An Improved User Identification Method Across Social Networks Via Tagging Behaviors

1st Dongsheng Zhao
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
161050057@hdu.edu.cn

2nd Ning Zheng
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
nzheng@hdu.edu.cn

3rd Ming Xu
School of Computer-
Science and Technology
Hangzhou Dianzi University
Hangzhou, China
mxu@hdu.edu.cn

4th Xue Yang
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
153050004@hdu.edu.cn

5th Jian Xu
School of Computer Science and Technology
Hangzhou Dianzi University
Hangzhou, China
jian.xu@hdu.edu.cn

***Abstract*—User Identification problem is concerned with identifying the same person with multiple virtual identities across social network sites(SNSs). Most of the existing approaches pays close attention to the similarity of profile attributes, generate-contents and linkages of friends or simply combination of these features. Only one method analyzes the feasibility of user tags in User Identification problems, but does not analyze the particularity and the inconsistency of tags belong to users among different social networks. In this paper, an improved user identification method across social networks via tagging behaviors is proposed that a new symmetric variant of BM25 (BM25 is a bag-of-words retrieval function that ranks a set of documents based on the query terms appearing in each document, regardless of the inter-relationship between the query terms within a document) using the semantic relationships between inconsistent tags among different social networks. By using extracted features from the inconsistent tagging behaviors, profile attributes and SVM supervised learning techniques, a classifier is developed for performing user identity matching between two social network sites. Evaluation on Douban and Weibo real world data-set showed that the accuracy of the proposed method is 30% higher than that of the common tag-based approach.**

***Index Terms*—Social Network sites, User Identification, Semantic relatedness, Tag analysis, Machine Learning, Data mining**

I. Introduction

In recent years, the blossom of social network sites(SNSs) of various kinds has completely changed our life by providing everyone with the easy and fun of sharing our various information like never before(e.g., micro-blogs, images, videos, reviews, activities). However, different SNSs provide different services. To better take advantage of services provided by each social network [1], users tend to join multiple SNSs. User linkage was firstly formalized as connecting corresponding identities across communities [2]. Cross-social network user identification definition as shown in Fig. 1. Implications of linking user identities: First, the information integration of users dispersed in multiple social networks is conducive to fully understand users' interests and provide better recommendations or services; Second, [3] users with accounts on multiple SNSs allow us to integrate users' behavior patterns and solve problems that cannot be handled by data from only one site, such as cold-start and data sparsity issues in many predictive tasks. For example, a newly established social network site may not have enough historical data for recommendation or prediction system; Third, [4] identifying users across SNSs can help researchers to confirm users' identities, improve the consistency of user information and provide trust mechanisms among users. Fourth, identifying users across SNSs is helpful for searching malicious users' from multiple social network sites identities in security domain.

Several profile matching techniques [5], such as image recognition, syntactic, semantic, statistical, linguistic and network-based measures exist, all of which are built for solving specific problems according to their target domain. Different profile attributes are not public or available on different social network sites. In this paper, we only use information that does not involve private data(e.g.,username, domain, location), which is available from different social network sites. Reference [6] studied the feasibility of exploiting individual tagging practices to identify a user and link her/his social network accounts. But they did not consider the user's inconsistent tagging behaviors, such as users tag their published pictures with "food" on one SNS, but tag "cooking" on the same pictures on another one. We perform semantic relevance analysis on these tags for issues of users different tagging behaviors on different SNSs.

This work is supported by the cyberspace security Major Program in National Key Research and Development Plan of China under grant No.2016YFB0800201, the Natural Science Foundation of China under grant 61572165 and 61702150, the State Key Programof Zhejiang Province Natural Science Foundation of China under grant No.LZ15F020003, the Key research and development plan.

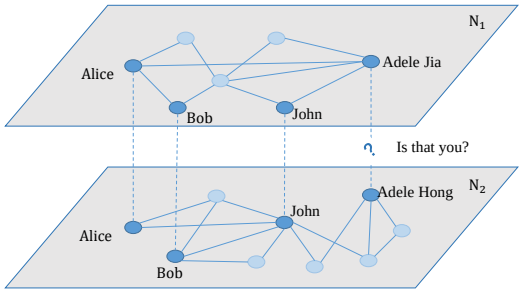

Fig. 1. The structure of matching users. N_1 and N_2 represents different two social networks respectively. The solid lines represents the friendship between users within the social network. The dotted lines represents the user's two accounts in different social networks.

Word2vec[1] is a very popular open source tool based on deep learning. It can learn the vector representations of words in the high-dimensional vector space and calculate the cosine-distances between words. That is to say, the tool can find the semantic relationships between tags. We apply it in a symmetric variant of BM25 [7] to calculate similar scores for users' inconsistent tagging sets on different SNSs. We combine users' profiles and tags to identity two user accounts belong to one real individual on different SNSs.

Our algorithm performs identity matching between two user identities on two different social networks for two scenarios:(a) matching users across two social networks - given two user identities from different SNSs, we decided whether these two user accounts belong to one real individual; (b) searching for a user - given a user's information on one SNS, we find his or her identity, with a similar name on another SNS; The main contributions of this paper are shown as follows:

1) In this paper, we proposes a new symmetric variant of BM25 that a semantic-based BM25. The semantic-based BM25 be used to calculate the semantic similarity scores of the two tag sets.
2) The paper proposes profile and tagging behaviors based user identification method with a greedy matching strategy.
3) We validate our proposed ideas and evaluate our algorithm through a comprehensive experimental study, using a real world data-set collected from Douban and Weibo. We compare it against measures like TF, TF-IDF, BM25 and variants of BM25. The experimental results show that our method outperforms common approaches and produces high-quality results.

The remaining of this paper is organized as follows. Section II briefly reviews the related work in the field of the user identity resolution. In Section III, we formally state the problem definition. Section IV introduces our methodology and implementation details. The experimental results and analysis are demonstrated in Section V. Finally, in Section VI, we draw our conclusions and sketch possible future evolution of our research.

[1]https://code.google.com/p/word2vec/

II. Related Work

In this section, we focus on summarizing research related to identifying individuals cross SNSs. User linkage was firstly formalized as connecting corresponding identities across communities [2] and a web-search-based approach was proposed to address it. Previous research can be categorized into four types: user-profile-based, user-network-based, user-generated-content-based and user-behavior-model-based. User-profile-based methods [5], [8]–[10] collected user profiles from multiple social network sites and then converted user profiles in vectors, of which each dimension corresponds to a profile field (e.g., username, birthday, hometown, etc.). Based on the completeness and connectivity of network topology structures, user-network-based methods [11], [12] categorized networks into two type: local network and global network. The local network was built from the ego-networks of user identities. The ego-network for each user identity was obtained through the one-hop neighborhoods(e.g. following/followee/friend relationships). The global network often indicated arbitrary merging graph such as a large sample or even complete social networks. User-generated-content-based methods [4], [8] collected personal identifiable information(e.g. users' core interests, users' writing style, users' unique footprints, etc.) from public pages of user-generated content. User-behavior-model-based methods [10], [13], [14] analyzed behavior patterns and built models from usernames, writing styles and activity trajectories.

Following traditional ways of classifying data mining and machine learning models, existing research can be summarized into three models: supervised, semi-supervised and unsupervised models. Supervised model [4], [9], [15] require a large number of labeled data sets to train the model. To overcome the difficulty of labeling, [14] proposed a semi-supervised multi-objective framework to jointly model inconsistent behaviors and structures. Reference [13] further proposed an unsupervised model, which focused on task to decide whether cross-platform user identities with the same username belongs to the same natural person.

Prior works were built for solving specific problems according to their target domain, and did not consider the user's tag data in social networks. Reference [6] focused on linking users in tagging systems and proposed a method to linearly combine the edit distances of usernames and the similarities between the tags provided by users. Their method simply considered tagging behaviors and didn't analyze the specificity and inconsistency of the tags. The proposed method has achieved very good performance.

III. Problem Statement

In this section, we introduce some basic notations and definitions for user identity linkage task. Usually such profile data is structured as attribute-value pairs(e.g., username:brown mark, screen-name:seclab, location:hangzhou, etc.). SNSs are web application in which users generate contents(e.g., photos, videos, blogs, essays, etc.) and annotate it with a list of freely chosen keywords called tags. These tags are convenient for

users to categorize and retrieve their contents and other users' contents, especially in the era of big data. In our proposed approach we study the inconsistent tagging behaviors of users in different SNSs. The basic notations are defined as below:

- Let User Identity u represent the unique social account representation on a social network for a real natural person $\mathcal{P}$. It can be composed of three components: Profile, Tag, Content. Profile $\vec{p}_u$ refer to a set of user description features such as username, location, description among other attributes. We refer to the keywords that users use to mark their own generated-content as tags. Users can easily retrieve and sort a number of data on social networks. Tag $\vec{t}_u$ includes a set of keywords that used to mark information generated by users and facilitate retrieval of the information. Content $\vec{c}_u$ includes a set of attributes that represent the activities that user is involved in and includes time and text.
- We denote Social Network as $\mathbb{N}, \mathbb{N} = (\mathcal{P}, \mathcal{F})$ where $\mathcal{P} = \{u_1, u_2, \ldots, u_n\}$ denotes the set of all user accounts on $\mathbb{N}$ and $\mathcal{F} = \{f_1, f_2, \ldots, f_n\}$ denotes the set of all user attributes on $\mathbb{N}$ and $f_i = \{\vec{p}_u, \vec{t}_u, \vec{c}_u\}$ denotes the set of a user's attributes.

DEFINITION 1. *User Identification across Social Networks. Given two online social networks $\mathbb{N}^s$ (source site) and $\mathbb{N}^t$ (target site), the task of user identity linkage is to predict whether a pair of user identities u_i^s and u_j^t chosen from $\mathbb{N}^s$ and $\mathbb{N}^t$ respectively belong to the same real natural person, the user identification procedure attempts to learn an identification function $\mathcal{G} : \mathbb{N}^s \times \mathbb{N}^t \longrightarrow \{0, 1\}$ such that:*

$$\mathcal{G}(u_i^s, u_j^t) = \begin{cases} 1, & \text{if } u_i^s \text{ and } u_j^t \text{ belong to same person,} \\ 0, & \text{otherwise.} \end{cases}$$

It is noteworthy to point that to solve this problem straight forwardly by checking every user pair without any attribute filtering requires a high cost of computing. In this paper, the problem of user identification is divided into two steps: at first, we select the candidate matching pairs according to the similarity of user's screen-name and then determine the matching users.

IV. Our Algorithm

In this section we focus on the method of identifying users across social network sites based on their tagging behaviors and profiles. We determine whether the candidate pairs really match into a binary classification. The framework for user identification across SNSs is shown in Fig. 2.

A. Matching Users based on Profile

We consider profile attributes that are publicly available on both networks, such as username, domain, location, etc. The following algorithms is used to calculate the similarity between the two strings:

- LCS similarity - The Longest Common Sub-string(LCS) [16] repeatedly finds and removes the longest common substring in the two compared strings up to a minimum length. A similarity measure can be calculated by dividing the length of the longest common sub-string by the minimum length of the two original strings. This algorithm is more suitable for Strings with prefix or suffix (e.g., seclab, 666seclab, seclab666). For example, the longest length of the common sub-strings of the two strings "seclab" and "seclab605" is 6, the LCS("seclab", "seclab666") $= \frac{6}{6} = 1.0$.
- Levenshtein similarity - edit distance [17] is defined to be the smallest number of edit operations(inserts, deletes, and substitutions) required to change one string into another. The edit distance can be converted into a similarity measure(between 0 and 1), by dividing it by the maximum length of the two strings, such as LS("seclob666l", "seclab666") $= 1 - \frac{2}{10} = 0.8$.
- Jaccard similarity - The Jaccard [18] measure simply calculates the ratio of equal tokens in the union of tokens of both strings. The range of the Jaccard similarity is between 0.0 and 1.0, 1 being an exact match and o otherwise. First, the string be divided into a single character. Second, Jaccard similarity be used to measures character set of two strings respectively. For example, the Jaccard similarity of "mark brown" and "brown mark" is 1.0.
- CLL similarity - The CLL similarity is a combination of Longest Common Sub-string and edit distances that form a another string similarity measure $(\frac{editdistances}{editdistance+LCS})$. The range of the CLL similarity is between 0.0 and 1.0. For example, the edit distance and longest common substring of the strings "lab" and "seclab666" are 6, 3 respectively, $CLL($ "lab", "seclab666" $) = \frac{3}{6+3} = 0.33$.

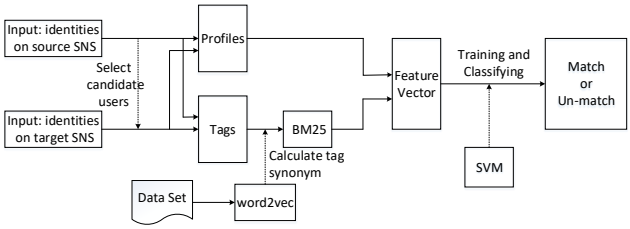

Fig. 2. Framework for users identification across SNSs.

B. Matching Users via Semantic-Based Tags

Tagging behaviors are influenced highly by the social network site's domain and design choices. For example, people tag different music items, images and diaries on Douban, and tag different images, micro-blogs and even themselves on Weibo. Although these tags are marked with different content, they are all related to the user's interests. It gives us an opportunity to analyze these unstructured tags to identify the identities of different social network sites that belong to the same user in real life. If a document contains a tag 10 times, it is not 10 times as relevant as appearing just once. Similarly, if a user use a tag 10 times, it is not 10 times as relevant as

using a tag just once. For example, in our experimental dataset "humanity" is used for one-third of the diaries in Douban, conversely, "punk" is used only a few times. Hence, "punk" is a very discriminative tag when matching against Douban, while "humanity" is hard to distinguish. Thus, we use BM25 together with a social network site specific IDF and average number of tags. Fortunately, the social network site has already counted the total number of each tag marked by users, which is represented by the TF value.

$$\mathrm{IDF}(q_i) = max(0, \log \frac{N - n(q_i) + \frac{1}{2}}{n(q_i) + (1 - \frac{1}{2})}) \quad (1)$$

$$\mathrm{TF}(q_i, \vec{t}_t) = \frac{f(q_i, \vec{t}_t) \cdot (k_1 + 1)}{f(q_i, \vec{t}_t) + k_1 \cdot (1 - b + b \cdot \frac{|\vec{t}_t|}{avgdl})} \quad (2)$$

$$w(\vec{t}_s, \vec{t}_t) = \sum_{i=1}^{|\vec{t}_s|} \mathrm{IDF}(q_i) \cdot \mathrm{TF}(q_i, \vec{t}_t) \cdot \frac{f(q_i, \vec{t}_s) \cdot (k_2 + 1)}{k_2 + f(q_i, \vec{t}_s)} \quad (3)$$

where $n(q_i)$ is the total number of users using the tag q_i on the SNS, $avgdl$ is the average number of tags used by all users. However, because of the inconsistency of user tagging behaviors, this method does not improve the matching very well. Whats more, the application scenario of BM25 is text, which is different with the situation of a small number of tags in this paper, so it cannot be applied directly. Although [6] leveraged social network specific IDF with BM25(Best Match), their results are still not good enough. For example, users tagged their published pictures with "food" on Douban, but "cooking" on the same pictures on Weibo, f("food","cooking") $= 0$. So We propose a semantic-based tag matching method. The semantic-based matching extends the tag value matching process, so that tag value that do not have a 1-1 relationship but are related regardless, will also be taken into consideration. For example, f("food","cooking") $= 1$. Then $f(q_i, \vec{t}_t)$ in (2) can be defined as:

$$f(q_i, \vec{t}_t) = \sum_{p \in \vec{t}_t} sem(q_i, p) \quad (4)$$

$$sem(q_i, p) = \begin{cases} tf(p), & \text{if } syn(q_i, p) > t \\ 0, & \text{otherwise.} \end{cases} \quad (5)$$

where $tf(p)$ is the frequency of a tag p used by a user, $syn(q_i, p)$ is used to calculate the relevance of two tags or whether two words are synonyms. The word2vec model is used to convert q_i and p into word space vectors, then the cosine-distance between the two vectors is calculated.

C. Matching Algorithm

We assume that each user has at most one identity in a social network, such as one user can only match at most one identity in another social network. We are inspired by a Stable Marriage Problem(SMP) to propose a stable bi-directional matching algorithm, that is, both sides of the match are optimal.

Algorithm 1: Picking out the user's candidate matching user from $\mathbb{N}^t$

Input: $\mathbb{N}^s$, $\mathbb{N}^t$
Output: *Candidate matching user sets for each user in* $\mathbb{N}^s$

$Candidates \leftarrow \{\}$;
foreach $u \in \mathbb{N}^s$ **do**
  $u_{top-k} \leftarrow$ select user from $\mathbb{N}^t$ based on name;
  $\vec{t}_{vector} \leftarrow \{(u, u_{top-1}), (u, u_{top-2}), \cdots\}$;
  $labels, scores \leftarrow classification(\vec{t}_{vector})$;
  $u_{candidate} \leftarrow u_{top-k}, (labels, scores)$;
  $Candidates \leftarrow \{u, u_{candidate}\}$;
end foreach
return $Candidates$

We first look for candidate matching pairs that are optimal solutions to each other. Algorithm 1 is used to select candidate matching users from the target SNS for each user in the source SNS, and multiple candidates of one user have been sorted in descending order according to the matching score. Algorithm 2 prunes the results of Algorithm 1, judges the error as the candidate matching pairs, and extracts the correct matching pairs. Algorithm 2 adopts a greedy strategy and extracts the best matching pair until the end of the algorithm convergence. For example, (s1,s2,s3) are in social network site SNS1 and (u1,u2,u3) are in social network site SNS2, respectively, Algorithm 1 yields output: $\mathcal{D}_{candidates}$ = {s1:{u1:2(the matching score), u2:1}, s2:{u1:2, u2:1}, s3:{u1:3, u2:2, u3:1 }} and $\mathcal{W}_{candidates}$ = {u1:{s1:1}, u2:{s1:3, s2:2, s3:1}, u3:{s2:3, s3:1}}. The output of Algorithm 1 is given to Algorithm 2 as input, then Algorithm 2 converges after five iterations and yields output the final stable matching results. The results of each step of the iteration are:

1: $\mathcal{M}_{match} = \{s1 : u1\}$;
2: $\mathcal{D}_{candidates}$ = {s2:{u2:1}, s3:{u2:2, u3:1}}, $\mathcal{W}_{candidates}$ = {u2:{s2:2, s3:1}, u3:{s2:3, s3:1}};
3: $\mathcal{M}_{match} = \{s1 : u1, s2 : u2\}$;
4: $\mathcal{D}_{candidates}$ = {s3:{u3:1}}, $\mathcal{W}_{candidates}$ = {u3:{s3:1}};
5: $\mathcal{M}_{match} = \{s1 : u1,\ s2 : u2,\ s3 : u3\}$.

V. Experiments

In this section, we present our experimental campaign aimed at determining the performances of our approach.

A. Experiment Setup

Data Collections: In order to gain real data, we collect data from two large and popular social networks, Douban and Weibo, by making use of Web Browser Automation. We develop web crawlers that specifically crawl user information on Weibo and Douban web sites. The crawler automatically extracts the user profile, tags and text content from the two SNSs. Table I lists the information used for each social plat-

Algorithm 2: Function stable bi-directional matching algorithm

Input: $\mathbb{N}^s$, $\mathbb{N}^t$, $\mathcal{D}_{candidates}$, $\mathcal{W}_{candidates}$
Output: a stable matching $\mathcal{M}_{match}$

$\mathcal{M}_{match} \leftarrow \{\}$;
while $\mathcal{D}_{candidates}$ *is not None or* $\mathcal{W}_{candidates}$ *is not None* **do**
  foreach $u \in \mathcal{D}_{candidates}$ **do**
    $u_c \leftarrow \mathcal{D}_{candidates}[u]_{top}$;
    if u_c *in* $\mathcal{W}_{candidates}$ **then**
      $u2 \leftarrow \mathcal{W}_{candidates}[u_c]_{top}$;
    else
      Delete $\mathcal{D}_{candidates}[u]_{top}$;
      if $\mathcal{D}_{candidates}[u]$ *is None* **then**
        Delete $\mathcal{D}_{candidates}[u]$;
      else
        continue;
      end if
    end if
    if $u2$ *not in* $\mathcal{M}_{match}$ **then**
      if $u == u2$ **then**
        $\mathcal{M}_{match} \leftarrow \{u : u_c\}$;
      else
        Delete $\mathcal{D}_{candidates}[u]_{top}$;
        if $\mathcal{D}_{candidates}[u]$ *is None* **then**
          Delete $\mathcal{D}_{candidates}[u]$;
        end if
      end if
    else
      Delete $\mathcal{W}_{candidates}[u_c]_{top}$;
      if $\mathcal{W}_{candidates}[u_c]$ *is None* **then**
        Delete $\mathcal{W}_{candidates}[u_c]$;
      end if
    end if
  end foreach
end while
return $\mathcal{M}_{match}$

TABLE I
A SUMMARY OF USER FEATURES USED FOR EACH DATA SET.

Data Set	*User Features*
Weibo	username; location; URL; description; tag; text
Douban	username; location; URL; description; tag; text

form. The dataset for the word2vec is based on the wikidata-corpus[2].

- *Weibo*[3]:Weibo is one of the most popular Chinese micro-blogging websites with 165 million daily active users, similar to a hybrid of Twitter and Facebook.
- *Douban*[4]:Douban is a Chinese social networking service web-site, with 300 million monthly active users, allowing registered users to record information and create content related to films, books, music, recent events and activities in China.

[2]https://dumps.wikimedia.org/zhwiki/latest/zhwiki-latest-pages-articles.xml.bz2
[3]https://weibo.com/
[4]https://www.douban.com/

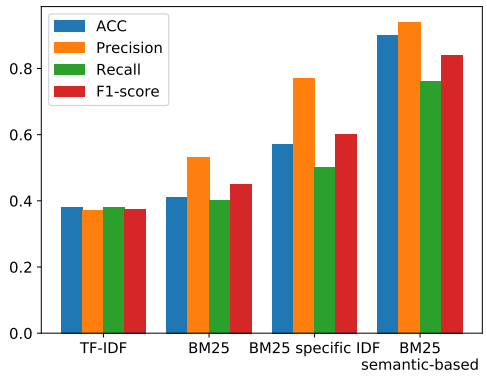

Fig. 3. Accuracy, Precision, Recall and F1-score by setting Douban as source SNS and Weibo as target SNS.

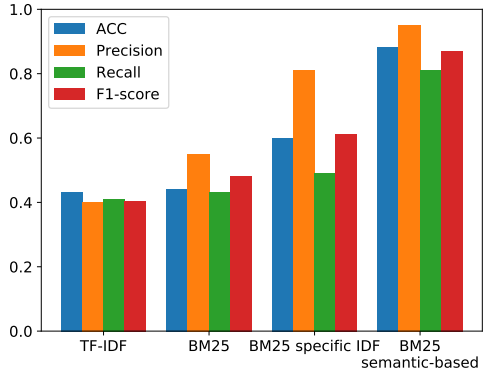

Fig. 4. Accuracy, Precision, Recall and F1-score by setting Douban as source SNS and Weibo as target SNS

It is very difficult to manually annotate the ground-truth user linkage pairs across different SNSs. Fortunately, on Douban, some users exhibit their Weibo identity or URLs in personal descriptions, which can be used as ground truth of user identity matchings between the two SNSs. We crawl 76415 accounts from Douban randomly, and find 6258 of them post their Weibo identity or URL in their home pages. After removing 98 canceled accounts, we finally get a total of 6162 accounts of user. In this paper, We randomly selected 1000 users from data set as testing set, and the remaining 5162 users data were used as training set.

A remarkable feature of the data set is that only a few tags occur in more than one SNSs: less than 40% of the same tags are used in two social network sites. But when we take the tag up to the semantic layer, up to 80% tags on both social networks are semantically similar. Interestingly, only 20% of the users have tags on Douban, while 80% of the users have tags on Weibo, which means that a small number of users on Douban prefer to use tags.

TABLE II
RESULTS OF TF-IDF, BM25, BM25 SPECIFIC IDF AND OUR ALGORITHM, ALL IMPROVEMENTS ARE SIGNIFICANT

Strategy	*wb → db*				*db → wb*			
	MRR	***S@1***	***S@5***	***S@10***	***MRR***	***S@1***	***S@5***	***S@10***
TFIDF	0.45	0.38	0.54	0.57	0.48	0.43	0.54	0.60
BM25	0.47	0.41	0.55	0.57	0.48	0.44	0.55	0.60
BM25 specific IDF	0.52	0.50	0.55	0.57	0.52	0.49	0.57	0.60
Semantic-based BM25	**0.73**	**0.70**	**0.75**	**0.75**	**0.77**	**0.72**	**0.81**	**0.82**

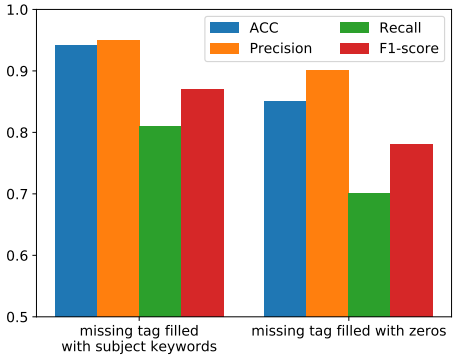

Fig. 5. Accuracy, Precision, Recall and F1-score by setting Weibo as source SNS and Douban as target SNS

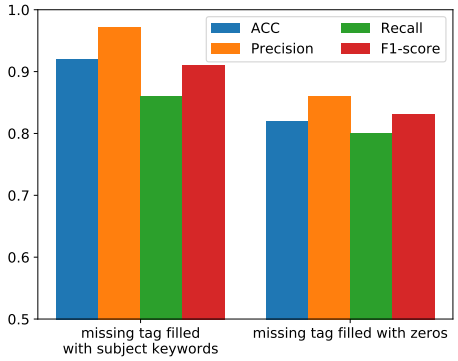

Fig. 6. Accuracy, Precision, Recall and F1-score by setting Douban as source SNS and Weibo as target SNS

Dealing with Missing Information. When generating the eigenvectors from users, it is not uncommon to observe a significant amount of information missing among SNSs. Therefore previous approaches [10], [13] constructed the models where a missing feature is automatically filled with zeros, unfortunately, this is not the case for our approach. We observe that tags are closely related to users' interests. Sometimes a user doesn't mark any tags, then the LDA model will be used to extract subject keywords from user-generate-content as his/her tags. Fig. 5 setting Weibo as the source social network and Douban as the target social network. Fig. 6 setting Douban as the source social network and Weibo as the target social network. In Fig. 6 and Fig. 5, the results indicate that when the user's tags are lost, we use the subject keyword extracted from user-generate-content as his/her tags to improve the performance of the algorithm and also increase the robustness of the algorithm.

Evaluation metrics. To measure the quality of the user match rankings we use MRR and S@k. MRR(Mean Reciprocal Rank) indicates at which rank the correct profile occurs on average. The success at rank k(S@k) stands for the mean probability that the correct identity occurs within the top k of the ranked results. Accuracy($\frac{|TP|+|TN|}{|TP|+|TN|+|FP|+|FN|}$), Precision($\frac{|TP|}{|TP|+|FP|}$), Recall($\frac{|TP|}{|TP|+|FN|}$) and F1-score($\frac{2 \cdot Precision \cdot Recall}{Precision+Recall}$) are the most common used evaluation metrics for user identity matching and user identity resolution. Thus in our experiments, we utilize these four metrics defined below to measure performances of our method and compare with other methods.

B. Comparison with Related Methods

For identifying users across SNSs based on their tagging behaviors, we experiment with standard techniques like TF-IDF, BM25 and [6] a variant of BM25 using site specific statistics and compare them against our semantic-based variant of BM25 using word2vec. As shown in Fig. 3, Fig. 4 and Table II, our method has a good performance whether we setting Weibo as a source SNS and Douban as a target SNS and vice versa. From Table II, the results of our method is 20% higher than other methods. In Fig. 4 and Fig. 3, our method has obviously improved on ACC, Precision, Recall and F1-score.

C. Performance Analysis with k and t varied

Fig. 8 and Fig. 7 shows that we have selected users of the top k user-name similarity scores from the data set as candidate users. The results indicate that the accuracy rate gradually decreases with the increase of k, but the precision changes very little. When we select 15 candidate users, the Accuracy is 85%. The possible reason is that the sample data set is not large enough.The positive and negative samples in the sample set used in our classification model are 1:1. In actual situations, when we search for candidate matching users based on a user name in another social network, we often search for more than one user. Experimental results show that users generally have similar names in different social network sites.

The performance of only using tagging behaviors to identify users across SNSs with varied t is shown in Fig. 9. The results indicate that the accuracy of the experiment increases with the

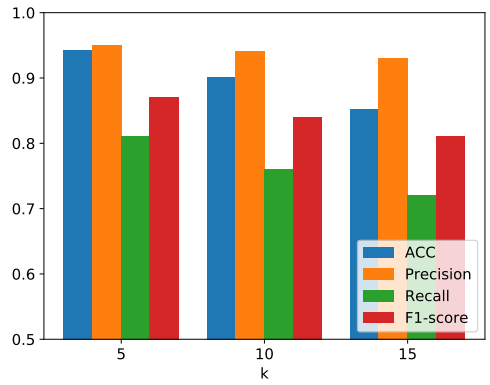

Fig. 7. Results with k varied by setting Weibo as source SNS and Douban as target SNS

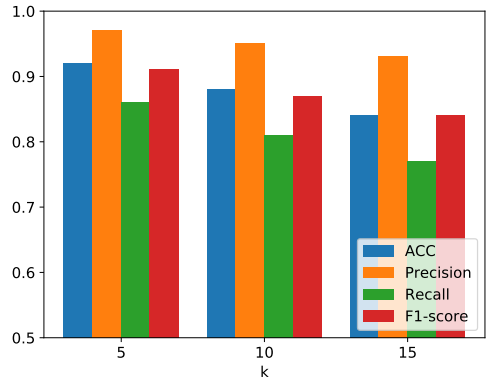

Fig. 8. Results with k varied by setting Douban as source SNS and Weibo as target SNS

increase of t and the accuracy reaches the maximum when $t = 0.8$. In our experiment, we finally chose the $t = 0.8$.

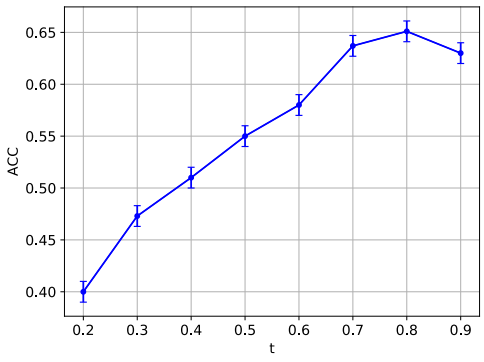

Fig. 9. Accuracy with t varied only use tags to identifying user.

VI. CONCLUSION & FUTURE WORK

In this paper we investigate whether users can be identified across SNSs by analyzing their tagging behaviors and profile attributes. According to our analysis, most people use different keywords to represent the same semantic tags in different social networks. According to this phenomenon, we propose a users identification method leveraging users' tagging behaviors and profiles to match users' identities. The experimental results indicate that our method can obtain better performance on accuracy, precision, recall, F1-score, MRR and S@k compared to existing methods. However, only two SNSs are adopted as data sources. In the further research, we will increase the number of SNSs and obtain more data sets to test and revise our method for the purpose of obtain more stable and reliable results.

REFERENCES

[1] K. Shu, S. Wang, J. Tang, R. Zafarani, and H. Liu, "User identity linkage across online social networks: A review," *SIGKDD Explorations*, vol. 18, no. 2, pp. 5–17, 2016.

[2] R. Zafarani and H. Liu, "Connecting corresponding identities across communities," in *ICWSM*. The AAAI Press, 2009.

[3] C. Li and S. Lin, "Matching users and items across domains to improve the recommendation quality," in *KDD*. ACM, 2014, pp. 801–810.

[4] Y. Nie, Y. Jia, S. Li, X. Zhu, A. Li, and B. Zhou, "Identifying users across social networks based on dynamic core interests," *Neurocomputing*, vol. 210, pp. 107–115, 2016.

[5] K. Cortis, S. Scerri, I. Rivera, and S. Handschuh, "An ontology-based technique for online profile resolution," in *SocInfo*, ser. Lecture Notes in Computer Science, vol. 8238. Springer, 2013, pp. 284–298.

[6] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff, "Identifying users across social tagging systems," in *ICWSM*. The AAAI Press, 2011.

[7] S. E. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[8] X. Mu, F. Zhu, E. Lim, J. Xiao, J. Wang, and Z. Zhou, "User identity linkage by latent user space modelling," in *KDD*. ACM, 2016, pp. 1775–1784.

[9] O. Peled, M. Fire, L. Rokach, and Y. Elovici, "Matching entities across online social networks," *Neurocomputing*, vol. 210, pp. 91–106, 2016.

[10] R. Zafarani and H. Liu, "Connecting users across social media sites: a behavioral-modeling approach," in *KDD*. ACM, 2013, pp. 41–49.

[11] S. Bartunov, A. Korshunov, S.-T. Park, W. Ryu, and H. Lee, "Joint link-attribute user identity resolution in online social networks," in *Proceedings of the 6th International Conference on Knowledge Discovery and Data Mining, Workshop on Social Network Mining and Analysis*. ACM, 2012.

[12] F. Buccafurri, G. Lax, A. Nocera, and D. Ursino, "Discovering missing me edges across social networks," *Inf. Sci.*, vol. 319, pp. 18–37, 2015.

[13] J. Liu, F. Zhang, X. Song, Y. Song, C. Lin, and H. Hon, "What's in a name?: an unsupervised approach to link users across communities," in *WSDM*. ACM, 2013, pp. 495–504.

[14] S. Liu, S. Wang, F. Zhu, J. Zhang, and R. Krishnan, "HYDRA: large-scale social identity linkage via heterogeneous behavior modeling," in *SIGMOD Conference*. ACM, 2014, pp. 51–62.

[15] P. Zhang, T. Lu, H. Gu, and N. Gu, "Identifying user identity across social network sites based on overlapping relationship and social interaction," in *ChineseCSCW*. ACM, 2017, pp. 25–32.

[16] C. Friedman and R. Sideli, *Tolerating spelling errors during patient validation*. Academic Press Professional, Inc., 1992.

[17] G. Navarro, "A guided tour to approximate string matching," *ACM Comput. Surv.*, vol. 33, no. 1, pp. 31–88, 2001.

[18] P. Jaccard, "Etude de la distribution florale dans une portion des alpes et du jura," *Bulletin De La Societe Vaudoise Des Sciences Naturelles*, vol. 37, no. 142, pp. 547–579, 1901.