Future Generation Computer Systems 83 (2018) 104–115

Contents lists available at ScienceDirect

Future Generation Computer Systems

journal homepage: www.elsevier.com/locate/fgcs

Matching user accounts based on user generated content across social networks

Yongjun Li [a,*], Zhen Zhang [a], You Peng [a], Hongzhi Yin [b], Quanqing Xu [c]

[a] School of Computer, Northwestern Polytechnical University, Xi'an, Shaanxi 710072, China
[b] School of ITEE, The University of Queensland, Brisbane, QLD 4072, Australia
[c] Data Storage Institute, A*STAR, Singapore 138632, Singapore

HIGHLIGHTS

- A User Identification Model solely based on User Generated Content across social sites is presented.
- A solution to UGC-based user identification is proposed.
- The importance of spatial, temporal and content similarity for user identification is identified.
- The user identification experiments based on three ground truth datasets are conducted.

ARTICLE INFO

Article history:
Received 1 November 2017
Received in revised form 13 December 2017
Accepted 20 January 2018
Available online 31 January 2018

MSC:
J.4
H.2.8
H.3.3

Keywords:
User identification
User generated content
Online behavior analysis
Social network
Machine learning
Algorithms
Experimentation

ABSTRACT

Matching user accounts can help us build better users' profiles and benefit many applications. It has attracted much attention from both industry and academia. Most of existing works are mainly based on rich user profile attributes. However, in many cases, user profile attributes are unavailable, incomplete or unreliable, either due to the privacy settings or just because users decline to share their information. This makes the existing schemes quite fragile. Users often share their activities on different social networks. This provides an opportunity to overcome the above problem. We aim to address the problem of user identification based on User Generated Content (UGC). We first formulate the problem of user identification based on UGCs and then propose a UGC-based user identification model. A supervised machine learning based solution is presented. It has three steps: firstly, we propose several algorithms to measure the spatial similarity, temporal similarity and content similarity of two UGCs; secondly, we extract the spatial, temporal and content features to exploit these similarities; afterwards, we employ the machine learning method to match user accounts, and conduct the experiments on three ground truth datasets. The results show that the proposed method has given excellent performance with F1 values reaching 89.79%, 86.78% and 86.24% on three ground truth datasets, respectively. This work presents the possibility of matching user accounts with high accessible online data.

1. Introduction

In the last decade, many popular social network sites have emerged and the number of monthly active users has also grown quickly. As of April 2017, Twitter has more than 319 million monthly active users, and Facebook has 1968 million monthly active users. Sina Microblog,[1] a popular Twitter-style Chinese microblog service, also has more than 313 million monthly active users [1]. These social sites have changed the way we interact with each other, and make it simpler to stay connected with friends.

Normally, people tend to use several Online Social Networks (OSNs) simultaneously. For instance, an individual uses Facebook to keep in touch with his friends, uses Twitter to post news, and uses Foursquare[2] for location-based social activities. As we expect, his social activities and connections are scattered on several networks, so his online data on a single site is often incomplete. If we accurately integrate these sites, we can build his better and more complete personal information to improve online services, such as community discovery, recommendation.

To integrate OSNs, it is essential to match user accounts across sites. Matching user accounts is also called user identification, and anchor linking [2]. There are some existing works which discuss

* Corresponding author.
E-mail address: lyj@nwpu.edu.cn (Y. Li).

[1] http://weibo.com.

[2] http://foursquare.com.

https://doi.org/10.1016/j.future.2018.01.041

possible solutions to this problem (see Section 2 for detail). Many existing works addressed this problem based on the rich user profile attributes [3,4], including name, birthday, hometown or work location, gender, education, profile photo. Due to personal privacy settings, it is costly or difficult to obtain the above attributes. On the other hand, these attributes are also easily faked for special purposes. The above limitation makes these existing schemes quite fragile [2]. Some researchers leveraged the friendship network to identify users [2,5,6]. Taking into account personal privacy, most of users usually make part of the friendship network public. Even if we can obtain the user friendship network, these connections are also sparse. These existing user identification methods based on friendship network are plagued by the above limitations [2]. Some researchers also employed User Generated Content(UGC) to identify users based on posting time, location and writing style. However, in existing works, the UGCs often are used with profiles or friendship network together to identify users, so these solutions face similar problems as described above.

The UGCs posted on different sites by one user usually contain rich information redundancies. Meanwhile, users often make some of their UGCs public and these public UGCs are easily obtained. Intuitively, we can identify users merely based on UGCs, and the UGC-based user identification could break through the above limitations. With its value and significance, the UGC-based method is surely very challenging. The first challenge is writing style, which is usually difficult to be extracted from short UGCs. The number of UGCs user posting publicly on different sites are seriously imbalance. In this work, we focus on UGCs across OSNs and present a novel framework to tackle user identification. As we know, this is one of the few works on user identification solely based on UGCs. This method could be applied jointly with other feature-based algorithms for better identification performance. This work makes four contributions as following.

(1) A UGC-based User Identification Model (U-UIM) across OSNs is presented. In our real life, an individual usually posts the same activities on different OSNs. The more similarities in two user's UGCs, the higher the probability that they belong to the same offline individual. We formulate the UGC-based user identification problem as the similarity measurement on two user's UGCs, and further subdivide into spatial similarity, temporal similarity and content similarity measurement. Then we employ the supervised machine learning algorithm to address this problem.
(2) We present a supervised machine learning based solution to user identification solely based on UGCs. We first present several algorithms to measure the spatial similarity, temporal similarity and content similarity of any two UGCs, respectively. Based on these algorithms, we extract the spatial features, temporal features and content features from UGCs, and then pour them into a cascaded three-level machine learning based solution framework. Finally, we obtain a three-level classifier. In this classifier, we fuse the information redundancies of spatial, temporal and content dimension for user identification.
(3) Several user identification experiments based on three ground truth datasets are conducted. We study whether the base classifiers have impact on user identification accuracy. The results show that GraBoosting works best on three datasets. This helps us select a better base classifier for our experiments. We compared the proposed solution with existing works. The results show that the proposed solution enables to provide better performance.
(4) The importance of spatial, temporal and content similarity for user identification is identified. Eight similarity measurement algorithms fall into four classes. The spatial similarity holds the 1st place, the common string related similarity is on the 2nd class, the word vector related similarity algorithms is on the 3rd class, the temporal similarity ranks final.

The rest of the paper is organized as follows. We first introduce the related works in Section 2. Then in Section 3, we describe the preliminary concepts, and give the problem formulation. In Section 4, we present the solution framework and UGC-based user identification across OSNs. Then Section 5 shows the experiment results on social networks. In Section 6, we conclude the paper.

2. Related works

In an OSN, a user usually creates an identity and constitutes its three major dimensions namely Profile, Content and Network. Each dimension is composed of a set of attributes which describes her and differentiates her from others [7]. Existing works on user identification are mainly based on these three dimensions or the hybrid dimensions.

In some existing works, the researchers presented methods that only used profile attributes to identify a user across sites. Liu et al. [8] matched user accounts in an unsupervised approach using usernames. Zafarani et al. [9,10] presented a MOBIUS method to identify the user across sites based on the naming patterns of usernames. Perito et al. [11] introduced the idea of using username to match multiple online accounts of a user across sites. Liu et al. [12] analyzed usernames' characteristics including length, special character, numeric character etc., and proposed a weighting function of user identification based on the above characters. However, username is not always available, and even in some situation, the username is a numeric string automatically assigned by sites. This makes these existing schemes fragile. Motoyama et al. [13] extended the profile attribute set, and used name, city, school, location, age, email etc. to match user accounts. Iofciu et al. [14] used the similarity between users' profiles to identify users. Abel et al. [15] aggregated user profiles and matched users across systems. Raad et al. [16] addressed the user identification by providing a matching framework based on all the profile's attributes. The proposed framework allowed users to give more importance to some attributes and assign each attribute a different similarity measure. They concluded that user accounts could be accurately matched based on a set of attributes. However, the profile attributes do not require exclusivity and are easily faked by users for different purposes.

Some existing works studied the user identification problem solely based on user network. Zhou et al. [2] proposed a friend relationship-based user identification algorithm. It calculates a match degree for all candidate user matched pairs, and only pairs with top ranks are considered as identical users. Narayanan et al. [17] solely used network structure to analyze privacy and anonymity, which is closely related to user identification issue. Korula et al. [18] presented a mapping algorithm based on the degrees of unmapped users and the number of common neighbors, using two control parameters to fine-tune performance. Owing to the privacy setting, in many cases, the users' friend networks are not public and accessible across sites. Researchers attempted hybrid approaches to solve this issue. Bartunov et al. [19] considered both the profile and friend network, and proposed an approach based on conditional random fields to identify users. Bennacer et al. [20] also used the friend network and the publicly available profile to iteratively match profiles across social networks. Malhotra et al. [21] used the user profile and friend network to generate the user's digital footprints, and applied automated classifiers for user identification based user's footprints. The above studies show that the friend network has forceful and robust features for user identification. However, this information is often sparse, because