accurate identification of users.

In summary, due to the user's access to the Internet, which includes individual browsing histories on different web sites, these records are signals that change over time [16]. In this paper, we
will abstract the discrete digital signals according to the user's web browsing behavior, and pay attention to the temporal distribution of the browsing behavior of users. Time-frequency analysis is one of the important tools for non-stationary signal analysis, it can help people understand the complex structure of signals and reveal their inherent laws [17]. In this paper, we utilize the time-frequency analysis method to extract the characteristics of the user's network browsing behavior and get the user's behavior pattern, in order to identify the user.

3. User Network Behavior Data

The user network behavior data is the basis of the experiment in this paper. The user's online browsing behavior records need to be preprocessed to get the attribute fields required by the experiment. Usually, each user accesses the LAN gateway firstly when surfing the Internet, and users browsing histories and Internet behaviors are all recorded by the gateway server. From 2018-5-1 to 2018-10-31, we had collected the data recorded in the Web server of a laboratory gateway and preprocessed the recorded data to obtain information such as user id, request access time, access resources(URL), site type and so on. That is mainly from the data set according to the "URL" to obtain the corresponding network site classification (see Figure 1, the classification criteria are not unique). In order to describe the subsequent statistical process more intuitively, the variable names are used to indicate the main types of sites that users visit in their daily life or learning (see Table 1).

![Preprocessed data](image)

**Table 1. Definition of Web Site Type Format.**

| Site name             | Variable name | Site name          | Variable name |
|-----------------------|---------------|--------------------|---------------|
| News portal           | S1            | Academic Forum     | S8            |
| Social networking sites| S2            | Movie website      | S9            |
| Search Engines        | S3            | Music website      | S10           |
| Enterprise site       | S4            | Online game        | S11           |
| Government website    | S5            | Online shopping    | S12           |
| Academic research     | S6            | Other sites        | S13           |
| Online learning       | S7            | Null               | 0             |

The content of this section is about the types of network sites where user behavior data is distributed over time. To this end, the time from 7:00 to 22:00 is divided into 90 time periods (every 10 minutes is a time period), each time period will have a larger weight of access sites, that is, the longest stay time. Therefore, through the preprocessed data, sorting out the distribution of the types of network sites on each time period every day, statistics for 90 days. Data with ID 192.168.1.132 is selected as shown in Table 2.
Table 2. Distribution of Web Site Types.

| User_id     | Time period   | Day 07:00-07:10 | 07:10-07:20 | 07:20-07:30 | ... | 21:40-21:50 | 21:50-22:00 |
|-------------|---------------|-----------------|-------------|-------------|-----|-------------|-------------|
| 192.168.1.132 | 1             | S1              | 0           | S1          |     | S2          | S13         |
| 2           | S3            | S2              | 0           | S10         | S1  |
| 3           | S1            | 0               | S3          | S9          | S3  |
| 4           | S2            | 0               | S6          | ...         | S3  | S2          |
| ...         |               |                 |             |             |     |             |             |
| 89          | 0             | S1              | S3          | S11         | S11 |
| 90          | S2            | S3              | S1          |             |     |             |             |

4. Feature Extraction

User feature extraction is the basis of user identification. According to the network site type distribution table of the user about the time period, two operations are performed: (1) Counting the probability of visiting each site every day for 90 days; (2) Counting the probability of visiting each site at each time period in 90 days. After statistics of user data, we use the time-frequency analysis method to select and extract important characteristics that are closely related to the user’s browsing behavior to accurately identify the user.

Discrete Fourier transform (DFT) is adopted in this paper, which is a time-frequency analysis method for feature extraction proposed in reference [16]. Therefore, the joint study of the probability of visiting each site every day and the probability of accessing each site at each time period in the 90 days, we can better explain the behavior record of the user on a certain day and at a certain time period.

Fourier transform is one of the most useful methods for mathematical physics and engineering. Discrete Fourier Transform (DFT) is the discrete form of the Fourier transform that transforms the signal sampling of the time domain into the sampling of the frequency domain.

In the discrete Fourier transform, \( x(n) \) is a finite sequence of length \( n \), the actual cosine and sine are just expressions of coefficients, as follows:

\[
X(k) = \sum_{n=0}^{N-1} x(n) \left[ \cos \frac{2\pi kn}{N} - j \sin \frac{2\pi kn}{N} \right], \quad k = 0, 1, \ldots, N - 1
\]

Where \( X(k) \) is the \( k \)th DFT coefficient, which is the amplitude of the \( k \)th component.

The DFT of the series is formally defined as follows:

\[
X(k) = \sum_{n=0}^{N-1} x(n) e^{-j\frac{2\pi nk}{N}}, \quad k = 0, 1, \ldots, N - 1
\]

Where \( e \) stands for the base of natural logarithms and \( j \) is the imaginary unit. In fact, DFT is the \( N \) points equispaced sampling of the spectrum \( X(k) \) of \( x(n) \) on\([0,2\pi]\), that is, the discretization of the sequence spectrum.

According to the collected data, the probability of users accessing each site every day (90 rows and 12 columns) and the probability of visiting each site at each time period (90 rows and 12 columns) as the input of discrete Fourier transform. Performing DFT on each column of data separately, which can obtain the corresponding amplitude-frequency map and phase-frequency map of each column of data, and extract the maximum amplitude value and corresponding phase value of each column of data as the feature vector of the user.

We take the user data with the ID of 192.168.1.132 as an example, select the probability data of the user visiting the search engine (S3) and the academic research (S6) every day in 90 days, and perform
discrete Fourier transform to obtain the corresponding amplitude-frequency plot and phase-frequency plot, which are shown in Figure 2. Figure 2(a) is the representation of search engine by DFT transformation. Figure 2(b) is the representation of academic research by DFT transformation.

![Probability Distribution of Accessing Search Engine Site](image1)
![Amplitude-Frequency Diagram](image2)
![Phase-Frequency Diagram](image3)

**Figure 2.** Amplitude-frequency diagram and phase-frequency diagram after DFT transformation.

![Probability Distribution of Visiting Academic Research Site](image4)
![Amplitude-Frequency Diagram](image5)
![Phase-Frequency Diagram](image6)

**Table 3.** Comparisons of visiting probability and amplitude between two types of websites.

| Site type       | Comparison items | 1     | 2     | 3     | 4     | … | 88    | 89    | 90    |
|-----------------|------------------|-------|-------|-------|-------|----|-------|-------|-------|
| Search Engines  | Visiting probability | 0.19149 | 0.14545 | 0.10909 | 0.08333 | … | 0.11111 | 0.18519 | 0.09091 |
|                 | Amplitude        | 0.00012 | 8.07E-05 | 0.0001 | 1.84E-05 | … | 1.84E-05 | 0.0001 | 8.07E-05 |
| Academic Research | Visiting probability | 0.3617 | 0.29091 | 0.36364 | 0.27083 | … | 0.27778 | 0.38889 | 0.12727 |
|                 | Amplitude        | 0.00027 | 0.00019 | 0.00033 | 0.00015 | … | 0.00015 | 0.00033 | 0.00019 |

By accessing the data of the two types of websites and Table 3, we can find that after DFT transformation for each column of data, the following characteristics in the frequency domain: the user accesses a site type frequently, and the number of daily visits is larger, after the discrete Fourier transform, the corresponding amplitude will be larger than that of other sites.

Therefore, in the result of transforming each column of data, the extracted feature is the amplitude and the phase corresponding to the component having the largest amplitude. The amplitude and phase of all the columns of the two probability tables are combined into a one-dimensional vector as the feature vector of the user. User feature vector with ID 192.168.1.132 is that User_feature_vector = \{0.000643 -2.0882 0.000016 -2.5473 0.00018 0.3519 0.000152 0.1383 0 0 0 0 0.000425 0.0048 \}
5. User Identification

Through the explanation of the contents of the previous section, the maximum amplitude \( (X_k) \) and its corresponding phase \( (\varphi_k) \) are obtained by DFT transformation for each user's access behavior, and the two values of the \( m \) columns are combined into a one-dimensional vector \( P \) to represent:

\[
P = [X_1 \varphi_1 X_2 \varphi_2 \ldots X_i \varphi_i \ldots X_m \varphi_m]
\]  

(1)

Where, the element \( X_i \) is the amplitude after the \( ith \) column transformation and the element \( \varphi_i \) is the phase after the \( ith \) column transformation, and the time interval \( \Delta t \) between the adjacent two elements is equal.

According to the extracted vector and the characteristics of the transform in the frequency domain, we can see that the meaning of \( X_i \) in the behavioral feature can be regarded as the amplitude of the probability change of the user visiting the current site. The meaning of \( \varphi_i \) in the behavioral feature can be regarded as the information of the time when users visit the current site behavior. Therefore, this one-dimensional vector can represent the user's browsing behavior pattern.

Preliminary prediction of anonymous users is implemented according to the known pattern categories, based on the thought of calculating the similarity between the test user and the behavior pattern of the known category. The Euclidean distance can reflect the absolute difference between the individual numerical features, it is suitable to analyze the difference in numerical value by the similarity distance from the perspective of dimension. Therefore, in our experiment, the matching process between the test user and the known pattern user feature vector is mainly based on the Euclidean distance calculation, and the distance \( D_i \) between the input test feature vector \( Y \) and the feature vector template \( P_i \) is defined:

\[
D_i = \|P_i - Y\|_2 = \left( \sum_{j=1}^{m} (p_{ij} - y_{ij})^2 \right)^{1/2} \quad (i = 0,1,\ldots,n), \quad (j = 1,2,\ldots,m)
\]  

(2)

Comparing the value of \( D_i \), which is took the minimum value. At this time, \( i \) is the number corresponding to the eigenvector of the known behavior pattern, which achieves the purpose of automatic recognition.

In the paper, the data of the first three months of the log record is used as the training data set, and the data of the last three months is used as the test data set. The test data is processed by the same method, and the feature vector of a test user data is extracted, which is Test_User_feature_vector = \{0.000176 1.1297 0.000072 1.2514 0.000384 0.9511 0.000202 0.8198 0.000145 -1.0685 0.000105 -1.9768 0.000035 -1.1516 0.002079 -0.7606 0.000018 -2.392 0.000587 -2.2448 0.000669 2.6279 0 0 0 0.001836 2.1892 0.000101 2.3017 0.000067 1.332 0.000237 2.4666 0.000398 -1.6326 0.000214 1.087 0.000278 1.6836 0.000028 -1.4569\}.

Before calculating the correlation between the test user and the known behavior pattern, the behavior pattern is normalized, then the behavior pattern of the test user data is calculated with each line feature vector of the known pattern category to obtain the matching degree between the two vectors, which is shown in Figure 3.
Users whose matching degree is less than 0.1 are selected as the criterion of user identification. It can be seen from the matching result graph that the matching degree of the 17th, 46th, and 90th users with the test data is less than 0.1. Therefore, users numbered 17, 46, and 90 are eligible for the article's selection of the test user recognition prediction.

The data of other test users are operated on the same way and obtained the number after each test user matching the known category behavior pattern, which took the matching degree less than 0.1 and arranged the user numbers according to the matching degree from small to large. Then the IP addresses of the test user and the matching user are compared in turn, if the IP address is the same, the user identification is correct.

The result in Figure 4(a) is the minimum matching degree between the test user and the matching user. The accuracy of the user identification is 40.90%. The result in Figure 4(b) is the top two rankings of the matching degree, and the accuracy of the user identification is 65.45%. The result in Figure 4(c) is the top three rankings of matching degree, and the user recognition accuracy rate is 77.27%.

By extracting the characteristics of the user behavior data, the probabilistic features of the test users and the known mode users can be initially obtained for visiting each site. Secondly, the user with the least matching degree is selected as the recognition result of the test user. Finally, it is judged whether the user's IP address is the same, and if the IP address is the same, the user identification is correct. It is proved by experiments that the method of discrete Fourier transform is used to extract the amplitude and phase of the network browsing behavior data as the feature vector of the user, which ensures the
high accuracy of the user recognition and reflects the behavior that the user may occur in a different
time. It also provides a reference for extracting the characteristics of user web browsing behavior.

6. Conclusion
While using the Internet, users also generate tens of thousands of message records. This paper mainly
describes the statistics of user's Web behavior data, the extraction of user behavior characteristics and
the establishment of the user identification process. First, the user behavior data is counted to obtain
two probability tables: (1) the probability of visiting each site every day in 90 days; (2) the probability
of visiting each site in 90 days at each time period. Then DFT is applied to each column of the two
table data to extract the user's eigenvectors, so as to match the user's eigenvectors with the known
patterns. Finally, the effectiveness and accuracy of the method for user identification are verified by
experiments.

Although the accuracy of user identification is low, the effectiveness of user online behavior
feature recognition can still be seen from time-frequency analysis. In the later research, we will try to
improve the user recognition rate by combining other user information, such as user location
information and the usage information of user mobile terminal.

7. Acknowledgments
This work is supported by the National Natural Science Foundation of China (No. 61462076).

8. References
[1] Niu Wen-Jia. (2016). User network behavior portrait: user network behavior portrait analysis
and content recommendation applications in large data. Publishing House of Electronics
Industry.
[2] Ju C, Wang J and Zhou G. (2018). The commodity recommendation method for online shopping
based on data mining. Multimedia Tools and Applications.
[3] Kušen, Ema, Strembeck M, Cascavilla G and Conti M. (2017). On the influence of emotional
valence shifts on the spread of information in social networks. Proceedings of the 2017
IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining
2017. ACM.
[4] Fang and Jing. (2014). Targeted advertising based on social network analysis. Applied
Mechanics and Materials, 488-489, 1306-1309.
[5] Li C and Jiang Z. (2016). A hybrid news recommendation algorithm based on user's browsing
path. IEEE/ACIS International Conference on Computer & Information Science. IEEE.
[6] Zhang J H and Zou Q. (2016). Group learning analysis and individual learning diagnosis from
the perspective of Big Data. IEEE International Conference on Cloud Computing & Big Data
Analysis. IEEE.
[7] Zhang Shu-Sen, Liang Xun, Qi Jin-Shan. (2017). A review on role identification methods in
social networks. Chinese Journal of Computers (3).
[8] Zheng F and Luo F. (2013). Network user identification method and application server
thereof. WO.
[9] Li Y, Zhang Z and Peng Y. (2017). A solution to tweet-based user identification across online
social networks.
[10] Esfandyari A, Zignani M, Gaito S and Rossi G P. (2016). User identification across online
social networks in practice: pitfalls and solutions. Journal of Information Science, 44(1).
[11] Feng S, Wang Q, Shen D, Kou Y, Nie T and Yu G. (2018). User identification across social
networks based on global view features. 2017 14th Web Information Systems and Applications
Conference (WISA). IEEE.
[12] Zhou X, Liang X, Du X and Zhao J. (2017). Structure based user identification across social
networks. IEEE Transactions on Knowledge and Data Engineering, 1-1.
[13] Zafarani R, Tang L and Liu H. (2015). User identification across social media. ACM
Transactions on Knowledge Discovery from Data, 10(2), 1-30.
[14] Wenxin LI, Zhang D and Zhuoqun XU. (2008). Palmprint identification by fourier
transform. *International Journal of Pattern Recognition and Artificial Intelligence, 16*(04), 417-432.

[15] Liu Xiang, Wang Bin-Jun, Wang Jing-Ya and Du Jin. (2015). Periodic behavior patterns implied in spatiotemporal trajectories. *Science Technology and Engineering, 15*(35), 197-203.

[16] Zhu C, Li B, Li A and Zhu T. (2017). Predicting depression from internet behaviors by time-frequency features. *IEEE/WIC/ACM International Conference on Web Intelligence*. IEEE.

[17] Wu Guo-Ning, Qi Jing-Jing and Zhou Ya-Tong. (2017). Time-frequency analysis method: research and prospect. *Journal of Image and Signal Processing*, 07(01), 24-35.