Understanding the User Display Names across Social Networks

Yongjun Li
School of Computer
Northwestern Polytechnical University
Xi' an, Shaanxi 710072,China
lyj@nwpu.edu.cn

You Peng
School of Computer
Northwestern Polytechnical University
Xi' an, Shaanxi 710072,China
yoyopeng@mail.nwpu.edu.cn

Zhen Zhang
School of Computer
Northwestern Polytechnical University
Xi' an, Shaanxi 710072,China
zz4955@163.com

Quanqing Xu
Data Storage Institute,
A*STAR
Singapore 138632, Singapore
xu_quanqing@dsi.a-star.edu.sg

Hongzhi Yin
School of ITEE
The University of Queensland
St Lucia Campus, Brisbane
QLD 4072, Australia
db.hongzhi@gmail.com

ABSTRACT

The display names that an individual uses in various online social networks always contain some redundant information because some people tend to use the similar names across different networks to make them easier to remember or to build their online reputation. In this paper, we aim to measure the redundant information between different display names of the same individual. Based on the cross-site linking function, we first develop a specific distributed crawler to extract the display names that individuals select for different social networks, and we give an overview on the display names we extracted. Then we measure and analyze the redundant information in three ways: length similarity, character similarity and letter distribution similarity, comparing with display names of different individuals. We also analyze the evolution of redundant information over time. We find 45% of users tend to use the same display name across OSNs. Our findings also demonstrate that display names of the same individual show high similarity. The evolution analysis results show that redundant information is time-independent. Awareness of the redundant information between the display names can benefit many applications, such as user identification across social networks.

Keywords

Online social network, Display name, Redundant information, Measurement and analysis

WWW 2017 Companion, April 3–7, 2017, Perth, Australia.
ACM 978-1-4503-4914-7/17/04.
http://dx.doi.org/10.1145/3041021.3051146

1. INTRODUCTION

Nowadays, online social networks have been very popular communication tools in our daily life. According to the statics report[1], until Sept. 2016, there are 1,712 million active users on Facebook, 500 million active accounts on Instagram, 313 million active users on Twitter. However, no one social network is universal. A person usually joins various social networks for different purposes. Liu et al.[10], found that an individual joined 3.99 social networks on average.

When an individual joins a social network site, he needs to select a display name for his account. Generally, due to the limitation of human memory[18], he often has consistent behavior when selecting his display names, which brings redundant information between his different display names. For example, a user whose display name is 'Bay Area Dad' on Foursquare, has display name 'San Francisco Dad' on Twitter, which both reflect his role in family. In this paper, we mainly measure and analyze the redundant information between the display names of the same individuals.

The existing works about redundant information focus on usernames, which is unique on a single site. Vosecky et al.[16], analyzed the similarity of two usernames by the vector-based name matching algorithm. Perito et al.[14], estimated uniqueness of a username by the entropy. Iofciu et al.[7], summarized the methods used for comparing two usernames. Liu et al.[4], analyzed usernames characteristic including length, special character, numeric character etc. Zafarani et al.[18], presented a MOBIUS method to analyze the usernames that belong to the same individual.

Nevertheless, the usernames are not always alphanumeric string and sometimes even not available. In some social networks, such as Foursquare and QQ[2] , the username is assigned by the site and is a numeric string. In this situation, it has little redundancy information between the usernames. On the other hand, the user's display name is often alphanumeric string and also is obtained easily. The display names an individual selects for different OSN sites often also have

[1]https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/
[2]QQ is a very popular instant messenger in China.

abundant redundancy information, so we focus on the measurement and analysis on the display names across social networks, and make the following four main contributions.

Display Name Acquisition Framework on Cross-OSNs. Based on the cross-site linking function of Foursquare, we developed a specific distributed crawler to extract the display names individuals selected for Facebook, Twitter and Foursquare, respectively. This is the foundation of measurement and analysis on display name.

Display Name Overview on Single OSN. We first give an overview on our real datasets. Our observation indicates that 1) the letter distribution of an individual display name is similar with his real name; 2) more than 45% of users tend to use the same display name across different social networks.

Display Name Redundant Information Analysis. We measure the display names' redundant information in three ways: the length similarity, the character similarity and the letter distribution similarity. We found that 1) there is no obvious feature of an individual's display name length. 2) The character similarity between a user's display names is very high. 3) The letter distribution of display names is very similar.

Display Name Evolution over Time. We divide our real data into nine datasets based on the chronological order of registration, and demonstrate whether our measured display name attributes are relevant with the user registration time. The results shows that the display name attributes are time-independent.

The structure of this paper is as follows. In Section 2, we present the existing works. In Section 3, we describe the data acquisition process and give an overview on our datasets. We detail the measurement and analysis on the display name in Section 4 and analyze data consistency as time evolution in Section 5. The cross-name discovery and discussion is presented in Section 6 and finally we conclude this paper in Section 7.

2. RELATED WORK

Over the past few years, researchers have studied many of the properties of various online social networks. Motoyama et al.[12], measured user's profiles on Facebook and Myspace for matching individuals. Wang et al.[17], compared a series of user activities across Facebook, Twitter, and Foursquare. Chen et al.[2], presented a holistic measurement on Foursquare based on its cross-site linking function. Ottoni et al.[13], studied the user behavior across Twitter and Pinterest and found that the global patterns of use across the two sites differ significantly. These existing works give a good view for us on cross-sites analysis.

We mainly measure display names across OSNs. When comparing two names, the similarity algorithms are most commonly used. Jaro distance[8, 5, 15, 3], Jaro-Winkler[1, 11, 9, 3] and TF-IDF algorithm[11, 3] were always employed to compute the similarity of two usernames. Buccafurri et al.[1], also used Levenshtein, QGrams, Monge-Elkan and Soundex algorithm to compute the similarity of two usernames. When analyzing the characteristics of usernames, Zafarani et al.[18], utilized Longest Common Substring, edit distance, Dynamic Time Warping distance, Jensen-Shannon divergence and n-gram algorithm etc. Liu et al.[4], proposed a similarity algorithm based on the Longest Common Substring. Jain et al.[8], adopted Cosine similarity to measure the similarity of two posts. Hussain F et al.[6], also introduced Cosine similarity into medications to identify and correct the misspelled drugs' names. The display name is a short string. We make some improvements based on the above basic algorithm for calculating the similarity of two display names.

Figure 1: Two foursquare user's public profile pages

3. DATA COLLECTION AND OVERVIEW

3.1 Collection Method

Currently, some social network sites support the cross-site linking function, such as Foursquare, Google+, Pinterest. This function allows a user to link his accounts to other social network sites. We choose Foursquare to obtain the user information because of its great popularity and unique numerical user ID. If we know the ID of a user, we can access his profile page with URL *https://foursquare.com/user/ID*.

Fig.1 shows the public profile pages of two users on Foursquare. One links his Facebook account, and the other links both Facebook account and Twitter account. These account links are user-authorized and have extremely high reliability. Based on this cross-site linking function, we might obtain an individual's display names on Foursquare, Facebook and Twitter, respectively.

We obtain the display names in three steps: 1) access a user's profile page via *https://Foursquare.com/user/ID*; 2) parse the obtained profile page to get the user's Foursquare display name, as well as Twitter link and Facebook link if this user has revealed them publicly; 3) extract his corresponding display names on Facebook and Twitter by API, respectively.

We obtain the real data in two stages. In the first stage, we access the profile pages corresponding to the first 100,000 IDs. In the second stage, we extend the scale of user's IDs to1.3 million. To solve the limitation of request number from the same IP, we develop a distributed crawler, in which each sub-crawler is responsible for crawling a part of IDs. In total, 1.3 million Foursquare IDs are crawled during April and May in 2016. The sizes of the real datasets we obtained are shown in Table 1. Overall, we successfully obtained 597,822 display names on Foursquare among 1.3 million IDs. The actual obtained ratio is only about 46%. This is mainly caused by the following reasons. First, some IDs have been deactivated, so the corresponding profile pages do not exist. Second, because some users have taken strong privacy protection, we cannot access their profile pages.

Table 1: Display name collection statistics

	Planned	Obtained
Foursquare	1,300,000	597,822
Facebook	327,609	288,480
Twitter	113,951	102,315
Facebook-Twitter	-	67,826

As shown in Table 1, we actually obtain 102,315 display names on Twitter and 288,480 display names on Facebook, respectively. The number of users, who have revealed both Facebook and Twitter URLs, is 67,826. Specifically, we find that 54.80% of users have disclosed their Facebook URLs, and only 19.06% of users exposed their Twitter URLs. The former disclosed ratio is nearly three times of the latter. We take a further analysis and find it is mainly caused by their popularity. The former number of active users is 5.47 times as the latter.

3.2 Data Overview

The dataset consisting of the display names obtained from Facebook is denoted by FB. Based on the same method, we could get the dataset TW and dataset FS. We first give an overview of the display names on three datasets.

Letter Distribution What is the relationship between the display names on social network and the names in our real life? Are their letter distributions consistent? We calculate the frequency of each letter in display names, and compare the obtained display names with the commonly used names in life. These real names are collected from the data hall[3] and named as common dataset.

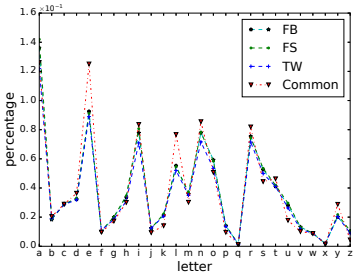


Figure 2: Letter Distribution Comparison with Real Names

Fig.2 presents the percentages of 26 letters on FB, FS, TW, and common datasets, respectively. Letters 'e' and 'l' appear more frequently in common dataset, and the percentages of other letters on four datasets are very similar. The higher percentages are followed by 'a','e','n','i','r','o','l','s','t', and 'm', accounting for about 70.4%, 71.21%, 65.63%, 75.67% of all characters on FB, FS, TW, common datasets respectively. We observe that the letter distribution of display names in three social networks is similar to in real life.

Ratio of Same Display Name We combine two display names that the same individual uses in two different sites as a pair and construct three datasets. These datasets are denoted by FB-TW, FS-TW, and FB-FS, respectively. We calculate the percentages of the same display names in three datasets, respectively. The results are illustrated in Fig.3. It should be mentioned that we ignore the letter case when we count the same display names. For example, "David Marks" and "david Marks" are defined as the same name. The percentages on FB-TW, FS-TW, and FB-FS are 47.84%, 45.84%, 63.68%, respectively. Liu et al[10] have found that 59% of individuals prefer to use the same username. For display name, we reach the similar conclusion that more than 45% individuals prefer to use the same display name across the social networks.

[3]http://www.datatang.com/data/12061

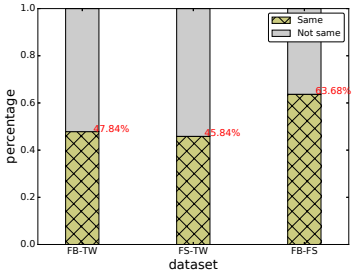


Figure 3: The Ratio of Same Name on Different Datasets

Based on an overview of the obtained data, we can see the letter distribution of the display names is similar with the real names and more than 45% of users usually use the same display name in several social network sites.

4. DATA ANALYSIS

In this section, we further measure and analyze the redundant characteristics between the different display names of the same individual from length similarity, character similarity and letter distribution similarity. We conduct our analysis on datasets FB-TW, FB-FS and FS-TW, respectively. In order to make our analysis more reliable and convincing,we compare display names of the same individual with display names of different individuals.Therefore, we construct three negative datasets randomly on the basis of positive datasets, abbreviated as negFB-TW, negFB-FS and negFS-TW, respectively. The size of negative datasets is the same as that of the corresponding positive dataset.

4.1 Length Similarity

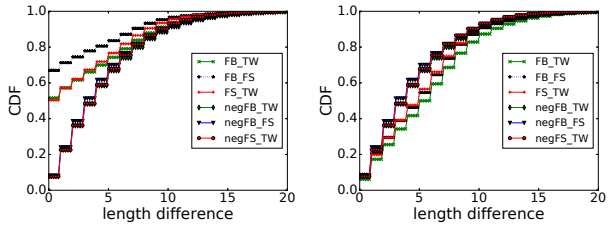
The length is one of the most basic attributes researchers have used. On dataset FB, the maximum length of the display name is 70, however, the length of the display name of the same individual selected for Foursquare is 40. One straightforward question is how much difference in length between different display names of the same user is. Therefore, we conduct a detailed measurement and analysis on the length difference of the display names.

We assume that $name_1$ and $name_2$ are two display names of an individual. The length difference of $name_1$ and $name_2$ is expressed by Eq. (1). The results are shown in Fig.4 (a).

$$\triangle Len_{name} = abs(len(name_1) - len(name_2)) \quad (1)$$

From Fig.4(a), we can see that only less than 0.5% of the length difference is larger than 20. More than 50% of the instances have length difference of 0 on positive datasets, while less than 10% on the negative datasets. This is mainly due to the fact that more than 45% individuals use the same display names on different social networks. For further observation, we remove these positive instances which two display names are completely same, and repeat the above measurement. The results are shown in Fig.4(b).

The CDF curves based on positive and negative datasets are very close, and the curves of negative datasets are just slightly higher than positive datasets. That is to say, the length difference of the positive instance has no significant difference with the negative instance. Besides, the curves on FB-TW, FB-FS, FS-TW are almost completely coincide.



(a) with name-pairs (b) without same name-pairs

Figure 4: The length difference distribution of the display names

This means the length difference have no significant correlation with the social networks.

4.2 Character Similarity

Two display names from different social networks are two special strings. Each string is composed of 1-3 words. We can use string similarity algorithm, as well as the characteristics of names, to measure the character similarity.

Length of LCS/Short Length The longest common substring problem is a good metric to measure the similarity of two different strings. When using the longest common substring, we also consider the affection of the username's length. We define this metric as the ratio of the length of the longest common string to the minimum length between two strings. Its value ranges from 0 to 1. The greater the value is, the more similar two display names are. Assume two display names are $name_1$ and $name_2$, respectively. This metric is expressed by Eq. (2).

$$Sim_{lcs} = \frac{len(lcs(name_1, name_2))}{min(len(name_1), len(name_2))} \quad (2)$$

We divide the value space of Sim_{lcs} into 11 slots. Based on Eq.(2), we calculate Sim_{lcs} of every pair of display names on all positive and negative datasets. The distributions of Sim_{lcs} are illustrated in Fig.5. Generally, the Sim_{lcs} values of the negative instances are concentrated at range [0, 0.2], and its proportion is larger than 91%. On the other side, the values of the positive instances are distributed in each slot, and the proportions of values located in [0.5, 1] are more than 64%, 64%, 74% on three datasets, respectively. By contrast, there are only less than 2% negative instances whose Sim_{lcs} values are larger than 0.5. After two same display names are removed from the datasets, in FB-FS and FS-TW, there are over 26% instances whose Sim_{lcs} values are 1.0 and over 20% instances on the FB-FS. However, on the negative datasets, there is no instance whose Sim_{lcs} value is equal to 1.0. Thus, it is clear that the users have their own fixed naming habit, rather than completely random selecting.

No. of common words /No. of short name words The display name can be divided into first name, last name or even middle name. Assume two display names are $name_1$ and $name_2$, respectively. Each name contains several words. We consider the number of the common words between two names, and is expressed by Eq.(3).

$$Sim_{word} = \frac{commonword(name_1, name_2)}{min(word(name_1), word(name_2))} \quad (3)$$

where $commonword(name_1, name_2)$ counts the number of the common words between $name_1$, and $name_2$; *word(name)* count the number of words contained in *name*.

We illustrate the results in Fig.6. The values of all negative instances are 0. However, nearly 50% of positive instances also have no common word between two display names, but it still has 30%, 27%, and 42% of positive instances with value 1.0 on three datasets, respectively. Besides, there are more than 20% of positive instances with value 0.5. The cases arise mainly because some individuals omit his first name or last name.

Edit Distance/Longest Length: The edit distance is a commonly used metric to evaluate the difference of two strings. Also, the edit distance of two names relates to the name length. The larger the names lengths are, the larger their edit distance may be. Therefore, we introduce the name length to this metric and express it by Eq.(4). The smaller the value is, the smaller the difference between two names is, that is, the larger the similarity is.

$$Sim_{edit} = \frac{edit(name_1, name_2)}{max(len(name_1), len(name_2))} \quad (4)$$

The results are shown in Fig.7. As we expected, the values of all negative instances are larger than 0.5. Conversely, the values of most positive instances is smaller than 0.5. The percentages of these instances are 54.36%, 53.90%, and 72.68% on three datasets, respectively. That is, if the edit distance of two display names is less than half of the longest name length, these two display names belong to the same individual with high probability.

Max of Best Match: Normally, an individual's display name consists of <first name, [middle name], [last name], [title]>. While not everyone writes all parts, some people omit the middle name, others omit last name, or even reverse the first name and last name. Thus, if we just compare the name as a whole, it will neglect the name's identical part. Therefore, we consider the max of best part match based on the longest common substring.

Suppose s_1 and s_2 are two strings. The similarity of s_1 and s_2 is expressed by Eq.(5).

$$Sim_{str} = \frac{len(lcs(s_1, s_2))}{(len(s_1) + len(s_2))/2} \quad (5)$$

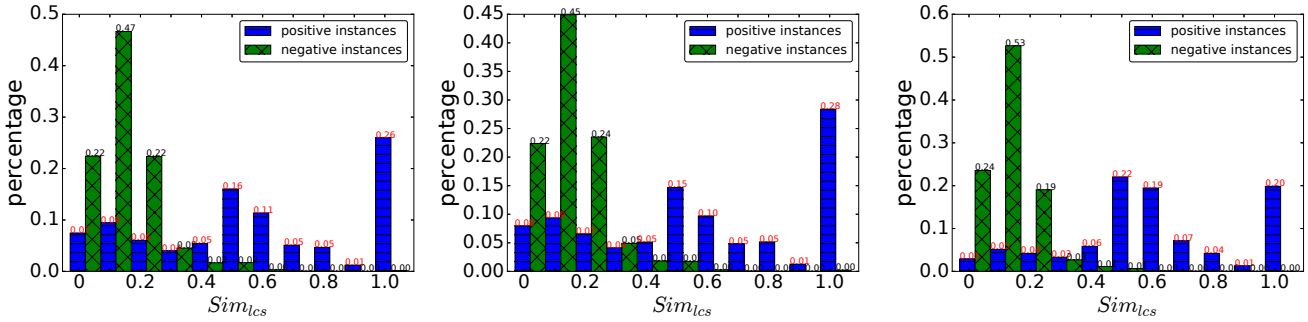
Suppose $name_1$ and $name_2$ are two display names. The detailed implementation steps of max of best match of $name_1$ and $name_2$ are shown as follows.

Step 1: Segment the two names into words, respectively, and get two name arrays Arr_1 and Arr_2;

Step 2: Calculate similarity of each word in Arr_1 with word in Arr_2 based on Eq.(5), and get a similarity matrix **A**;

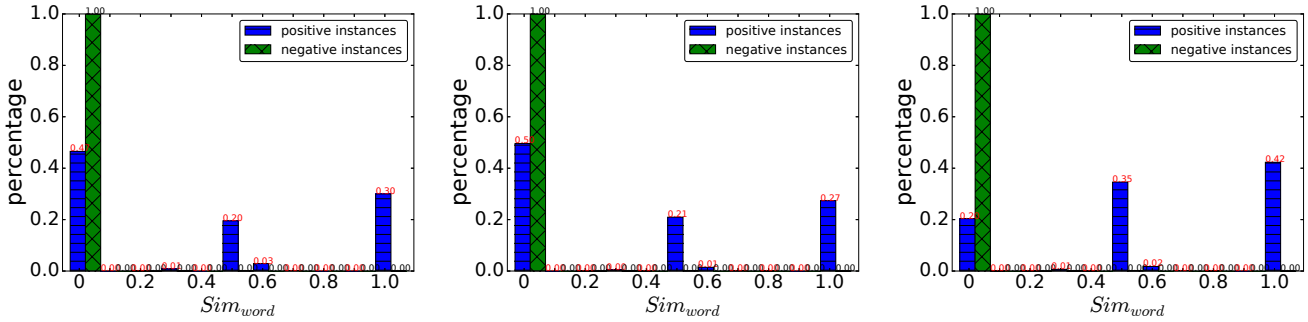
Step 3: Find the largest value in matrix **A**. The maximum of all values is the max of best match.

Fig.8 shows our measurement results on max of best match. We can see that most of the metric values on the negative instances are below 0.5 while on positive instances the metric values are larger than 0.5. The metric values of 53.43% of the positive instances on FB-TW are 1.0, and 50.37% of the positive instances on FS-TW have value 1.0. There are also about 20% of positive instances whose values are in [0.5, 0.9] on FB-TW and FS-TW. These users make some changes to their first names, or last names, or middle names. Compared with the name length, these changes are small. On the other



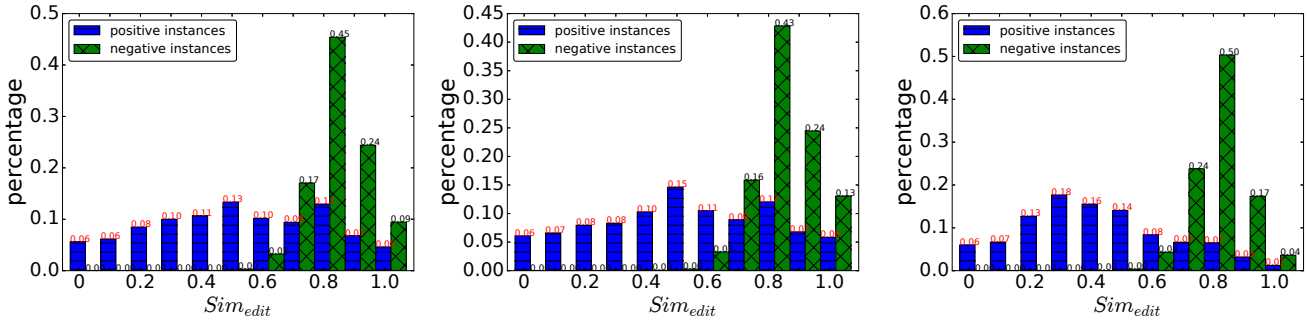
(a) FB-TW (b) FS-TW (c) FB-FS

Figure 5: Distribution of Sim_{lcs} on three Datasets



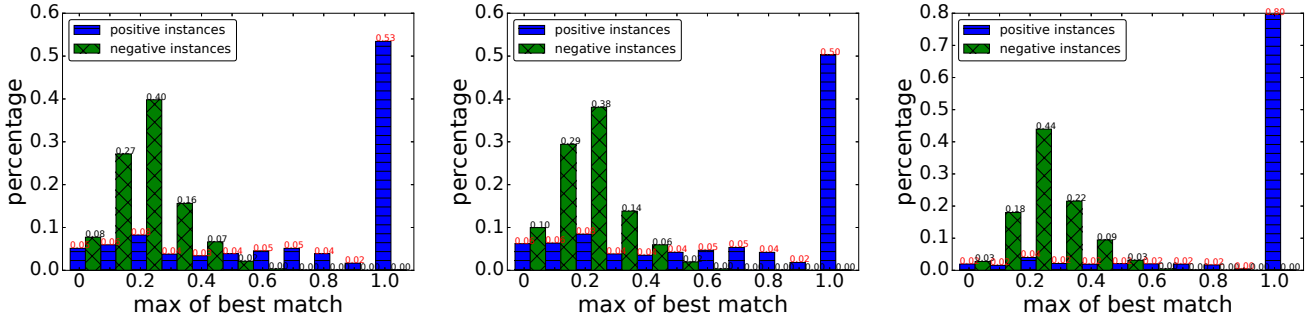
(a) FB-TW (b) FS-TW (c) FB-FS

Figure 6: Distribution of Sim_{word} on three Datasets



(a) FB-TW (b) FS-TW (c) FB-FS

Figure 7: Distribution of Sim_{edit} on three Datasets



(a) FB-TW (b) FS-TW (c) FB-FS

Figure 8: Distribution of max of best match on three Datasets

side, the percentage of the positive instances with values 1.0 is as high as 80% on FB-FS, which is higher than other two datasets. This mainly because the users always select the display names similar to their real names on Facebook and Foursquare.

4.3 Letter Distribution Similarity

The letter distribution presents the occurrence probability of each letter in a display name. For two similar display names, their letter distribution is also similar. For example, name "gate man" and name "man gate" have same letter distribution. For simplicity, here we only consider twenty-six English letters.

Jensen-Shannon Similarity Jensen-Shannon distance is used to calculate the difference between two probability distributions. We use JS distance to measure letter distribution difference between two display names. The smaller the distance is, the greater the similarity between two distributions is. Assume P and Q are the letter distributions of display name $name_1$ and $name_2$, respectively. The JS similarity of name $name_1$ and $name_2$ is expressed by Eq.(6).

$$\begin{aligned} Sim_{js} &= 1 - \frac{1}{2}(KL(P||M) + KL(Q||M)) \\ KL(P||Q) &= \sum_{i=1}^{P} P_i \cdot log\frac{P_i}{Q_i}, M = \frac{1}{2}(P+Q) \end{aligned} \tag{6}$$

where p_i is the occurrence probability of the i^{th} character. To avoid the situation that the logarithm does not make sense, we use a very low value e to smooth for the letters whose probability is zero. In this paper, we set the e to $2.2204460492503131e^{-16}$. The measurement results are illustrated in Fig.9.

For the negative instances, the Sim_{js} values are concentrated on the left and less than 0.7, while for the positive instances, their Sim_{js} values focus on the right, and show an increasing trend in [0.1,0.8]. On FB-TW and FS-TW, the values of only 5%, 6% of the instances are less than 0.1, which the distributions of letters are almost completely different. This is mainly due to the fact that the display names a user selected for different social networks are in different languages. From Fig.9, we also easily find that the percentage of the positive instances with values larger than 0.8 is about 45%, while the percentage of the negative instances is less than 2%.

Jaccard Similarity Jaccard similarity is used to compare the similarity of two sets. We consider the letters in a display name as a set and calculate the Jaccard similarity by Eq.(7). Fig.10 illustrates the results of Jaccard Similarity.

$$Sim_{jac} = \frac{len(set(name_1) \bigcap set(name_2))}{len(set(name_1) \bigcup set(name_2))} \tag{7}$$

where $set(name)$ is the set of letters in the name.

For the negative instances, the Jaccard similarity value is very small. The values of most of the negative instances are less than 0.5. For the positive datasets, the value distribution is more uniform. It should be noticed that we remove the positive instances with two same display names, but the values of the positive instances are still much larger than the values of the negative instances on average.

5. EVOLUTION ANALYSIS

In the above analysis, we only consider a single snapshot of the social network, neglecting their evolution over time. Is the redundant information between two display names consistent over time? In this subsection, we make the evolution analysis on the character similarity and letter distributions.

In Foursquare, the larger the user ID is, the later the registration time of this user account is. To obtain multiple snapshot of Foursquare, we divide the total ID into nine chunks, and crawl the first 150,000 IDs on each chunk. After repeat the data collection described in section 2.1, we obtain 9 chunks across Foursquare and Twitter. For the sake of convenience, the nine datasets across Foursquare and Twitter are denoted by $FS-TW_i$ (i=0,1,...,8). Similarly, we have datasets $FS-FB_i$ (i=0,1,...,8).

Fig.11 shows the evolution analysis results on $FS-TW_i$ (i=0,1,...,8). We find the letter similarity and character distribution described in section 4 are consistent on most datasets. The attributes on datasets $FS-FB_i$ (i=0,1,...,8) also remain unchanged over time. Because of lack of space, we do not show their distribution figures and length analyses in detail .

6. DISCOVERY

Through above analysis, we conclude:

(1) More than 45% of users tend to use the same display name in different OSNs. This is mainly because the users have limited memory and the needs of maintaining their personal image and reputation on different social network sites.

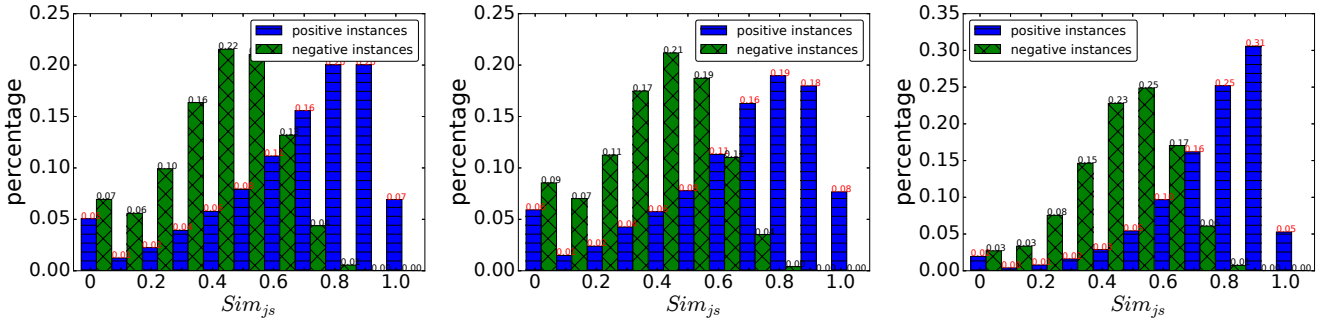
(2) For the positive instances, the letter similarity is striking, although two names are not exactly same. Specifically as follows:

- For more than 64% of the positive instances, the length of the longest common substring is more than half of the shorter name length. Moreover, there are 20% of the users whose one name is fully contained in the other name;
- 27% or more of users have the same surname or last name;
- There are 53% of positive instances whose edit distance is less than half of his longer name length;
- As for the best match of the positive instances, the values of more than 50% of users are 1.0.

A user usually selects different names in different OSNs for the purpose of privacy. Actually, most individuals just change part of their real names, and retain some of the basic information. This information tends to make the display names of a user having high character similarity.

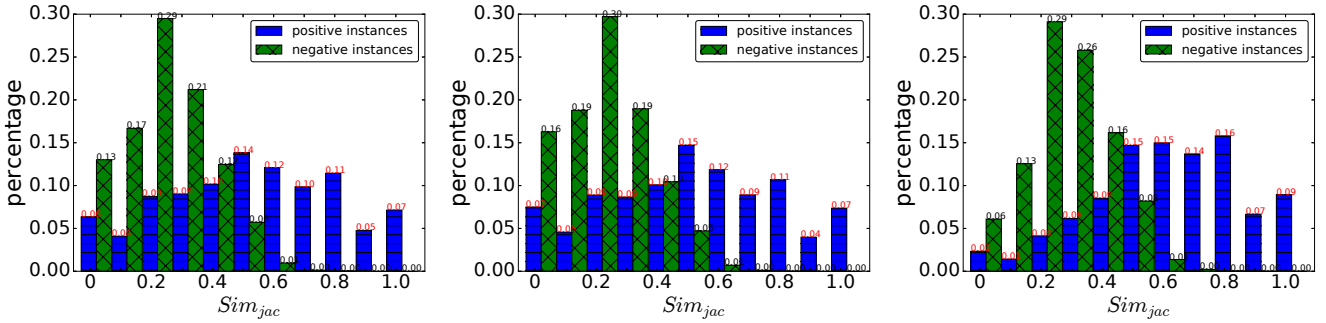
(3) The letter distributions of the positive instance are very similar.

- The Jensen-Shannon similarity of more than 45% of positive instances are larger than 0.8;
- The Jaccard similarities of more than 47% of positive instances are over 0.5. On the contrary, the corresponding percentage of the negative instances is only 2%.



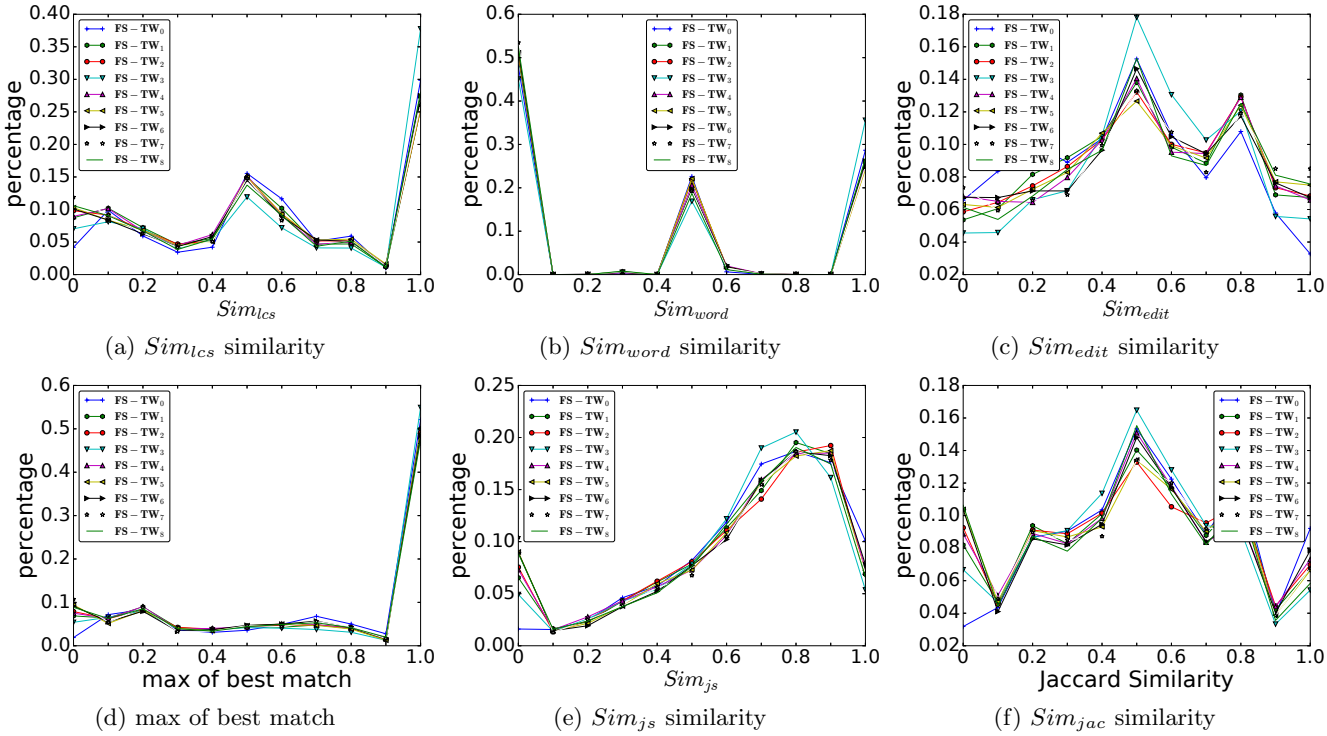
(a) FB-TW (b) FS-TW (c) FB-FS

Figure 9: Distribution of Sim_{js} on three Datasets



(a) FB-TW (b) FS-TW (c) FB-FS

Figure 10: Distribution of Sim_{jac} on three Datasets



(a) Sim_{lcs} similarity (b) Sim_{word} similarity (c) Sim_{edit} similarity

(d) max of best match (e) Sim_{js} similarity (f) Sim_{jac} similarity

Figure 11: Evolution analysis on character distribution on FS-TW

The alphabet distribution reflects the user preference for specific letter. Some letter can also reflect a user's country or region to a certain extent. Therefore, the closer the letter distribution of two display names is, the more likely the two names belong to the same user.

(4) The evolutionary analysis results show that the above attributes remain unchanged over time.

(5) The similarity of two display names from Facebook and Foursquare is generally more striking. This is mainly due to the user tend to choose his display name closer to his real name on these two social networks.

7. CONCLUSION

A display name is a name that an individual chooses shown to other avatars on an OSN site. By comparing the display names from the same users and the different users, we know that the character similarity and the letter distribution of the positive instances are very high. The results of our measurements demonstrate that the same individual on different OSNs tends to use the same display names or similar display names. The presented attributes are very helpful for identifying whether accounts belong to the same individual or not based on their display names.

8. ACKNOWLEDGEMENTS

This research was supported in part by Shaanxi Provincial Natural Science Foundation Research, China under grant No. 2014JM2-6104.

9. REFERENCES

[1] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino. Discovering links among social networks. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, pages 467–482. Springer, 2012.

[2] Y. Chen, C. Zhuang, Q. Cao, and P. Hui. Understanding cross-site linking in online social networks. In *Proceedings of the 8th Workshop on Social Network Mining and Analysis*, page 6. ACM, 2014.

[3] W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string metrics for matching names and records. In *Kdd workshop on data cleaning and object consolidation*, volume 3, pages 73–78, 2003.

[4] D.Liu, Q. Wu, W. Han, and B. Zhou. User identification across multiple websites based on username features. *Chinese Journal of Computers*, 38(10):2028–2040, 2015.

[5] O. Goga, H. Lei, S. H. K. Parthasarathi, G. Friedland, R. Sommer, and R. Teixeira. Exploiting innocuous activity for correlating users across sites. In *Proceedings of the 22nd international conference on World Wide Web*, pages 447–458. ACM, 2013.

[6] F. Hussain and U. Qamar. Identification and correction of misspelled drugs' names in electronic medical records (emr). *Science and Technology Publications*, 2:333–338, 2016.

[7] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media*, pages 522–525, 2011.

[8] P. Jain, P. Kumaraguru, and A. Joshi. @ i seek'fb. me': identifying users across multiple online social networks. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1259–1268. ACM, 2013.

[9] K. Kim, M. Khabsa, and C. L. Giles. Random forest dbscan for uspto inventor name disambiguation. *arXiv preprint arXiv:1602.01792*, 2016.

[10] J. Liu, F. Zhang, X. Song, Y.-I. Song, C.-Y. Lin, and H.-W. Hon. What's in a name?: an unsupervised approach to link users across communities. In *Proceedings of the sixth ACM international conference on Web search and data mining*, pages 495–504. ACM, 2013.

[11] A. Malhotra, L. Totti, W. Meira Jr, P. Kumaraguru, and V. Almeida. Studying user footprints in different online social networks. In *Proceedings of the 2012 International Conference on Advances in Social Networks Analysis and Mining (ASONAM 2012)*, pages 1065–1070. IEEE Computer Society, 2012.

[12] M. Motoyama and G. Varghese. I seek you: searching and matching individuals in social networks. In *Proceedings of the eleventh international workshop on Web information and data management*, pages 67–75. ACM, 2009.

[13] R. Ottoni, D. B. Las Casas, J. P. Pesce, W. Meira Jr, C. Wilson, A. Mislove, and V. Almeida. Of pins and tweets: Investigating how users behave across image-and text-based social networks. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media*, 2014.

[14] D. Perito, C. Castelluccia, M. A. Kaafar, and P. Manils. How unique and traceable are usernames? In *International Symposium on Privacy Enhancing Technologies Symposium*, pages 1–17. Springer, 2011.

[15] E. Raad, R. Chbeir, and A. Dipanda. User profile matching in social networks. In *Network-Based Information Systems (NBiS), 2010 13th International Conference on*, pages 297–304. IEEE, 2010.

[16] J. Vosecky, D. Hong, and V. Y. Shen. User identification across multiple social networks. In *2009 First International Conference on Networked Digital Technologies*, pages 360–365. IEEE, 2009.

[17] P. Wang, W. He, and J. Zhao. A tale of three social networks: User activity comparisons across facebook, twitter, and foursquare. *IEEE Internet Computing*, 18(2):10–15, 2014.

[18] R. Zafarani and H. Liu. 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*, pages 41–49. ACM, 2013.