Procedia Computer Science

Volume 80, 2016, Pages 376–385

ICCS 2016. The International Conference on Computational Science

Identifying Users across Different Sites using Usernames

Yubin Wang[1,2], Tingwen Liu[1,2,*], Qingfeng Tan[1,2], Jinqiao Shi[1,2], and Li Guo[1,2]

[1] Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China
[2] National Engineering Laboratory for Information Security Technologies, Beijing, China
{wangyubin, liutingwen, tanqingfeng, shijinqiao, guoli}@iie.ac.cn

Abstract

Identifying users across different sites is to find the accounts that belong to the same individual. The problem is fundamental and important, and its results can benefit many applications such as social recommendation. Observing that 1) usernames are essential elements for all sites; 2) most users have limited number of usernames on the Internet; 3) usernames carries information that reflect an individual's characteristics and habits *etc.*, this paper tries to identify users based on username similarity. Specifically, we introduce the self-information vector model to integrate our proposed content and pattern features extracted from usernames into vectors. In this paper, we define two usernames' similarity as the cosine similarity between their self-information vectors. We further propose an abbreviation detection method to discover the initialism phenomenon in usernames, which can improve our user identification results. Experimental results on real-world username sets show that we can achieve 86.19% precision rate, 68.53% recall rate and 76.21% F1-measure in average, which is better than the state-of-the-art work.

Keywords: user identification, username similarity, self-information model, abbreviation detection

1 Introduction

Identifying users across different sites, which tries to find the accounts that belong to the same individual, is a fundamental and important problem. This work can be applied in many applications, such as user profiling and personalized recommendation. Given a targeted user, ArnetMiner [15] enriches the user's profile by integrating the information extracted from the corresponding accounts elsewhere. In [11], users' auxiliary information on Twitter are exploited to address the typical problems in single network-based recommendation solutions to recommend YouTube video.

In this paper, we focus on addressing the important problem using usernames, owing to following three reasons. First, usernames are essential elements for all sites, while user attributes and social behaviors do not exist in some sites or hard to collect for researchers. In this case, prior identification approaches [2, 10, 5, 15] designed for social networks do not work well. Even

*Tingwen Liu is the corresponding author of this paper.

 Selection and peer-review under responsibility of the Scientific Programme Committee of ICCS 2016

doi:10.1016/j.procs.2016.05.336

if all user attributes and social behaviors needed are available, our work is still valuable as it can be used to improve prior social graph based approaches. Second, most users have limited number of usernames on the Internet, and these usernames usually have the same or similar naming rules. Because it is hard for users to memory too many different and casual usernames. Third, usernames may also reflect the characteristics and habits of an individual. For example, username `shmilyszw` in CSDN consists of `shmily` (an abbreviation of "See How Much I Love You") and `szw` (probably an abbreviation of someone's name).

For two given usernames, this paper tries to determine whether they belong to the same individual based on username similarity and username abbreviation. Username similarity is intended to define how much similar the two usernames are, and username abbreviation is to check if one username (or its substrings) is an initialism of the other username (or its substrings). We assume that two usernames with high similarity and initialism phenomenon are very likely to belong to the same individual.

Distance metrics, such as Levenshtein distance, are intuitive and easy-to-implement tools to quantify username similarity. However, they are not the best choice. Because a username usually consists of multiple relatively independent parts, while these distance metrics do not consider the permutation of the username parts. This paper introduces the self-information model to quantify the similarity between usernames. We extract 1296 content features and 77 pattern features for each username, which are integrated as a vector by the self-information model with the self-information of each feature as its weight. Then we quantify the similarity of any two given usernames as the cosine similarity between their self-information vectors.

We reduce the problem of detecting the initialism phenomenon into the problem of getting the minimum number of meaningless characters for each username. A meaningless character is the one that is not a member of any word in a given username after splitting the usernames to get some non-overlapped words. Note that there may be multiple different ways to split a username. The problem is NP-hard and addressed in this paper based on the dynamic programming strategy.

We make three key contributions in this paper. First, we quantify username similarity based on the self-information vector model and our proposed content features and pattern features. Second, we propose a dynamic programming algorithm to detect the initialism phenomenon between usernames. Third, we conduct experiments on real-world username sets and validate the effectiveness of our work.

2 Related Work

Prior work on identifying users across different sites can be divided into three categories: user attribute based approaches, social graph based approaches and hybrid approaches.

User attribute based approaches [12, 8, 13, 4, 14] are designed for these sites where social network structures are unavailable. As a result, we could only obtain and leverage the attributes of users, especially usernames, to identify users across different sites. Perito *et al.* [8] used a 5-gram Markov Chain model to compute the username observation likelihood as the estimation the uniqueness of the username. Their work is limited to only using this single feature to link different usernames. Zafarnai *et al.* [13] extended this work and conducted a more in-depth analysis of the features of the usernames. They proposed the methodology MOBIUS to model the features of usernames according to the users' behavioral patterns when creating usernames and employed machine learning for effective identification. Then Zafarnai *et al.* [14] generalized their work in [12, 13], and give a further detailed discussion on the problem of user identification across social media. Our work is a user attribute based approach, that identifies user across