Towards Multi-account Users Detection and Connection Based on Telecom Data

Hu Zhongshun [1, a], Chen fei [2,b] , Xie Chenhao[2,b], and Xiao Yanghua [2,b]

[1] Shanghai Ideal Information Industry(Group)co.,ltd

No.1835, South Pudong Road, Shanghai, China

[2]Fudan University, No. 825 Zhangheng Road, Shanghai, China

[a]huzs@ideal.sh.cn, [b]redreamality@gmail.com, [c]shawyanghua@gmail.com

Keywords: DPI, connecting users, similarity, user profile, behavioral similarity, co-occurrence

Abstract. Internet users tend to have a number of different types of Internet accounts, and yet there does not exist a perfect Internet account association approach. The telecom network operators use DPI technology can detect the part of user's Internet records, which makes it possible to realize the connection of the Internet account on the operator side. This paper illustrates the significance of extraction, connection of Internet accounts from the DPI data. We combine the edit distance and Jaccard coefficient two string similarity algorithms to propose a string similarity method which is suitable for calculating the similarity of the account information and use classification algorithm to analyze the similarity of the account from the user's behavior, and find the correlation between the accounts. On the basis of the account associated, we provide more detailed portrait of users. Finally, we show that the proposed method can effectively extract the accounts from DPI data and associate accounts experiment on China Telecom data.

Introduction

With the rapid development of the Internet, nowadays a lot of daily requirements (e.g., shopping, social networking and entertainment) are satisfied through the network. Most users have one or more accounts in different web application platforms. Telecom operators are able to collect web data of their users by DPI technology [1], which contain different accounts of many users. No other company can own such a big amount of data. In other words, DPI data contains the date from many big Internet corporations. When the complementary information on each site is integrated together, we can present a global image of users' interests and provide more user information to help improve online services such as network information verification.

Reza Zafarani et al.[2] classify human behavior into three categories according to human limitations, external factors and internal factors. They select user behavior and establish MOBIUS (Modeling Behavior for Identifying Users across Sites) to judge the correlation of users from different social networking sites [2]. Pasquale De Meo et al. deeply analyze the user behavior on different social sharing platforms. They analyze the trend of user behavior, user portraits based on tags and the semantic information of user portraits[3]. Papers [3] [4] carry out some correlation analysis between accounts from different social networking sites using a wealth of information, but the research is limited to social networking sites. Focus on the problem of the calculation of the similarity of Chinese characters.

This paper proposes the method that decrypting the plain code from DPI data, marking field meaning, and then extracting the Internet website account information. We propose the method of calculating the string similarity and calculating correlation between isomerism accounts under the Sino-British mixed circumstance, to obtain multi-user accounts and match the accounts with the users. We use multi-user accounts to get more information and more detailed description of real users.

Problem Statement

Account association is to determine whether any two accounts belong to the same user. If the accounts belong to the same person then a direct correlation is established. The multi-account association has the following problems:

Account heterogeneity. Social networking sites are rich in information, similar in structure, and many fields of the accounts are the same. Exuberant account information indicates correlation between accounts. Account information structure differs a lot between different types of sites, and the accounts have less common properties, which provide abundant features from different aspects such as the user behavior, timestamp, browser information. The account heterogeneity demands the dots to be connected from every aspects mentioned above.

Diversity of characters. Different types of accounts have different limitations in different fields. Some fields are pure English, others are pure Chinese, others are English and Chinese mixed, and others (such as QQ nickname) even allow special characters. The unique problem set requires much processing techniques.

Account Matching Strategy

Ordered by reliability from strong to weak, we match accounts via the following three indications.

Matching against Strong Indications. This strategy compares strong association information between accounts, such as QQ number, email, phone number. Exact match for the strong association of information between any two accounts can be considered accurate to associate the two accounts, and similar strong match-related information is meaningless, like QQ number, phone number and no practical significance.

Strong association information may be considered an exact match between the accounts, on the one hand as an output of the account associated with the module, on the other hand as positive and negative samples for an account associating other similar algorithms, to train the model, and to evaluate algorithm.

Matching against content similarity. This strategy compares the string similarity of personal information between accounts. We propose a method that can handle many types of string, so the original string calculated can contain Chinese, English and other characters. Combining edit distance and Jaccard coefficient, we propose an improved similarity, which adapts to the computation of multiple character types in heterogeneous accounts, in order to determine the similarity between accounts extracted from the same broadband.

Algorithm A-1 Account information string similarity algorithm

Input：two normalized string s1, s2
Output：whether the two strings are similar

```
if s1,s2 is identical:
    return true
if one of them is the substring of the other:
    return true
if s1,s2 are both in Chinese:
    if EDsim(s1, s2) > t1: return true
    if Jaccard(s1, s2) > t2: return true
Cast s1、s2 to phoneticize
if EDsim(s1, s2)>t3: return true
else:
    return false
if s1, s2 are both in English:
if EDsim(s1, s2) > t4: return true
else:
        return false
#Chinese mix up with English
if EDsim(s1, s2) > t5
        return true
else:
        cast s1,s2 to phoneticize
if EDsim(s1,s2) >t6
        return true
return false
```

Algorithm A-1 is the pseudocode that analyzes similarity algorithm normalization, where s1, s2 represent two strings, specifically refers to two fields of account information used to compare.t1-t6 represent thresholds of different situations and different similarity algorithm. The edit distance similarity is defined by Eq.1:

$$\text{EDsim} = \frac{EditDistance}{\min(s1,s2)}, \quad (1)$$

where s1, s2 are strings to be compared and min(s1, s2) is the smaller length of the two. Jaccard coefficient can be calculated as Eq.2:

$$Jaccard = \frac{S1 \cap S2}{S1 \cup S2}, \quad (2)$$

With the string similarity algorithm above, we can determine whether the two accounts belong to the same user by computing the string similarity of the personalized information on the different accounts under the same ADSL account, so the association between the two accounts can be established.

Matching user behavior similarity. This strategy compares behavior similarity between accounts. The following are the three types of behavior-based features:

Co-occurrence times ratio: As for a person who has a different account, total co-occurrence frequency will be greater than those between different people's accounts. Co-occurrence is counted if distance of two accounts log timestamps is less than a given time window Δt, co-occurrence frequency is recorded as coTimes. The co-occurrence times ratio is defined as the Eq.3 :

$$\text{coTimesRate} = \frac{coTimes}{min(n1,n2)}, \quad (3)$$

where coTimesRate representatives the ratio of the co-occurrence counts, coTimes refers to the number of co-occurrence, and n1, n2 denote the number of records for two accounts respectively, min (n1, n2) represents choosing the smaller one as the denominator.

Same UserAgent proportion: The one frequently used Internet often only a few devices, such as a computer, a cell phone. First, we traverses all the records of two accounts, and count the number of records whose UserAgent are identical. Then, we use the Eq.4 to calculate UserAgent ratio:

$$\text{sameUARate} = \frac{sameUA}{min(n1,n2)}, \quad (4)$$

where, sameUA keeps the records of two accounts with same UserAgent, and n1, n2 represents the number of records of the two accounts respectively, where the smaller is as the denominator .

Ratio of co-occurrence with same UserAgent: We consider if the frequency of logs from two accounts appear in same time windows with same UserAgent is high enough, the two accounts can belong to one user, co-occurrence of two accounts with same UserAgent is calculated by Eq.5:

$$\text{Co\_UARate} = \frac{Co\_UA}{min(n1,n2)}, \quad (5)$$

where, Co_UA is the number of co-occurrence in two accounts logs with same UserAgent, and n1, n2 represent the number of records of two accounts respectively, where the smaller one is as the denominator .

Experimental Results and Evaluation

Making pairwise comparison with the different accounts under the ADSL broadband, we determine that both accounts are the same people's account when the similarity of personalized information is greater than a threshold or share exactly the same intensity information such as QQ number, phone number, registered mailbox name. We report the results in Table 1.

Table 1 Account Information Matches

Domain	Amount
AD possesses Weibo and QQ at the Same Time	1306
QQ	8973
Weibo	2156
Account Pairs with Same Registered Information	246
Account Pairs we found using Similarity Algorithm	1805

Among them, we use QQ-Weibo account pairs with the same registration information as a positive sample to evaluate the similarity algorithm alone. We artificial evaluated 246 account pairs with same information and found 72 account pairs are similar while the rest account pairs dissimilar. The algorithm determined that 68 accounts are similar. The algorithm can determine *94.4%* of the artificial determination similar account pairs, indicating that the similarity algorithm is close to the artificial.

Conclusions

We deep mining and analyze Telecom DPI big data, parse plaintext Internet account information from the combination of Telecom DPI data and Internet data, obtain a wealth of account information. For domestic Internet account information which is in Chinese, English and special characters mixed actual locale, we propose a method for calculating string similarity of account information. Using Telecom DPI big data, we propose a method to associate accounts from different types of sites, extract account information from Telecom DPI data, narrow account relevancy range, use account information and account behavior to extract common features of different types of accounts, and finally implement associations between different types of accounts. The telecommunication data mining analysis system derived from this paper has been transplanted to the telecommunication cluster and the server. The system analyzes from DPI data while a small number of accounts remain unresolved, the account logs recognition rate is still room for improvement. For the account association accuracy and coverage has not reached perfection, in terms of accuracy and coverage there is further room for improvement.

Acknowledgements

This paper was supported by the National NSFC(No.61472085, 61171132, 61033010), by National Key Basic Research Program of China under No.2015CB358800, by Shanghai Municipal Science and Technology Commission foundation key project under No.15JC1400900.

References

[1] Dharmapurikar, S., Krishnamurthy, P., Sproull, T., & Lockwood, J. (2003, August). Deep packet inspection using parallel bloom filters. *In High performance interconnects, 2003. proceedings. 11th symposium on (pp. 44-51)*. IEEE.

[2] Zafarani, R., & Liu, H. (2013, August). Connecting users across social media sites: a behavioral-modeling approach. In P*roceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 41-49). ACM.

[3] Meo, P. D., Ferrara, E., Abel, F., Aroyo, L., & Houben, G. J. (2013). Analyzing user behavior across social sharing environments. ACM *Transactions on Intelligent Systems and Technology (TIST)*, 5(1), 14.

[4] Pennacchiotti, M., & Popescu, A. M. (2011). A Machine Learning Approach to Twitter User Classification. *ICWSM*, 11(1), 281-288.

[5] Jingting Wang. A study on Chinese character clustering and similarity. Modern Technology of Library and Information 2011, 27(2): 48-53 (in Chinese)