Research on the Relationship Between Chinese Nicknames and Accounts in Social Networks

Yi Han, Xiangyu Liu, Yanhui Du^(✉), and Tianliang Lu

People's Public Security University of China, Beijing 100038, China
duyanhui@ppsuc.edu.cn

Abstract. With the increasing integration of mobile Internet into people's daily life, generally a user will register several different network applications at the same time. Therefore, there are many virtual identities belonging to one person on the Internet, and the similarity analysis of cross-platform network identities is of great significance in the field of network security. This paper studied the Chinese user nicknames and virtual identity recognition in domestic social networking. Considering that account nicknames, to some extent, can reflect the characteristics of the account owners' naming habits and preferences, and that the information of nickname is easier to obtain than other registration information, we collected the nickname information of users who registered on three application platforms: WeChat, Weibo and Alipay. At the same time, according to the characteristics of account nicknames and their probability distribution, we determined the characteristic indicators that can be used to calculate the similarity of nicknames. Finally, this paper optimized the Jaro distance and Jaro-Winkler distance algorithms, and proposed an identity algorithm suitable for domestic social networks, especially calculating the similarity of Chinese nicknames, and verified the effectiveness of the algorithm on the basis of large-scale real data.

Keywords: User nickname · Similarity · Jaro distance · Virtual identity

1 Introduction

With the rapid development of mobile Internet, various social software has greatly enriched people's lives, and the scale of Internet users in China is also growing. By December 2020, the total number of Internet users in China has reached 940 million [1]. Research shows that about 84% of Internet users have more than one social networking account [2]. At the same time, with the characteristics of multiple identities, virtualization and anonymity of social networks becoming more and more obvious, some lawless elements spread extreme remarks through the network, which seriously endangered social order.

In recent years, a lot of attention has been paid to the study of social network user identity, and the analysis of network account association can provide necessary information for the safety supervision of public opinion in social networks. Considering that user nicknames are easier to obtain than other registration information, and to some extent, it

can reflect the naming habits and feature preferences of account holders. Therefore, this paper collects user nicknames from domestic mainstream social networking sites, and proposes a method for calculating account similarity based on Chinese user nicknames.

- This paper investigates the similarity of social network accounts, and finds that most of the current research results are oriented to foreign social platforms such as Twitter, Facebook, Myspace [3–5], which is not applicable to account analysis of the top Apps in the domestic application market, and designs a prototype system that can push users' nicknames in batches by using the address book matching and friend recommendation functions of App, and constructs a database for studying the similarity of domestic SNS accounts.
- This paper summarizes some features of Chinese nicknames, analyzes the consistency of the features of the same natural person on different platforms, and selects character features as the only index of the algorithm. Specifically, the paper converts Chinese characters into English characters by the way of phonetic conversion, which solves the problems of homophones and homophones in account similarity judgment.
- This paper proposes a similarity algorithm suitable for Chinese nicknames, aiming at the common characters in different strings, such as “head-head”, “tail-tail” and “head-tail”. Improve the calculation weights when the calculation weight when the characters of “tail-tail” and “head-tail” are consistent, and the effectiveness of the algorithm is verified on the experimental data set.

In the first section, the paper introduces the relevant research work. The second section introduces the method of obtaining account nickname data and its implementation process; The third section sorts out some characteristics of Chinese accounts, and analyzes the consistency of characteristics of accounts on different platforms; In the fourth and fifth sections, an account similarity algorithm suitable for Chinese nicknames is proposed, and the effectiveness of the proposed algorithm is verified by real data Finally, the sixth section summarizes the full paper.

2 Related Work

2.1 Research Status

At present, researchers usually design network account association analysis based on three different dimensions: user attribute, user behavior and user relationship. Among them, user attribute features mainly include user nickname, avatar, birthday, place of residence, personality signature, etc. User behavior characteristics mainly include online speech, behavior track, participation topic, etc. The characteristics of user relationship mainly include attention, interaction between fans and friends, joining groups and so on. In the research field of user nickname association analysis, most of them are related to English-speaking users such as Twitter and Facebook, and have achieved good results. In view of the unequal information that can be collected by the above-mentioned user characteristics in different social networks, this paper mainly studies the research status of user nicknames in China and abroad.

Reza Zafarani et al. [6] listed 414 potential features related to user names, and proposed a calculation method of multi-platform virtual identity association by calculating the most important 10 features. Siyuan Liu et al. [7] give weight to user names, and then carry out association analysis on user virtual identities based on statistical analysis method. Perito et al. [8] used n-gram model to measure the uniqueness of user names, and finally analyzed the similarity of user names by editing distance. Dong Liu et al. [9] made statistics according to the user's explicit length, the frequency of using special characters, the use of numeric characters, the combination mode of keyboard input mode and user name, and obtained probability distribution and feature analysis, and then proposed a method of identity verification. Y Li et al. [10] have done more detailed feature engineering on the username of Chinese users, considering the existence of the user name in simplified and traditional Chinese or English letters and homophones, a language mapping method based on Pinyin is proposed. Siyuan Liu et al. [11] mentioned the situation of Chinese characters in the study of association analysis of virtual accounts, but in fact, they did not carry out in-depth discussion and did not propose corresponding solutions.

2.2 Existing Problem

Although the research on user-oriented nickname association analysis has important research significance, the research based on Chinese user nicknames is still in its infancy, and literature and research results at home and abroad are extremely scarce. At present, Chinese-oriented research fields mostly focus on natural language analysis, semantic understanding and other corpus studies with a certain length, but there are few short Chinese analyses on user nicknames, and the feature extraction and correlation analysis techniques of Chinese user nicknames are quite different from those of English users. Therefore, the research on nickname association of Chinese users is facing great challenges, and there are also many problems to be solved, among which the most prominent ones are as follows:

1. Traditional Chinese characters. When registering account nicknames, mainland users prefer simplified Chinese, while users in Taiwan Province, Hong Kong and Macau generally use traditional Chinese.
2. Disassembly of Chinese characters. For example, the same natural person has registered usernames "弓长佩" and "张博佩" on different social platforms, so how to establish account association with user names with such characteristics is a problem worthy of in-depth study. For another example, accounts such as "萌正太King" and "默念你的爱" also take the dismantling of Chinese radical as a part of nicknames. How to effectively identify and classify them is also of great significance in the analysis of users' nickname features.
3. Homophonic. The problem of homophonic words has always been a hot spot in the field of Chinese research, and there is also the problem of homophonic words in the research of nickname of Chinese users. For example, "橙丸村" and "陈婉春" are different network accounts belonging to the same natural person. How to solve Chinese homophones and homophones is also one of the problems that need to be solved in this paper.

4. **Emoji expression.** With the gradual liberalization of character restrictions on user nicknames in different social platforms, some post-90s, post-95s and even post-00s network users have added personalized characters to their nicknames, and many personalized nicknames such as "西瓜红烧才好吃🍉", "Eva👩的母后" and "董婷💖🐼🐼🐼" have brought great difficulties to Chinese account analysis.

2.3 Research Opportunities

In the early days, the data sources of virtual identity account research were scarce, and most experimental data sets were mainly collected by web crawlers to obtain attribute information related to users, such as name, gender, birthday, occupation and so on. Most of this kind of characteristic information is incomplete, and its authenticity is difficult to verify, so it is difficult for researchers to obtain it for research on high-quality dataset. As for user nicknames, literature [12] has high credibility in associating different usernames of the same user through social investigation, but this method consumes manpower and time, and the respondents are generally unwilling to provide their account information to others. There are also some social platforms that allow registration-free login through other website accounts, but the jump view function is limited on the personal homepage of this account, and researchers cannot obtain the account information associated with the target account on other platforms. In addition, some public data sets, such as Google+, Facebook, Twitter, etc., are mainly English-speaking and native-speaking users in the West and Europe, and are not the research objects of this paper.

In recent years, with the promotion of real-name registration system, a user of social networks in China, more and more social platforms need to bind their mobile phone numbers when registering. Although the mobile phone numbers are not public, it virtually provides the possibility for the research of virtual identity association analysis. Literature [13] introduces that most network applications provide the function of "address book matching", and through this function, the mapping relationship between mobile phone number and virtual account number is established. In addition, literature [14] introduces many cases of personal privacy leakage, and literature [15] puts forward a method of obtaining account related information by using network App communication traffic, and makes use of it. Therefore, the above-mentioned real data related to the user's mobile phone number and nickname provide a new idea for the experimental data research of virtual identity account association analysis.

3 Data Collection and Implementation

In order to construct nickname samples in China, this paper selects three applications with the highest download volume in Android mobile phone application market: Weibo, WeChat and Alipay as data sources (Table 1), and filter users who register and customize nicknames in the three applications at the same time as seed samples.

In addition, the paper randomly selects 7296 accounts from the leaked user data of Sina Weibo in March 2020, and forms the mapping relationship between nicknames and mobile phone numbers of Weibo users according to the corresponding user nicknames associated with Weibo account id.

Table 1. Statistics of monthly active users of different network applications

Network applications	Number of monthly active users	Date	Reference
Weibo	530 million	2021.5.10	Second-quarter earnings in 2020
Wechat	1.242 billion	2021.5.20	Second-quarter earnings in 2020
Alipay	711 million	2020.6	Prospectus of Ant Group

3.1 Information Acquisition and Integration Analysis

Literature [8] introduces the method of network account matching. The paper matches the address book of the above 7296 mobile phone numbers, and obtains the account nicknames of WeChat and Alipay. The specific methods are as follows (Fig. 1):

Fig. 1. The flow chart of matching mobile phone number with network account

- Step 1 Import the mobile phone number to be processed into the mobile phone address book, and view/add friends in the address book through the target application.
- Step 2 App uploads the address book information to the application server in the form of original text or abstract information.
- Step 3 Server returns the user account information matching with the mobile phone number in the address book (registration binding).

The statistics of account number matching results are shown in Table 2.

Table 2. Network platform account matching results

Total	Weibo	Wechat	Alipay	All the three platforms
7296	7296	5683	6055	5050

3.2 Acquisition Module Design and Implementation

This paper designs and implements the matching function based on mobile phone address book. The prototype system for obtaining user nicknames is divided into three modules, namely, address book import module, information acquisition module and content extraction module. The address book import module is responsible for automatically loading the target mobile phone number and inputting it in the standard format readable by the address book; The information acquisition module is responsible for uploading the address book information to the target application server and acquiring the address book user account information from the server; Based on the analysis of the target application, the content extraction module realizes the automatic collection of nickname information of user accounts in the address book.

The prototype system acquisition terminal realized in this paper is based on Google Pixel equipped with Android native system version. This choice is mainly due to the following considerations: First, the Android system has a high market share of mobile devices, and it has a wider application value to use Android system as the target system. Second, most networks apps developers provide Android version of application software, so there will be no shortage of application collection coverage; More importantly, Android system is an open source software project, and the system carried by Google Pixel is a native operating system, so there is no version customization and secondary development, which greatly facilitates the design and implementation of the prototype system.

The specific method is as follows: firstly, the mobile phone number to be analyzed is processed into VCF format file, empty the original address book in the mobile phone; And then Import VCF format file into the phone address book; The information acquisition module pre-analyzes the trigger mode of the address book reading function in the target application, automatically triggers the function, and realizes the acquisition of address book user account information from the server; The content extraction module obtains user nicknames by analyzing the target application interface, and the module can extract user nicknames in batches; Finally, the user nickname and the corresponding mobile phone number are stored in the database (Fig. 2).

Fig. 2. The frame diagram of acquisition module

4 Data Collection and Implementation

4.1 Universal Feature

As a hot research direction of social network, account association technology usually adopts account characteristics including: 1) length characteristics. The length of most usernames is within a specific range, which is generally not too short or too long, which can visually show the length characteristics of user names. 2) Character type. Chinese characters are the main nicknames of domestic social network users. Compared with the western language system with only 26 English letters, any character or a group of characters in the nicknames of domestic network users have unique symbols for judging a natural person. 3) Special characters. Some nicknames containing numbers and symbols can reflect the naming habits and preferences of users. 4) Combination mode. Combination mode can also be used as one of the characteristic indexes in judging nickname similarity, that is, the same natural person should have the same or similar combination mode on different platforms. In order to further apply the above features to the Chinese language environment, the paper also needs to make statistics on the above features, as shown in Table 3.

Table 3. Consistency analysis of user nickname features in different platforms

Length characteristic	Character type	Special characters	Combination mode
0.047	0.652	<0.001	0.076

4.2 Feature Selection

Through the statistics of the above characteristics of WeChat, Weibo and Alipay, it is found that users register nicknames on different platforms. The length characteristics,

special characters and combination patterns are not consistent. The main reasons are as follows: the character types supported by nicknames of users on different platforms are inconsistent, and the expressions of special characters and emoji are quite different among different operating systems, coding methods and input methods; Weibo nicknames have the unique characteristics of the whole network. In order to avoid duplication with other nicknames, users often add numbers, letters and other characters, resulting in long overall characters and diverse combination modes. However, this kind of situation rarely occurs in nicknames of WeChat and Alipay. Therefore, length features, special characters and combination patterns are not considered as similarity calculation factors.

In terms of character features, due to the uniqueness of Chinese characters, users are usually accustomed to using fixed or similar Chinese characters. Therefore, in the above statistical results, most users show good consistency in character features. In addition, every Chinese character in Chinese has Chinese Pinyin, and the Chinese Pinyin Scheme published in 1958 clearly stipulates that Chinese Pinyin adopts the internationally accepted Latin alphabet. Therefore, this paper takes character features as the core index to calculate similarity, converts Chinese into English writing form by the method of phonetic conversion, and then calculates account similarity by calculating text similarity. This method can effectively solve the conversion between simplified Chinese and traditional Chinese characters, and at the same time, it can play a better conversion effect in homophonic words with the same pronunciation and similar pronunciation.

5 Algorithm Design

5.1 Jaro Distance

Jaro distance is one of the most commonly used methods to judge the similarity of short texts based on the number and sequence of commonly used characters. It was originally used to judge whether two names of health records are the same in census, so it is the best selection for matching user names. This section is further optimized based on Jaro distance algorithm to improve nickname recognition ability suitable for domestic network users.

For any two strings, the matching window under Jaro algorithm is defined first as follows: the comparison between characters should be limited to a certain range or condition, besides if two characters are equal in this range, indicates a successful match; If it exceeds this range, the match fails. In Jaro algorithm, it is defined as formula (1), that is, mw does not exceed the value of the following expression:

$$w \leq \frac{\max(|S_i|, |S_j|)}{2} - 1 \quad (1)$$

Where, $|S_i|$ and $|S_j|$ represent the lengths of the string S_i and S_j . At the same time, if the matching sequence characters are reversed, the number of characters in the reversed sequence is recorded as the transposition number (tn for short) in Jaro algorithm. Therefore, based on the logic of matching window mw and transposition number tn , the Jaro distance calculation formula of string S_i and S_j is as follows:

$$JD_{ij} = \begin{cases} 0 & \text{if } m = 0 \\ \frac{1}{3} \left(\frac{m}{|S_i|} + \frac{m}{|S_j|} + \frac{m-t}{m} \right) & \text{otherwise} \end{cases} \quad (2)$$

Where m represents the number of matching characters of two strings based on the matching window logic; t represents $\frac{1}{2}m$ under transposition number logic.

5.2 Jaro-Winkler Distance

Given the importance of English prefixes, Jaro-Winkler distance is further modified basis on Jaro distance, meaning that, if two strings of the first few characters are the same, they will be more similar. The formula of the algorithm is shown as formula (3):

$$JWD_{ij} = JD_{ij} + lp(1 - JD_{ij}) \quad (3)$$

In which JD_{ij} represents Jaro distance based on Jaro algorithm; l indicates the number of common prefix characters of two strings (maximum no more than 4); p is a scale factor constant, that describes the contribution of common prefix to similarity. The larger the p is, the greater the weight of common prefix is (the maximum value not more than 0.25, and the default value is 0.1).

It can be seen that jaro-winkler distance algorithm is more friendly to prefix matching, but there is still a certain degree of misjudgment when calculating the similarity of user nicknames. For example, “芸” and “白芸” are nicknames of the same natural person on different platforms, but their Jaro-Winkler similarity is 0, so they are judged to be a group of unrelated accounts, which is a misjudgment. From here, we can make a conclusion that Jaro-Winkler distance algorithm has an obvious error in nickname similarity analysis, in particular when a group of nicknames keywords are at the end of a string, and the likelihood of this occurrence is relatively high. Therefore, the Jaro-Winkler distance is further optimized.

5.3 Text Algorithm

Taking into account the habit of bidirectional combination of last name and first name in nicknames, the matching weights of key characters of “first-first”, “tail-last” and “first-last” should be considered at the same time. Therefore, the two algorithms are further revised in this paper, and the modified distance formula is as shown in formula (4). The paper is called Jaro-Winkler-Plus distance algorithm:

$$JWPD_{ij} = JD_{ij} + \max\left(l_h p_h, l_t p_t, \max(l_{ih}, l_{jh}) \frac{(p_h + p_t)}{2}\right) (1 - JD_{ij}) \quad (4)$$

In which: JD_{ij} is Jaro distance based on Jaro algorithm.

l_h represents the number of character with common prefix of two strings, where l_{ih} and l_{jh} represent the number of common S_i prefix and S_j suffix, and the number of common S_i suffix and S_j prefix, l_t represent the number of common suffix of two strings, and satisfy the minimum value of arbitrary $l_i (i = h, t, ih, jh)$ is no less than the minimum number of characters 1 (such as ‘a’) and the maximum number is no more than the maximum number of characters 6 (such as “Zhuang”).

p_h is a common prefix scaling factor constant, and p_t is a common suffix scaling factor constant, any $p_i (i = h, t)$ satisfy $0.1 \leq p_i \leq 0.165$, in order to ensure $\max\left(l_h p_h, l_t p_t, \max(l_{ih}, l_{jh}) \frac{(p_h + p_t)}{2}\right) \leq 1$ Therefore, under the premise of $l_i \leq 6$, the maximum value is also reduced from the initial 0.25 to 0.165.

It can be seen that Jaro-Winkler-Plus distance has obvious advantages in measuring text similarity compared with Jaro and Jaro-Winkler distance, which is shown as follows: First, it is better suitable for different habits of different users writing surnames + firstnames (such as "张三疯"、" " and "疯子张"), especially Chinese people pay special attention to the last names, and are willing to use it repeatedly in social media or daily communication; Second, adding "tail-to-tail" and "head-to-tail" influencing factors on the basis of "head-to-head" instead of summary and induction, but adopts the principle of giving priority to the maximum value can avoid the excessive influence of head-to-head strategy on total Jaro distance.

6 Experiment and Analysis

6.1 Data Description

The data sources used in this paper are 5050 groups of user data processed in Sect. 3.1, all of which are registered on Alipay, Weibo and WeChat, and their nicknames are customized. In order to facilitate the calculation, this paper converts the nicknames of each user on three platforms, that is, the phonetic sequences without tones are recorded as strings. By random selection, this paper constructs three experimental data sets, each of which contains 5000 positive examples (user name pairs belonging to the same natural person) and 5000 counterexamples (randomly combining user name pairs of different natural persons).

6.2 Index Evaluation

Precision, Recall and F-Score are used as evaluation criteria to measure the performance of the algorithm. Specific definitions are shown in formulas:

$$Precision = tp / (tp + fp) \quad (5)$$

$$Recall = tp / (tp + fn) \quad (6)$$

$$F - Score = 2 \times pc \times rc / (pc + rc) \quad (7)$$

F-Score is the harmonic mean of Precision and Recall, and it is the total evaluation index of the algorithm performance. The three evaluation indicators in the above formulas are based on the following three indicators: tp (true positive) refers to the number of account pairs that are correctly judged as the same user by this algorithm. fp (false positive) is the number of account pairs incorrectly judged as the same user. fn (false

negative) represents the number of account pairs that are judged to be different users but are the same user. In our experiment, a threshold is required to be set to determine whether two nicknames belong to the same user account. If the distance is greater than the threshold, it is determined that the two nickname strings participating in the comparison belong to the same natural person. If the distance is less than or equal to the threshold, it is determined that the two nickname strings participating in the comparison do not belong to the same natural person. Here, the paper needs to flexibly adjust the threshold to balance the Precision and Recall. As shown in Fig. 3, Fig. 4 and Fig. 5, with the increase of the judgment threshold, Precision of the three groups of data also increases correspondingly, while Recall decreases obviously. The main reason for this result is that when the threshold is increased, the more severe the judgment condition is set, the number of username pairs that the algorithm judges to be a match also decreases. In addition, the number of misjudged username pairs also increased, resulting in a decrease in Recall.

Fig. 3. Precision value

Fig. 4. Recall value

Fig. 5. F-Score value

Through the trend analysis of the value, it seems that 0.5–0.55 is a reasonable threshold range. Here, the paper defines 0.5 as the threshold for judging account similarity.

6.3 Comparison of Methods

After the threshold is determined, Jaro distance and Jaro-Winkler distance are used to calculate nickname similarity, About 7.07% of the data distance increases slightly, but the influence degree does not change for judging whether they belong to the same user, Therefore, the results of the two distance algorithms are consistent. The algorithm proposed in this paper not only greatly improves the distance value, but also improves the judgment accuracy by about 9.12%. This shows that the nickname similarity calculation method proposed in this paper can better quantify user nicknames, and better identify and discover different network accounts belonging to the same natural person in Chinese environment (Table 4).

Table 4. F-Score of different algorithm

Algorithm	F-Score		
	Data Set 1	Data Set 2	Data Set 3
Jaro	0.6085	0.6238	0.6262
Jaro-winkler	0.6085	0.6238	0.6262
Jaro-winkler-plus	0.6921	0.7082	0.7319

In this paper, the proposed nickname similarity calculation method is applied to the user data of the three platforms of WeChat, Weibo and Alipay. Since the nickname custom rules of each platform are quite different, this paper can distinguish account similarity only by analyzing the character characteristics, indicating that this research direction has good value potential. In the future, better judgment results will be obtained by combining more platforms and larger-scale user data and extracting more feature attributes that can be used to calculate similarity, such as gender, age, birthday, hobbies and other features.

7 Conclusion

In view of the difficulty in identifying multiple virtual identities of domestic social network users, and the lack of data resources in domestic research on virtual identity accounts, considering that account nicknames can reflect the naming habits and feature preferences of holders to a certain extent, and the nickname information is more open and transparent than other information, this paper studies the similarity analysis of online accounts. First of all, the paper uses the “address book matching” and “friend recommendation” functions of social software to collect nickname data from three mainstream apps in China; Secondly, the paper puts forward some common characteristics of account nicknames, and makes statistics and consistency analysis on the nicknames registered by the same user on three platforms, and determines the characteristic indexes for calculating account similarity; Finally, Jaro distance algorithm is improved, and an account similarity calculation method suitable for Chinese nicknames is proposed.

Through experimental data verification, it is concluded that the proposed similarity calculation method based on account nickname character features is 9.12% more effective than the traditional Jaro distance algorithm, which is better applicable to the Chinese language environment dominated by Chinese characters and provides effective support for the identification of domestic netizens. In the next step, we can make a more in-depth study on the similarity determination of network virtual identity by combining the characteristics of users in other aspects, such as gender, age, friend relationship and other account attributes.

References

1. The 46th China Statistical Report on Internet Development[R/OL], 29 September 2020. <http://www.cnnic.net.cn/hlwfzyj/hlwxyzbg/hlwjtjbg/202009/P020200929546215182514.pdf>
2. Bartunov, S., et al.: Joint Link-attribute user identity resolution in online social networks. In: Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis, pp. 104–109. Beijing, China (2012)
3. User identification across multiple social networks based on information entropy. *J. Comput. Appl.* **37**(8), 2374–2480 (2017)
4. Motoyama, M., Varghese, G.: I seek you-searching and matching individuals in social networks. In: Proceedings of the 11th International Workshop on Web Information and Data Management, pp. 67–74. Hong Kong, China (2009)
5. You, G.W., Hwang, S.W., Nie, Z., Wen, J.R.: SocialSearch: enhancing entity search with social network matching. In: Proceedings of the 14th International Conference on Extending Database Technology, pp. 515–519. Uppsala, Sweden (2011)
6. Zafarani, R., Liu, H.: Connecting users across social media sites: a behavioral-modeling approach. In: Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM (2013)
7. Kumar, S., Zafarani, R., Liu, H.: Understanding user migration patterns in social media. In: AAAI’11
8. Perito, D., Castelluccia, C., Kaafar, M.A., Manils, P.: How unique and traceable are usernames? In: Fischer-Hübner, S., Hopper, N. (eds.) *Privacy Enhancing Technologies. PETS 2011*. LNCS, vol. 6794, pp. 1–17. Springer, Berlin, Heidelberg (2011). https://doi.org/10.1007/978-3-642-22263-4_1

9. Liu, D., Wu, Q.Y., Han, W.H., Zhou, B.: User identification across multiple websites based on username features. *Chin. J. Comput.* **38**(10) (2015)
10. Li, Y., et al.: Connecting Chinese users across social media sites. In: *ICMIA*, pp. 1273–1279 (2015)
11. Liu, S., Wang, S., Zhu, F., Zhang, J., Krishnan, R.: HYDRA: large-scale social identity linkage via heterogeneous behavior modeling. In: *Proceedings of the ACM SIGMOD* (2014)
12. Liu, J., et al.: What's in a name an unsupervised approach to link users across communities. In: *Proceedings of the International Conference on Web Search and Web Data Mining*, pp. 495–504. Rome, Italy (2013)
13. Cheng, Y., et al.: Research on user privacy leakage in mobile social messaging applications. *Chin. J. Comput.* **37**(1), 87–100 (2014)
14. Jin, P.-Y.: The leakage and protection of personal privacy data in the era of big data. *J. Tongji Univ. (Soc. Sci. Sect.)* **31**(3), 18–29 (2020)
15. Yue, H.-Z., Zhang, Y.-Q.: Vulnerability analysis and exploitation of location privacy leakage in webcasting platforms. *Chin. J. Comput.* **37**(1), 87–100 (2014)
16. Zhang, J.: Design of patient identity matching method and implementation of EMPI system (Social Science Section) **31**(3), 18–29 (2020)

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.