Connecting Chinese Users Across Social Media Sites

Yanan Li[1, a *], Junxing Zhu[2,b], Zhongcheng Zhou[3,c], Bin Zhou[4,d] And Xiaobo Wu[5,e]

[1,2,3,4] College of Computer National University of Defense Technology

Changsha, China

[5]School of software engineering Yantai Vocational College

Yantai, China

[a]toward-behind@163.com, [b]daishanshen5036@Sina.com, [c]ceaserz@163.com, [d]binzhou@nudt.edu.cn, [e]happywxb@163.com

Keywords: virtual identity; Chinese; username; bigdata; pinyin.

Abstract. The usage of social network usernames in the research of social identity linkage has been proved, especially for the English usernames. However, how to properly connect Chinese user identities by matching the usernames remains to be explored. Since a Chinese user may name or rename his/her usernames in different ways (e.g. using a Chinese username or translating it into English, using simplified Chinese characters or traditional ones, converting some words in a given username to their homophonic words, etc.), it is more difficult to connect Chinese users than connect English users by matching usernames. This paper proposes a kind of language mapping method which can translate different type of Chinese words of a given username into their corresponding Pinyin words. However, the number of user identities in a given social network can be very large, thus the username matching process between two social networks is very costly. Basically, we use the Hadoop and Spark frameworks to conquer the efficiency problems. We also have a study on various username matching algorithms, and figure out the features that are useful in Chinese username matching.

Introduction

Social networks are more and more important because of their important roles as user participations in one kind or another. They not only enrich our daily life, but also have great commercial potentials by their huge amount of users and data. However, comparing to their real world lives, social network users have less responsibilities for what they said and done due to the anonymous feature of social networks. People publish personal opinions with different accounts in different social platforms. Due to the fragmentary characteristic in social network platforms, the users' information is fragmented and the commercial potential of social network has a great deficit. If we could connect user identities in multiple social network platforms, it will enrich the information about the users and help us to improve the situation of single information source. There are several advantages by connecting user identities in multiple social platforms: (1) *Completeness*. The user profiles [1] are more specific and complete, then it will be better for describing the users. (2) *Accuracy*. Based on the proverb--"a thousand lies are needed to hide one", the accuracy of data which is received by connecting user identities in social network platforms is higher.

The usernames of social users are easy to be obtained and its importance to the research of social identity linkage has been proved in [5]. In this paper, with the help of the big data processing tools, we try to connect Chinese user identities by matching the usernames. The section 2 introduces the related works. Section 3 describes the connecting Chinese user identities by matching the usernames, followed by the experiment in section 4. Section 5 concludes this research with directions for future work.

Connecting Chinese User Identities By Matching Usernames

Each of social network platforms requires a user to register a unique username. But our memory is limited [6] and the authors of [7] have a conclusion that people tend to choose same or similar usernames when they are registering with social network platforms. There are many algorithms that can calculate the similarity of usernames, such as edit distance, longest common substring, longest common subsequence, jaccard similarity coefficient, cosine similarity and Jensen-Shannon divergence. But which algorithm can achieve the best effect or what the optimal threshold is on each algorithm in counting the comparability of Chinese users' usernames should be tested. In this paper, the problems and modes in matching process are the emphasis in our research. The problems of language mapping and the related matching algorithms in Chinese are as follows.

Language Mapping Method. Social network has a large-scale application around the world due to its properties. There are different languages and characters in different areas, which makes users to have different tendencies to select their usernames in social network. Chinese users may use English or hybrid of English and Chinese as their usernames due to the popularity of English. This paper just consider the characters of Chinese and English, other languages or symbols (excluding the space) are regarded as special characters.

Chinese includes simplified Chinese and traditional Chinese. A Chinese user selects simplified Chinese or traditional Chinese as his/her username because of the preferences and regions. At the same time, there exists wrongly written characters or homophone when a user selects his username. Intuitively, it is nature to use pinyin to handle with the above problems. Besides, it can let us manipulate Chinese in an English-like way, which is more suitable for counting the comparability of usernames.

A Chinese character may map to several English characters by pinyin, and that leads to two questions which need to be verified in experiments.

1. After converting a Chinese character to pinyin, each character of pinyin is considered as a basic unit. For example, the pinyin-- “dao dan” , has 6 basic units in that situation, which are ‘d’ ,’a’, ‘o’, ’d’, ’a’ and ’n’.

2. After converting a Chinese character to pinyin, the pinyin is considered as a basic unit. For example, “dao dan”, it has two basic units, which are ‘dao’ ,’dan’.

The space character is always meaningless. But when a Chinese user selects username in social network, whether the space character is useless or not and what the importance is also need to verify.

Which matching algorithm is more effective and which is more suitable for Chinese also need to test. So the following are what we will verify.

1. Getting rid of the space in username, after converting, each character of pinyin is a basic unit (Char).

2. Getting rid of the space in username, after converting, the pinyin is a basic unit (Unit).

3. Containing the space in username, after converting, each character of pinyin is a basic unit (CharSpace).

4. Containing the space in username, after converting, the pinyin is a basic unit (UnitSpace).

Matching Algorithms. Here are the algorithms, which are used by the paper.

1, Edit Distance. Edit Distance is also called Levenshtein distance, it is a way to quantify how similar two strings are by counting the minimum number of operations, which are required to transform one string into the other. The allowed operations are the removal or insertion of a single character, or the substitution of one character for another.

2, Longest Common Substring. Longest common substring is to find the longest string that is a substring of two or more strings and the longest string must be continuous.

3, Longest Common Subsequence. Longest common subsequence is to find the longest subsequence common to all sequences in a set of sequences in a set of sequences. Unlike substring, subsequences are not required continuous.

4, Jaccard Similarity Coefficient. The Jaccard similarity coefficient, also known as the Jaccard index, is a statistic used for comparing the similarity and diversity of sample sets. The Jaccard similarity coefficient measures similarity between finite sample sets, and is defined as the size of intersection divided by the size of the union of sample sets. The equation is

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|}. \tag{1}$$

5, Cosine Similarity. Cosine similarity is a measure of similarity between two vectors of an inner product space that measures the cosine of the angle between them. The cosine of 0° is 1, and it is less than 1 for any other angle. It is thus a judgement of orientation and not magnitude: two vectors with the same orientation have a cosine similarity of 1, two vectors at 90°have a similarity of 0. It is particularly used in positive space, where the outcome is neatly bounded in [0, 1]. The figure of cosine similarity is following,

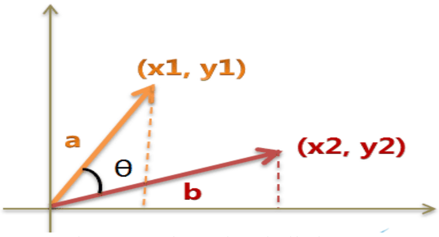


Figure 1. The cosine similarity

And the equation is

$$\cos\theta = \frac{a*b}{\|a\|*\|b\|} = \frac{\sum_{i=1}^{n}(x_i*y_i)}{\sqrt{\sum_{i=1}^{n}(x_i)^2}*\sqrt{\sum_{i=1}^{n}(y_i)^2}} \tag{2}$$

6, Jensen-Shannon divergence. In probability and statistics, the Jensen-Shannon divergence [9] is a popular method of measuring the similarity between two probability distributions. It is based on the Kullback-Leibler divergence, with some notable and useful differences, including that it is symmetric and it is always a finite value. The equation is the following:

$$D_{JS}(\mathrm{P}\,||\mathrm{Q}) = \frac{1}{2}\left(D_{KL}(P\,||M) + D_{KL}(Q\,||M)\right) \tag{3}$$

Where M(i)=1/2 (P(i)+Q(i)) and $D_{KL}(P\,||M)$ is the Kullback-Leibler divergence of P and M. Finally, the equation of the Jensen-Shannon divergence [9] is

$$D_{JS}(\mathrm{P}\,||\mathrm{Q}) = \frac{1}{2}\left(\sum_{i=1}^{n} P(w_i)\log_2\frac{P(w_i)}{M(w_i)} + \sum_{i=1}^{n} Q(w_i)\log_2\frac{Q(w_i)}{M(w_i)}\right) \tag{4}$$

Where w_i is the i-th letter in a string and satisfies $\sum_{i=1}^{n} P(w_i) = 1$.

Experiments

Data Preparation. The data used in the experiments are from Sina and Twitter. On the basis of existing data, we extracted 488906 Twitter users from about 236 million users based on whether it contains Chinese characters in users' self-introduction by Map/Reduce. Excluding Japanese users in extracted data, then we got 381914 Twitter users as Chinese users in Twitter. We found that the nicknames of Twitter users are similar to their usernames in Sina during the data processing. So we made an experiment with the usernames in Sina and nicknames in Twitter. We merged the Twitter users who have the same nicknames and obtained 289620 users finally because the nickname was not unique.

We utilized the websites of "about.me" and "GitHub", guaranteed to belong to the same individual. Then, 486 pairs are collected. Meanwhile, we also gained 240 pairs depending on the Sina interlinkage in Twitter user profiles. At a result, we have 726 pairs which called "associated dataset". Then it intersected with 289620 Twitter users and got 289836 Twitter users at last.

Normalization. We employed Spark in our experiment because the computational complexity of each matching algorithm is huge, which exceeds the processing power of traditional uniprocessing. In the 6 matching algorithms, the return value of edit distance, longest common substring and longest common subsequence is an integer, while others are double. Meanwhile, the smaller the value, the more similar in both edit distance and Jensen-Shannon divergence, then it was not conducive to measure consistency of algorithms. Therefore, we normalized the measurement of the algorithms.

Assuming user-sina as a Sina user's username and user-twitter as the nickname of a Twitter user, then we have a conclusion,

1, the ratio of edit distance

$$ED=1.0-\frac{ed*2}{(user-sina).split(" ").length+(user-twitter).split(" ").length} \tag{5}$$

Where ed is an integer of the return by edit distance.

2, the ratio of longest common substring

$$LCS=\frac{lcs*2}{(user-sina).split(" ").length+(user-twitter).split(" ").length} \tag{6}$$

Where lcs is an integer of the return by longest common substring.

3, the ratio of longest common subsequence

$$LCX=\frac{lcx*2}{(user-sina).split(" ").length+(user-twitter).split(" ").length} \tag{7}$$

Where lcs is an integer of the return by longest common Subsequence.

4, Jensen-Shannon distance

$$JS= 1.0 - js \tag{8}$$

Where js is the return by Jensen-Shannon divergence and the equation of jaccard similarity coefficient is not changed, cosine similarity as well.

Result and Analysis. In order to evaluate the effectiveness of our approach, we take serval measurements into account, such as precision, recall and F-Measure. The measurements are depicted as follows,

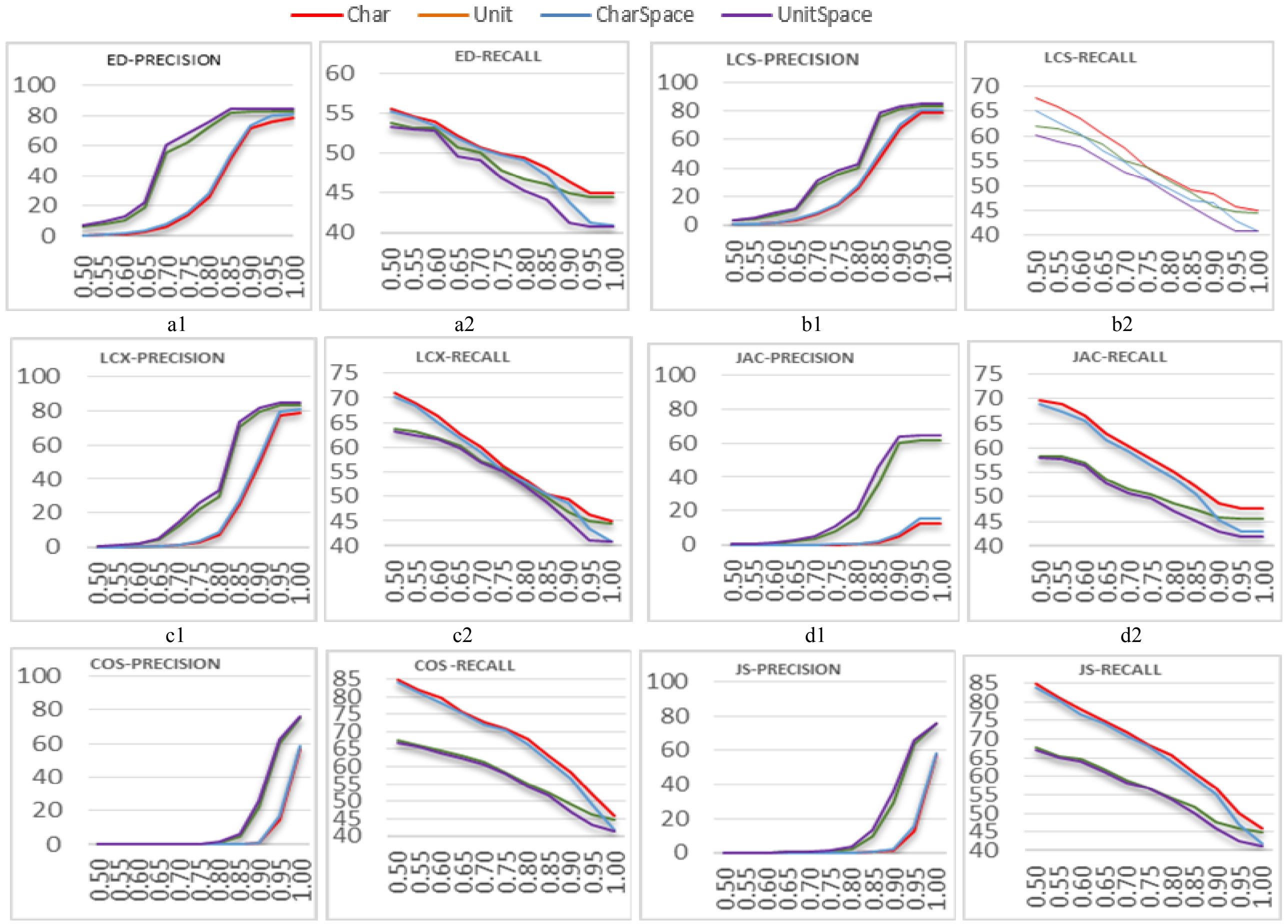
$$\mathrm{P}(Precision) = \frac{A}{A+B} \tag{9}$$

$$\mathrm{P}(Recall) = \frac{A}{A+C} \tag{10}$$

$$\mathrm{P}(F-\mathrm{Measure}) = \frac{P(Precision)*P(Recall)*2}{(P(Precision)+P(Recall))} \tag{11}$$

Where A is the number of positive instances by retrieving and B is the number of negative instances by retrieving and C is the number of positive instances which has not been retrieved.

The results are shown in Fig.2. Fig.2(a1) shows the precision of the ratio of edit distance. As can be seen from it, UnitSpace and Unit have higher precision, which proves that the precision in pinyin as a basic unit is higher than a letter as a basic unit after converting in a username containing Chinese characters. We can also see that a username containing the space achieves slightly higher precision from either a comparison between UnitSpace and Unit or CharSpace and Char. And it proves that the space is not useless and it represents some behavior habits of a user sometime. Fig.2(a2) shows the recall of the ratio of edit distance, we can see the recall decreases linearly with the increase of threshold value, and it is higher when letter is set as a basic unit. There also have some similar conclusions in other algorithms.



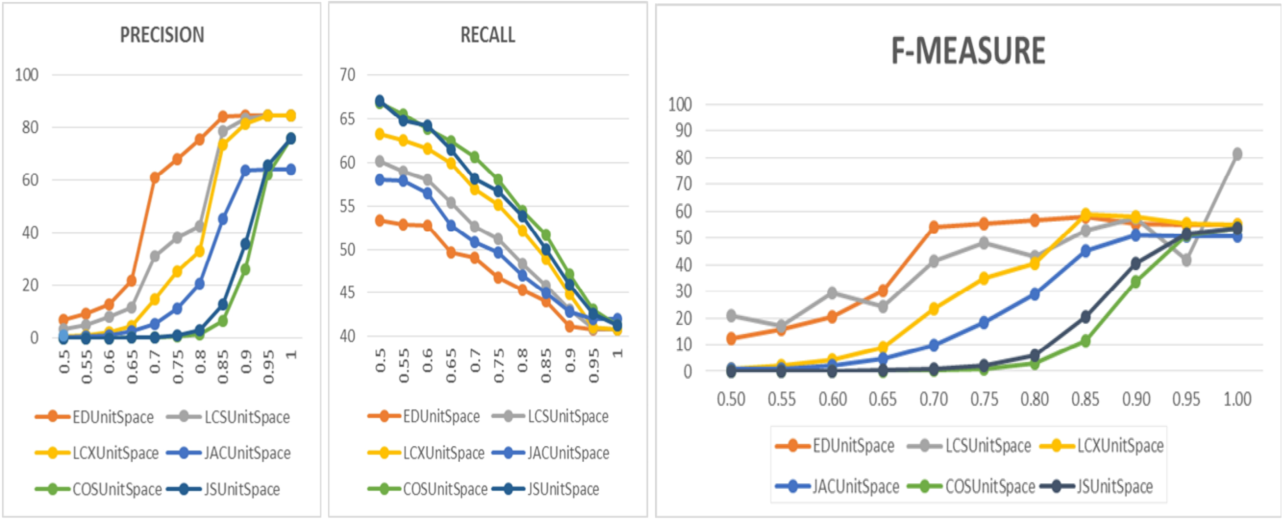
a1 a2 b1 b2

c1 c2 d1 d2

e1 e2 f1 f2

Figure 2. The precision and recall of 6 algorithms

Fig.3(a) shows the precision of all algorithms with type of UnitSpace. According to the result in it, for the precision, the ratio of edit distance, the ratio of longest common substring and the ratio of longest common subsequence outperform other algorithms. When the threshold value is in [0.85, 1.00], the ratio of edit distance has the best performance. But when the threshold value is in [0.85, 1.00], the precision of the 3 algorithms tends to be consistent. Fig.3(b) shows the recall of all algorithms. It can be seen that the recall of all algorithms decreases linearly and they tend to be consistent when the threshold value is greater than 0.9. Jensen-Shannon distance and cosine similarity have the highest recall when the threshold value is in [0.5, 0.9]. And Fig.3(c) shows the f-measure achieved by all algorithms. The ratio of edit distance has the highest f-measure when the threshold value is in [0.65, 0.85]. But the value of the ratio of longest common substring is the highest when the threshold value is in [0.5, 0.65]. When the threshold value is in [0.85, 1.00], the ratio of edit distance and the ratio of longest common subsequence tend to be consistent.



a b c

Figure 3. The comparison of 6 UnitSpace algorithms

From above figures, we have a conclusion that the space has a definite meaning and we should not delete it from the username simply. And the precision achieved by the method of pinyin as a basic unit is higher than letter as a basic unit, but the recall is opposite.

Literature References

In this section, we focus on summarizing research related to identifying individuals in social media. Nie and Zhou [3] studied the social identity linkage across two social media sites, Twitter and Sina. They proposed a fast exclusion method depending on users' opinions on specific events and the core interests of users, then users were quantified based on the core interests. At last, they used the Kullback–Leibler divergence to ensure the linkage. Siyuan Liu [4] made an effective research on social identity linkage based on user profiles and blogs. Rong [8] employed the machine learning technique to make a study on social identity linkage. Zhu [2] presented the adaptive piecewise Hash technique to match the content of MIME in email so that the authors can judge the identities of senders. Reza [12] made an analysis on social identity linkage in multiple communities. Reza [5] analyzed the features of usernames across social medias and extracted 414 features from username sequences and achieved high precision in experiment. But there still has some deficiencies: (1) The method that proposed can be used in many languages, but not including Chinese. (2) The features extracted from the paper are suitable to one candidate username and a sequence of prior usernames, which are the usernames in social media sites and belong to an individual in a real world. But there is not a detailed introduction about the situation that when connecting two users. (3)The experimental data is sufficient in that paper. But in practice, it will be better to employ big data processing tools, which that paper has not mentioned.

Then we propose the following solutions to overcome above shortcomings: (1) We study comprehensively for some problems in connecting Chinese user identities by matching the usernames. Specifically, we employ the language mapping method to handle with Chinese-English transformation, homophonic and the conversion between simplified Chinese and traditional Chinese. (2)In practice, the sequence of prior usernames introduced by Reza [5] is not easy to be obtained, and then we just talk about the situation, which only has two Chinese users from two different social network platforms(Sina, Twitter) .(3)We use the big data processing tools in experiment for the computational complexity in practice.

Summary

In social identity linkage, the usernames of social users are important which have been proved in prior research. However, how to properly connect Chinese user identities by matching the usernames remains to be explored. This paper has a discussion about social identity linkage on Chinese users based on usernames and it shows the effectiveness of various matching algorithms. But there are a lot of problems in practice if only using usernames, and even the person with same username may not be an individual in real world. There is no doubt that the username is very important in the study of social identity linkage. Chinese is different from alphabet languages, such as English. We employed the language mapping method to convert Chinese character to pinyin. But this method is limited. Such as a famous person, “Li Kaifu”, his Sina username is “Li Kai Fu” [10], and Twitter username is “Kai-Fu Lee” [11]. That situation need consider the transformation and the position of family name when converting it to pinyin and it will be a part of our future research.

Acknowledgements

The authors acknowledge the financial support by State Key Development Program of Basic Research of China (No. 2013CB329604), “863” program (No. 2012AA010401, 2012AA010402), NSFC(Grant No.61202362), project founded by China Postdoctoral Science Foundation(2013M542560). The author is grateful to the anonymous referee for a careful checking of the details and for helpful comments that improved this paper.

References

[1] Vosecky, Jan, Dan Hong, and Vincent Y. Shen. "User identification across multiple social networks." Networked Digital Technologies, 2009. NDT'09. First International Conference on. IEEE, 2009.

[2] Zhu Junxing, and Aiping Li. "An Advanced Spam Detection Technique Based on Self-adaptive Piecewise Hash Algorithm." Web Technologies and Applications. Springer International Publishing, 2014. 148-157.

[3] Nie, Yuanping, et al. "Identifying Users Based on Behavioral-Modeling across Social Media Sites." Web Technologies and Applications. Springer International Publishing, 2014. 48-55.

[4] Liu, Siyuan, et al. "Hydra: Large-scale social identity linkage via heterogeneous behavior modeling." Proceedings of the 2014 ACM SIGMOD international conference on Management of data. ACM, 2014.

[5] Zafarani, Reza, and Huan Liu. "Connecting users across social media sites: a behavioral-modeling approach." Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining. ACM, 2013.

[6] J. Yan, A. Blackwell, R. Anderson, and A. Grant. The Memorability and Security of Passwords-Some Empirical Results. U. of Cambridge Tech. Rep., 2000.

[7] Zafarani, Reza, and Huan Liu. "Connecting Corresponding Identities across Communities." ICWSM. 2009.

[8] Zheng, Rong, et al. "A framework for authorship identification of online messages: Writing‐style features and classification techniques." Journal of the American Society for Information Science and Technology 57.3 (2006): 378-393.

[9] J. Lin. Divergence Measures based on the Shannon Entropy. IEEE Transaction on Information Theory, 37(1):145–151, 1991.

[10] Information on http://weibo.com/kaifulee

[11] Information on https://twitter.com/kaifu_li

[12] Zafarani R, Liu H. Connecting Corresponding Identities across Communities.[J]. International Conference on Weblogs & Social Network, 2009.