What's in a Name? An Unsupervised Approach to Link Users across Communities*

Jing Liu
Harbin Institute of Technology
Harbin 150001, P.R. China
jliu@ir.hit.edu.cn

Fan Zhang
Nankai University
Tianjin 300071, P.R.China
zhangfan555@gmail.com

Xinying Song
Microsoft Research
Redmond, WA 98052, USA
xinson@microsoft.com

Young-In Song
NHN Corporation
Seoul, Korea
youngin.song@gmail.com

Chin-Yew Lin
Microsoft Research Asia
Beijing 100080, P.R. China
cyl@microsoft.com

Hsiao-Wuen Hon
Microsoft Research Asia
Beijing 100080, P.R. China
hon@microsoft.com

ABSTRACT

In this paper, we consider the problem of linking users across multiple online communities. Specifically, we focus on the *alias-disambiguation* step of this user linking task, which is meant to differentiate users with the same usernames. We start quantitatively analyzing the importance of the *alias-disambiguation* step by conducting a survey on 153 volunteers and an experimental analysis on a large dataset of About.me (75, 472 users). The analysis shows that the *alias-disambiguation* solution can address a major part of the user linking problem in terms of the coverage of true pairwise decisions (46.8%). To the best of our knowledge, this is the first study on human behaviors with regards to the usages of online usernames. We then cast the *alias-disambiguation* step as a pairwise classification problem and propose a novel unsupervised approach. The key idea of our approach is to automatically label training instances based on two observations: (a) rare usernames are likely owned by a single natural person, e.g. **pennystar88**[1] as a positive instance; (b) common usernames are likely owned by different natural persons, e.g. **tank** as a negative instance. We propose using the n-gram probabilities of usernames to estimate the rareness or commonness of usernames. Moreover, these two observations are verified by using the dataset of Yahoo! Answers. The empirical evaluations on 53 forums verify: (a) the effectiveness of the classifiers with the automatically generated training data and (b) that the rareness and commonness of usernames can help user linking. We also analyze the cases where the classifiers fail.

Categories and Subject Descriptors

H.4 [**Information Systems Applications**]: Miscellaneous

General Terms

Algorithms, Experimentation

Keywords

Social network, User linking, Pairwise classification

1. INTRODUCTION

In recent years, the unprecedented amount of data from online communities such as online forums, community-based question answering (cQA), and micro blogs is becoming available for fundamental research on social networks such as user influence estimation [32], user expertise estimation [17], community structure analysis [23], etc. However, most work has concentrated on single communities (i.e. without crossing websites) due to the lack of explicit links between users across communities. Building such links is not a trivial matter as we will show later.

Zafarani et al. [33] first formalized the task of user linking as linking users across multiple communities who belong to a single natural person in the real world. Linking users across sites can enable many applications. For example, it can be used for understanding user migration patterns in social media [14] and allows community owners to learn how to retain or increase site traffic; it can be used for aggregating the public profile data of users (belonging to one natural person) from different communities to solve the cold-start problem in recommendation or personalization [1, 19]; it can protect users from privacy risks arising from large amounts of publicly available user information [18, 19]; it can enable users to keep up-to-date with their online friends from different communities in an integrated environment [29]; it can be used to create cross-community expert recommendation systems and mine influential people on a global social graph.

The task of user linking aims to solve the problem of linking users across multiple online communities. These users are identifiable by their usernames. This task can be divided into two steps. In the first step, we determine the set of the same usernames appearing in multiple communities and develop algorithms to decide if a username in this set is owned by a single natural person. We call this step the *alias-disambiguation* step, i.e. differentiating users under the same usernames. In the second step, we deal with a natural person using different usernames across sites. We call this step the *alias-conflation* step, i.e. finding a natural person using different usernames. In this paper, we focus on the first step, i.e. the *alias-disambiguation* step and

*This work was done when Jing Liu and Fan Zhang were visiting students at Microsoft Research Asia, and Young-In Song worked as a researcher at Microsoft Research Asia.

[1]We use pseudo usernames in this paper to protect people's privacy.

WSDM'13, February 4–8, 2013, Rome, Italy.

leave the *alias-conflation* step for future work. Our survey of 153 volunteers and an analysis of a data set crawled from About.me[2] (75,472 users) show that the solution of *alias-disambiguation* can address a major part of the user linking problem in terms of the coverage of true pairwise decisions (46.8%, see Section 3.3). Specifically, 89.17% of participants in our survey reported that they prefer to use one main username across multiple communities. This implies that most of their online activities can be linked through their main usernames. A similar finding was also reported by [33] that 59% of their sampled users identified themselves with the same usernames across social networks. In addition, this problem cannot be trivially solved, since the baseline of treating two identical usernames always belonging to a single natural person only achieves an accuracy of 56.44% on a data set of 122 usernames.

We cast the *alias-disambiguation* step of user linking task as a pairwise classification problem. Given any two users from two different communities under the same username, a classifier is learned to decide whether these two users belong to a single natural person. Supervised methods are very effective but require manually annotated training corpora. To learn such a classifier without manually labeled data, we propose a novel unsupervised approach to automatically generate training data and then apply a standard machine learning method to learn classifiers. Our evaluation results show the classifiers trained with the automatically generated training data are effective. This method achieves an accuracy of 92.08% over a set of forum data consisting of 53 travel forums, 7.2 million threads, and 1.94 million users. Simply assuming the same usernames are always owned by a single natural person has an accuracy of 56.44%.

The key idea of our approach for automatic acquisition of training data is based on the following two observations:

- Rare usernames are likely owned by a single natural person, e.g. **pennystar88** (positive case).
- Common usernames are likely owned by different natural persons, e.g. **tank** (negative case).

Using n-gram probability of a username that estimates rareness or commonness of a username, training instances can be automatically assigned positive (low n-gram probability) or negative (high n-gram probability) labels. These two observations are supported by the data collected from Yahoo! Answers[3] in which a single username can be used by multiple natural persons. Please see Section 4 for more details.

The main contributions of this paper are:

- We demonstrate the importance of the *alias-disambiguation* step by conducting a survey and an experimental analysis on a dataset of About.me (see Section 3). To the best of our knowledge, it is the first study on human behavior on the usage of online usernames.
- We show that the rareness of a username measured by its n-gram probability is a good indicator of how likely it belongs to a single natural person (see Figure 4).
- We demonstrate how to automatically create a labeled training data set with the knowledge of the n-gram probability of a username (see Section 4).
- We verify the effectiveness of the classifiers trained with the automatically generated training data. We also examine the cases where our classifiers would fail. (see Section 6).

The remaining of this paper is organized as follows. Section 2 summarizes related work. In Section 3, we formally define the user linking problem and show the importance of *alias-disambiguation* step. Section 4 analyzes the relationship between n-gram probability of a username and its ownership, and shows how to utilize the n-gram probability of a username to automatically create a labeled training data set. Section 5 describes the features and models we propose. Section 6 describes evaluation setups, reports the experimental results and gives detail analysis. We conclude this paper in Section 7 and discuss future work.

2. RELATED WORK

2.1 User Linking across Communities

To the best of our knowledge, there has not been much research conducted on the problem of user linking. Zafarani et al. [33] firstly formalized the problem and proposed a web search based approach to address it. This approach is mainly based on two assumptions: (a) the URL of a user profile page contains the corresponding username; (b) a user profile page usually contains another username that is used by the same natural person on another community. However, our experiments suggest that these two assumptions do not hold for 75.47% of the cases in the data we collected. (see Section 6).

Iofciu et al. [12] 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. The proposed techniques are dependent on specific types of social networks (e.g. tagging services).

[29, 25, 19] collected user profiles from multiple social networks and proposed representing user profiles in vectors, of which each dimension corresponds to a profile field (e.g. username, description, profile image, location, etc.). Once the profile vectors are generated, both unsupervised and supervised approaches can be applied to link users. Vosecky et al. [29] used (unsupervised) comparison algorithms to compute the similarity scores between the user vectors, and those with scores larger than a pre-defined threshold are deemed to be the same person. [25, 19] used similarity vectors derived from annotated users as training instances, and upon which supervised classifiers are trained. The supervised approaches achieve high accuracy with regards to the user linking task. However, the types of identifiable personal information [22] are very different from site to site. Since it is impossible to manually label training instances for each online community, the above mentioned supervised approaches are not easily scaled. To address this challenge, we propose a novel unsupervised approach to automatically generate training instances, which can be adapted to any type of online communities trivially.

Another major limitation of the existing techniques is their dependencies on user profile pages to be publicly available [29, 25, 19], which is actually not the case for many online communities. Detailed analysis can be found in our experiments in Section 6. In our approach, we only collect personal identifiable information from the public user gen-

[2] `https://about.me`
[3] `http://answers.yahoo.com`

erated content (UGC) pages, which are accessible in most online communities.

2.2 De-anonymization on Social Networks

Researchers from data security and privacy area considered that linking the records from different anonymized databases may expose sensitive privacy information of the users [3, 22]. The main findings can be summarized into two categories: (a) It is pointed out that rare attribute values in high-dimensional sparse data sets can help de-anonymization [20, 10]. (b) [21, 15] found that an anonymized network can be successfully re-identified by only utilizing the structures of the social networks, because the online friends (neighbors) of a natural person are usually a similar group of people on different social graphs. In this paper, we have similar hypothesis on structure features for user linking (see Section 5).

2.3 Authorship Classification

Authorship classification is a task that identifies the authors of articles according to their writing styles by analyzing the corresponding article content. The findings in authorship classification can also help user linking. Novak et al. [24] proposed a language model based approach for authorship classification in online forums. In this paper, we use language model for feature extraction in the user linking task. Rao et al. [27] found that the usages of function words are indicative of authorship identification in mailing lists. We also make use of function words for user linking.

2.4 Entity Resolution

User linking is similar to several problems that have been studied for decades across multiple research communities. Examples include *coreference resolution* in natural language processing [6, 4, 28], where different mentions of the same underlying entity in free texts should be linked; *record linkage* in databases [5, 9], where two records from databases referring to the same real-world object should be identified; *people name disambiguation* in information retrieval [26, 13], which aims to assign documents to their authors with the same name, given thousands of documents belonging to different persons with the same name. Such problems fall under the umbrella-term *entity resolution* [5].

The state-of-the-art systems for addressing these problems mainly adopt two types of supervised approaches: (a) pairwise classification [4, 28] and (b) supervised and semi-supervised clustering [6, 5, 26, 13]. Given a username shared by multiple users, a graph of the corresponding users (nodes) can be constructed. However, the graphs of many usernames only contain two or three nodes (Figure 5). The clustering approaches cannot work well in such cases. Hence, in this paper, we cast the *alias-disambiguation* step as a pairwise classification problem. The supervised approaches usually outperforms unsupervised approaches in *entity resolution* tasks [9, 26]. However, it is expensive to manually annotate data sets for training. To address this challenge, we propose a method for automatic acquisition of training data.

To summarize the relationship to previous methods, our approach (1) automatically generates training instances by utilizing two characteristics of usernames; (2) models the *alias-disambiguation* step as a pairwise classification problem; (3) and explores the problem by using public UGC pages instead of the user profile pages, which are private in many online communities.

3. PROBLEM STATEMENTS

In this section, we first formally define the two steps of user linking: (a) the *alias-disambiguation* step and (b) the *alias-conflation* step (Section 3.1). In this paper, we focus on the first step, the *alias-disambiguation* step, and leave the *alias-conflation* step for future work. We demonstrate the importance of the *alias-disambiguation* step for the user linking problem by conducting a survey (Section 3.2) and an analysis on About.me (Section 3.3). The survey was completed by 153 volunteers including college students and full-time employees, who were invited by email or instant message (IM). The survey results provided us insight into human behavior with regards to the usage of online usernames. Specifically, 89.17% of the participants in our survey reported that they prefer using one main username across multiple communities. This implies that most of their online activities can be linked through their main usernames. Inspired by this observation, we further conducted an analysis on a large data set (75,472 users) crawled from About.me. The analysis results show that the *alias-disambiguation* step can address a major part of the user linking problem. Hence, the *alias-disambiguation* step is a valid starting point to solve the user linking problem.

3.1 Problem Definition

Let C denote a set of all communities, and $c \subseteq C$ denote a community. For each community c, there is a set U_c of all users who have registered in c. For a user $u \subseteq U_c$, let $id(u)$ denote the natural person to whom the user u belongs, and let $I(u)$ present all the publicly available information about the user u, including the username (denoted as $I(u).username$), the avatar (denoted as $I(u).avatar$), the location (denoted as $I(u).location$), all the posts written by him or her (denoted as $I(u).posts$), etc. The task of linking users across multiple communities can be divided into two steps: (a) the *alias-disambiguation* step and (b) the *alias-conflation* step. The formal definitions of these two steps are given as follows:

Alias-disambiguation step of user linking. The aim is to differentiate users under the same usernames. Formally, given two users $u \subseteq U_c$ and $u' \subseteq U_{c'}$ from two different communities c and c', and all the publicly available information about them $I(u)$ and $I(u')$, with the constraint that two users sharing the same username, the objective is to learn a function f to decide whether these two users belong to the same natural person. That is,

$$f(u,u') = \begin{cases} 1 & \text{if } id(u) = id(u') \\ 0 & \text{otherwise} \end{cases} \quad \text{subject to } I(u).username = I(u').username \tag{1}$$

Alias-conflation step of user linking. The goal is to find a natural person using different usernames. Formally, given two users $u \subseteq U_c$ and $u' \subseteq U'_c$ from two different communities c and c', and all the publicly available information about them $I(u)$ and $I(u')$, with the constraint that two users having different usernames, the objective is to learn a function g to decide whether these two users belong to the same natural person. That is,

$$g(u,u') = \begin{cases} 1 & \text{if } id(u) = id(u') \\ 0 & \text{otherwise} \end{cases} \quad \text{subject to } I(u).username \neq I(u').username \tag{2}$$

Table 1: The distribution of #usernames used by participants

#usernames	1	2	3	4	>4
%participants	13.04%	28.99%	27.54%	12.32%	18.12%

In this paper, for any user u, we obtain his or her publicly available information $I(u)$ from public UGC pages, rather than user profile pages that have been used in other work [29, 25, 19] since user profile pages are private in many communities. We will show this in Section 6.

3.2 Survey

We conducted a survey to explore the way in which people manage their usernames across online communities. To the best of our knowledge, it is the first study on human behavior with regards to the usage of online usernames.

Survey Content In addition to collecting the basic demographic information about participants, the survey asked a series of questions, starting with whether participants have ever participated in at least one online community. If they have done so, they were asked several follow-up questions about their frequency of using online communities and the number of usernames they use across all online communities in which they participate. Additionally, we asked participants who reported using more than one username whether they used one as a main username, and the reason they prefer to use one main username. We also asked the reason they use more than one username.

Participants This survey was completed by 153 people. All of the participants were invited via email or IM. 56.86% of the participants were male and 43.14% were female. 54.9% of the participants were aged $18 \sim 26$, 42.48% aged $27 \sim 35$ and only 1.96% aged 36 and over. 43.14% were university students and 56.86% were full-time employees. Most of the participants are heavy users of the internet, e.g. 67.97% spent approximately more than 20 hours on the internet per week on average, and 25.49% spent $5 \sim 20$ hours per week on average.

138 participants (90.20%) reported that they had participated in online communities before. 60.87% reported that they participate in online communities everyday. 15.94% participate weekly (e.g. $1 \sim 2$ times one week). 15.94% participate monthly (e.g. $1 \sim 2$ times one month). 9.42% participate rarely (e.g. $1 \sim 2$ times one year). We also collected the types of online communities in which they have participated. 69.93% have participated in social networks (e.g. microblog); 65.36% have participated in online forums; 54.25% have participated in blog community; 46.41% have participated in question and answering community.

Analysis Table 1 shows the distribution of the number of usernames used by the participants in online communities. We can observe that most participants (81.88%) used a small number of usernames $(1-4)$ across online communities. Specifically, 89.17% of the participants who used more than one username reported that they used one as their main username across multiple communities. This implies that the *alias-disambiguation* step can link most of their activities in online communities. Additionally, they prefer to use

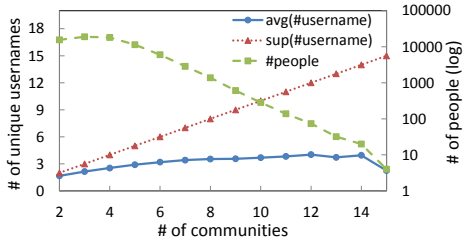

Figure 1: # of unique usernames w.r.t # of communities

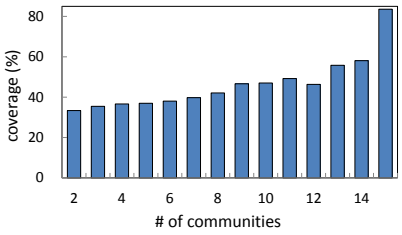

Figure 2: Coverage of true pairwise decisions in *alias-disambiguation* step

one main username across multiple communities due to two main reasons: (a) 79.53% reported that they wanted to minimize their efforts in remembering usernames across multiple communities. (b) 15.75% reported that a unique username would help them build their online reputation and make them more easily identified by other people. This observation tells us that people try to reduce the number of usernames they use. However, 86.96% of participants use more than one username (see Table 1). Two main reasons behind this phenomenon includes: (a) 58.26% reported that their preferred (main) usernames already had been used by other people at some communities, so they have to choose other usernames; (b) 21.74% reported that they would like to keep their online privacy by using different usernames in different communities to avoid de-anonymization.

3.3 Analysis on About.me

About.me is an online name card service, where people manually aggregate the links to their profile pages in other communities into one single personal page, which points to everything they do around the web. We crawled 75,472 public personal pages from About.me that show at least two involved communities and extracted all the usernames they used in those communities. There were 15 different communities in our data set (e.g. Twitter, LinkedIn and Flickr). On average, each person participated in 3.92 communities and had 2.44 usernames.

In Figure 1, people are categorized into 14 buckets according to the number of communities they participated in. The dash line (green) gives the number of people in each bucket. The dotted line (red) gives the upper bound of the number of unique usernames people can use in each bucket (i.e. the number of communities one participates in). The solid line (blue) indicates the average number of unique usernames people use in each bucket (i.e. the number of communities they participated in). From Figure 1, it can be seen that the average number of unique usernames used by one person only slightly increases as the number of the involved communities increases (the solid line). This implies people

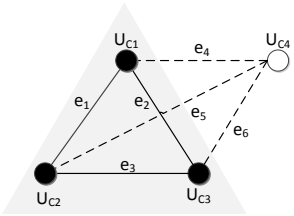

Figure 3: Example

tend to keep only a few usernames, i.e. 2 - 4, even if they participate in multiple online communities.

To quantitatively analyze how important solving the *alias-disambiguation* step is to approach the user linking problem, we conducted a further analysis of About.me. Linking all users belonging to one natural person is viewed as a number of separate pairwise decisions, where a true pairwise decision (tp) correctly links two users belonging to one natural person. Let $\sup(tp_1)$ and $\sup(tp_2)$ be the upper bound of the number of true pairwise decisions in the *alias-disambiguation* step and in the whole user linking problem, respectively. The ratio of $\sup(tp_1)$ and $\sup(tp_2)$ (called *coverage*) is the proportion of the true pairwise decisions in the user linking problem that can be covered by perfectly solving the *alias-disambiguation* step. This *coverage* is the best contribution the *alias-disambiguation* step can make to the user linking problem. It can be computed as follows:

$$coverage = \sup(tp_1)/\sup(tp_2) \quad (3)$$

Taking Figure 3 as an example, one natural person participates in four communities, c_1 (as user u_{c_1}), c_2 (as user u_{c_2}), c_3 (as user u_{c_3}) and c_4 (as user u_{c_4}). She uses one username across c_1, c_2 and c_3 (i.e. the usernames of u_{c_1}, u_{c_2} and u_{c_3} are the same), and uses another one in c_4. The user linking problem is to link the four nodes (u_{c_1}, u_{c_2}, u_{c_3} and u_{c_4}) together. Each of the 6 edges (e_1, e_2, ..., e_6) in Figure 3 indicates one true pairwise decision. In Figure 3, $\sup(tp_1)$ is 3 (i.e. 3 edges, e_1, e_2 and e_3), and $\sup(tp_2)$ is 6 (i.e. 6 edges, e_1, e_2, ..., e_6). Hence, the *coverage* of the *alias-disambiguation* step in Figure 3 is 50% (i.e. 3/6).

In Figure 2, people are categorized into 14 buckets according to the number of communities they participated in, and the bucket sizes illustrate the *coverage* value discussed above. The average *coverage* is 46.8%. Hence, it is important to solve the *alias-disambiguation* step. If the data about activity levels of users can be used, we believe that the *coverage* weighted by activity levels will be higher since 89.17% of our survey participants using more than one username reported that they use one main username in most online communities.

From Figure 2, we can observe that the *coverage* becomes higher with the increment of the number of communities people participated in. This implies that if one person participates in more communities, she prefers to spend less effort in managing her online usernames by using fewer usernames.

4. AUTOMATIC ACQUISITION OF TRAINING DATA

As we described in Section 1 and 2, we cast the *alias-disambiguation* step of user linking task as a pairwise classification problem. Supervised methods are very effective but require manually annotated training corpora. Moreover, the types of personal information are very different from site to site. It is impossible to manually label training instances for every online community. To address these challenges, we propose a novel unsupervised approach to automatically label training data for the user linking task.

The key idea of our approach is based on the following observations:

- Rare usernames are likely owned by a single natural person, e.g. **pennystar88** and **travelbag62**.
- Common usernames are likely owned by different natural persons, e.g. **tank** and **blues**.

In Section 3, we already observed that some people tend to use one main username across different communities to make it easier to remember usernames or to build their online reputation. To achieve these goals, they have to make their author names unique, for example, **pennystar88**. Hence, it is likely that two users sharing a rare username belong to a single natural person, so that the most rare usernames can be utilized for automatically labeling positive training instances. In contrast, it is unlikely that two users sharing a common username belong to a single natural person. Since there is a high probability that multiple natural persons prefer the common username. Hence, the most common usernames can be utilized for automatically labeling negative training instances. We propose using the n-gram probability of a username to estimate the rareness or commonness of a username (Section 4.1). Based on these two observations, we propose an algorithm to automatically label positive (low n-gram probability) or negative (high n-gram probability) training instances (Section 4.2).

4.1 N-gram Probabilities of Usernames

In this section, we use the n-gram probability of a username to estimate the rareness or commonness of the username, which is similar to [16]. Usually, when people select their usernames, they would like to use combinations of word sequences (one or more words) as their usernames. The word sequences may present people's real names, their birthdays, their hobbies, etc. For example, **pennystar88** might be composed of **penny**, **star** and **88** or **pen**, **ny**, **star** and **88**. The n-gram probability of a username is the n-gram probability of the word sequence of which the username consists. If the n-gram probability of a username is very low, the username is likely a rare username. If the n-gram probability of a username is very high, the user name is likely a common username.

Since it is not allowed to have spaces in usernames at many sites, we should first segment usernames and then estimate their n-gram probabilities. For example, **travelbag62** is segmented into **travel**, **bag** and **62**. We then treat the username segmentation task as a standard word breaking problem. Following the method proposed by [30], the task of username segmentation is

$$\hat{s} = \arg\max_{s \subseteq \Omega} P(s|a) = \arg\max_{s \subseteq \Omega} P(a|s)P(s) \quad (4)$$

Here, $s = (w_1, w_2, w_3, \ldots, w_{|s|})$ is a segmentation of a username a (without spaces) and the $\hat{s}$ is the objective segmentation with the maximum a posteriori, where $|s|$ denotes the number of words in the segmentation s. Let $|a|$ present the number of characters in a, and the size of the set of all possible segmentations Ω is $2^{|a|-1}$. In addition, $P(a|s)$

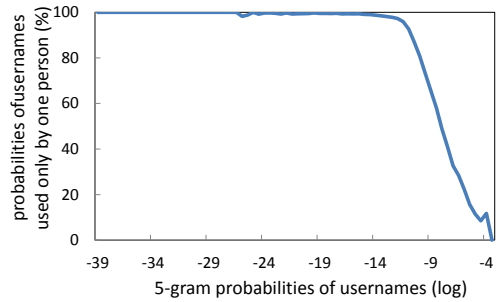

Figure 4: The distribution of username n-gram probability

and $P(s)$ are called the transformation and the segmentation prior model, respectively. In this paper, we use the word synchronous beam search approach proposed by [30] to estimate $P(a|s)$ and $P(s)$ and use the web n-gram service provided by [31] to estimate the n-gram probability. In this paper, we use the 5-gram based on the title corpus provided by the web n-gram service.

4.2 Automatic Acquisition of Training Data

Based on the two observations described above, we propose two intuitive assumptions for automatic acquisition of training data:

- **Assumption 1**: if the n-gram probability of a username is very low, it is likely that only one natural person uses it as his or her username.
- **Assumption 2**: if the n-gram probability of a username is very high, it is likely that more than one natural person uses it as a username.

We verify these two assumptions by using the data crawled from Yahoo! Answers, where different users are allowed to use the same username. Multiple users sharing the same username in Yahoo! Answers can be differentiated by their unique profile page URLs. There are 69 million question answering threads and 14 million unique users in our data set. We first counted the frequency of each unique username (i.e. the number of users using a given username) in the whole data set, and then sampled $299,716$ unique usernames used by $673,037$ unique users to verify our assumptions. In Figure 4, all the sampled usernames are categorized into buckets according to the n-gram probabilities of usernames after word breaking (estimated by the method in Section 4.1). The solid line shows the probabilities of usernames (with certain n-gram probabilities) used by only one person in Yahoo! Answers. From the solid line in Figure 4, we can observe that, (a) when the n-gram probability of a username is very low, it is likely that the username would be used by only one person; (b) when the n-gram probability of a username is very high, it is unlikely that the username would be used by one person. The experimental results verify our assumptions. Additionally, it can be observed that the solid curve in Figure 4 can be fitted by a logistic function of the n-gram probability of a username. Then, the fitted logistic function can be used for new usernames. We will show the details in Section 5.2.

Based on these two assumptions, we propose an approach to automatically label training data for user linking. Alg. 1 shows the details of our approach. The key idea of our approach is to utilize the n-gram probabilities of usernames for automatic acquisition of training instances, where each instance is a pair of users sharing the same username. We first estimate the n-gram probability of each username by using Alg. 1 (Line 2 $\sim$ 4). A given instance (i.e. one pair of users) with a username that has very low n-gram probability is assigned a true positive label, since it is likely that the username is used by only one person (Line 6 $\sim$ 8). In contrast, an instance with very a high n-gram probability username should be assigned a negative label, since it is very likely that the corresponding username is used by more than one person as their username (Line 9 $\sim$ 11).

Algorithm 1 Automatically Labeling Training Data

Input: A set S of n pairs of users sharing the same usernames from m communities
Output: A set Ω of labeled training instances

1: $\Omega \leftarrow \{\}$
2: **for each** $d = (u_c, u_{c'})$ **in** S **do** ▷ d is a pair of users sharing the same username
3: $d.username \leftarrow I(u_c).username$
4: $d.p \leftarrow$ EstimateNgramProb($d.username$) ▷ segment the username and then estimate its n-gram probability
5: $L \leftarrow$ SortByUsernameNgramProb(S) ▷ L is a list of paired users and is in ascending order
6: **for each** $d = (u_c, u_{c'})$ **in** SelectTopOnePercent(L) **do**
7: $d.label \leftarrow$ **positive** ▷ assign positive label
8: $\Omega \leftarrow \Omega \cup \{d\}$
9: **for each** $d = (u_c, u_{c'})$ **in** SelectLastOnePercent(L) **do**
10: $d.label \leftarrow$ **negative** ▷ assign negative label
11: $\Omega \leftarrow \Omega \cup \{d\}$
12: **return** Ω

5. OUR APPROACH TO LINKING USERS

We view the *alias-disambiguation* step of the user linking task as a pairwise classification problem. Given any two users (from two different communities) sharing the same username, the objective is to learn a classifier which decides whether these two users belong to one person. As a test case, we focus on linking users from different online forums. Our approach for user linking includes two main phases: automatic acquisition of training data (Section 4.2) and a standard classification procedure. In Section 5.1, we describe the features used for user linking. We introduce the classification models in Section 5.2.

5.1 Features

In this paper, we do not consider the user profile pages for feature extraction, since many communities keep user profile pages private (See details in Section 6.5). As we described in Section 3.1, we only extract features from public UGC pages (e.g. forum thread pages). Taking an online forum as a test case, there are three categories of features extracted from forum thread pages: (a) user meta data based features; (b) social relationship based features; (c) post content based features.

5.1.1 User Meta Data based Features

Usually, there is some user meta data (e.g. avatar, location) displayed on the content pages (e.g. forum thread pages).

Avatar We observed that some people would like to put the same avatars on different communities they participated in. One possible motivation is that it can help people build their online reputation and enable others to recognize them through the unique avatars. We use a standard downsampling method of digital image processing to check whether two avatars (images) are the same [11].

Location If the locations provided by two users are the same or part-of relationship, it is likely that these two users belong to one person. We use Google Map API[4] to check whether two locations are the same or part-of relationship.

Signature We observe that some users prefer to use the same or similar signature on different sites. We employ Jensen-Shannon divergence [8], which is a symmetric measure of distance between two probability distributions, to measure the difference between two signatures.

5.1.2 Social Relationship based Features

Co-Author In online communities, users have social relationships with each other via interactions. In an online forum, we define two users (from one forum) as co-authors, if they participated in one forum thread. We observed that two persons may co-author with each other in two different forums. Hence, given two users (sharing the same username) from two different forums, we extract a real value co-author feature, which counts the number of their common co-authors with the same usernames.

5.1.3 Post Content based Features

Function words Rao et al. [27] found that the usage of function words is an effective feature for authorship classification. Since function words are specific English words that are usually independent of the topic of the content, and the usage of function words correlates with the writing style of an author. We employ Jensen-Shannon divergence to measure the difference between two frequency distributions of function words.

The first n and the last n words One of our interesting observations is that some users would like to use certain words at the beginning or the ending of posts. For example, a user with a username **bob** usually used "cheers, bob ☺" as the last words in a post. Some users would like to use "hi all," when asking questions. We use Jensen-Shannon divergence to measure the difference between the frequency distributions of the first n and the last n words by two users ($n = 3$).

5.2 Classification Models

In this section, we propose two models for the pairwise classification phase. In **Model 1**, a support vector machine (SVM) model with RBF kernel [7] is trained upon the automatically labeled training data (Section 4.2), with the features introduced in the previous section. Given a set of training instances $T = \{(\mathbf{x_i}, y_i)\}$, where $\mathbf{x_i} \in \mathbb{R}^n$ is a feature vector with a label $y_i \in \{1, 0\}$. The learned model makes decisions based on the output of the function $f(\mathbf{x})$,

$$f(\mathbf{x}) = sign(\mathbf{w} \cdot \mathbf{x} + b) \tag{5}$$

where $\mathbf{x}$ is a given testing instance, and the parameters $\mathbf{w}$ and b are optimized on the training set. It is noteworthy that the username n-gram probability is not included in the feature set of **Model 1**. Since the training instances are labeled according to the n-gram probabilities, this feature can easily dominate when appearing in the feature set.

[4] `http://code.google.com/apis/maps`

Table 2: Statistics of top forums

Forum	#Threads	#Users
Disboards	$1,937K$	$169K$
Tripadvisor	$1,026K$	$522K$
Cruisecritic	$923K$	$233K$
Lonelyplanet	$699K$	$265K$
Bikeforums	$565K$	$114K$
Advrider	$312K$	$74K$

However, the n-gram probability of a username should be an effective indicator for the user linking task. We propose incorporating it in our proposed **Model 2**. Recall the prior knowledge of the username n-gram probability learned from Yahoo! Answers (Figure 4 in Section 4.2): (a) If the n-gram probability of a username is very low, it is highly likely that the username is used by only one natural person; (b) With the increment of the n-gram probability, the likelihood of the corresponding username being used by only one natural person becomes lower. As we mentioned in Section 4.2, the solid curve in Figure 4 can be fitted by a logistic function. Given the n-gram probability θ of a username, a probability $P(y = 1|\theta)$ can be estimated, which indicates how likely the username is used by only one natural person. The logistic function adopted is defined as:

$$P(y = 1|\theta) = \frac{1}{1 + e^{-\alpha(\theta - \beta)}} \tag{6}$$

where θ is the n-gram probability of the given username. Parameters α and β are learned from Yahoo! Answers ($\alpha = 0.6270$ and $\beta = -7.4212$). To utilize both the priori knowledge of the n-gram probability and the learned SVM model which makes use of other features, we design a new objective function $f'(\mathbf{x}; \theta)$ in **Model 2**:

$$f'(\mathbf{x}; \theta) = \begin{cases} 1 & \text{if } P(y = 1|\mathbf{x}; \theta) > \mu \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

where

$$P(y = 1|\mathbf{x}; \theta) = \lambda P(y = 1|\theta) + (1 - \lambda) P(y = 1|\mathbf{x})$$

Unlike in Equation (5), $P(y = 1|\mathbf{x})$ is the probabilistic output of the SVM model (i.e. **Model 1**), which can be derived from LIBSVM [7]; λ controls the contributions of the n-gram probability priori and the SVM predictions, and it is set to $P(y = 1|\theta)$ in this paper. When a username n-gram probability is low, **Model 2** relies more on the username related information $P(y = 1|\theta)$ due to the high $P(y = 1|\theta)$ value; when the n-gram probability becomes higher, **Model 2** is more dependent on the predictions of the SVM model $P(y = 1|\mathbf{x})$, which takes other features into consideration.

6. EXPERIMENTS

6.1 Data Set

We collected 53 travel related forums and crawled the public forum thread pages from these forums (until June 2011). We manually developed one template for each forum to extract the structured data (e.g. username, avatar, location, signature and post content) from the forum thread pages.

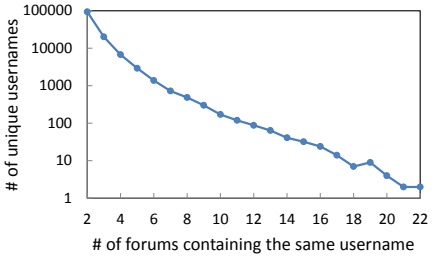

Figure 5: The frequency distribution of usernames

The data set includes 7.2 million forum threads and 1.94 million users. Table 2 shows the number of threads and the number of users in the biggest 6 sites. It should be noted that one username only can be allowed for one person to use in each of the 53 forums. There are totally 325, 139 pairs of users sharing the same usernames extracted from the data set and 127, 088 corresponding unique usernames. Our approach (Alg. 1) generated 1462 training instances (i.e. pairs of users), including 961 positive instances and 601 negative instances.

Figure 5 illustrates the frequency distribution of the same username being used in multiple forums. Each point in the plot specifies the number of unique usernames (corresponding to the y axis) which are observed in x forums (corresponding to the x axis). From the distribution it can be seen that most of the usernames that are used across forums only occur in a few forums (2 or 3), therefore resulting in sparse connections between users sharing the same usernames. Since clustering techniques are more useful for tasks involving richer and denser relationships, the clustering techniques widely used in *entity resolution* tasks [6, 5, 26, 13] are not suitable in these cases. Hence, we view this task as a pairwise classification problem.

6.2 Ground Truth

To evaluate our approaches, we hired one assessor to annotate the ground truth. The annotation of ground truth for this task is expensive. We developed a tool that helps the assessor quickly explore (a) user profile pages that have more detailed information about users and (b) users' post contents to make judgements. The information that the assessor can use to annotate data includes but is not limited to:

- The personal information about users on profile pages, e.g. avatar, personal interests, occupation, location;
- The information describing users themselves or their families in post contents or signatures, e.g. "... my family of five. Two adults, and three children ...", "... I'm planning a wedding October ...", etc;
- Similar questions posted on different forums with close dates;
- The places where (or the time when) they traveled according to their post contents or signatures;
- The URLs of blogs or homepages linked from their profile pages or signatures;
- Two people know each other, so they (frequently) interact with each other in different forums.

We randomly sampled and annotated 122 user pairs. There are 57 pairs labeled as positive instances, 44 pairs labeled as negative instances and 21 instances we cannot determine.

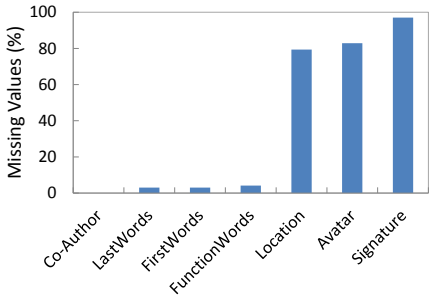

Figure 6: The percentage of missing values for each feature

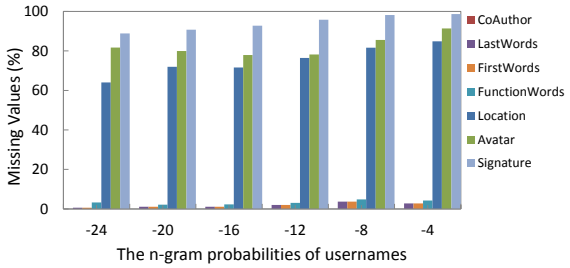

Figure 7: The percentage of missing values for each feature w.r.t n-gram probabilities of usernames

6.3 Evaluation Metrics

We employ the standard evaluation metrics in information retrieval to evaluate: accuracy (denoted as Acc), precision (denoted as P), recall (denoted as R) and F_1-measure (denoted as F_1), which are defined as follows:

$$Acc = \frac{tp + tn}{tp + tn + fp + fn}$$

$$P = \frac{tp}{tp + fp}, R = \frac{tp}{tp + fn}, F_1 = \frac{2PR}{P + R}$$

where tp indicates the number of true positives, tn indicates the number of true negatives, fn indicates the number of false negatives and fp indicates the number of false positives.

6.4 Missing Values of Features

We examine the problem of missing values of features in the dataset. Figure 6 shows the percentage of missing values for each feature in our dataset. It can be observed that some of the features (e.g. location, avatar and signature) have large proportion of missing values. Actually, a similar problem has been reported in [19]. However, Malhotra et al. [19] avoid this problem by selecting the communities where there is no large proportion of missing values.

The presence of missing values in a dataset can affect the performance of a classifier trained on the dataset containing missing values [2]. Several methods have been proposed to treat missing data. Acuna et al. [2] carried out experiments and found that case deletion (CD) is the most effective method for a dataset where there is a large proportion of missing values. This method discards all instances (cases) with missing values for at least n features (we use $n = 2$).

Table 3: Overall results

Method	Prec.	Rec.	F_1	Acc.
Baseline	0.5644	1.0000	0.7215	0.5644
Model 1	0.8571	0.6316	0.7273	0.7327
Model 2 ($\mu = 0.8$)	0.9455	0.9123	0.9286	0.9208
Model 2 ($\mu = 0.7$)	0.9000	0.9474	0.9231	0.9109
Model 2 ($\mu = 0.6$)	0.8636	1.0000	0.9268	0.9109
Model 2 ($\mu = 0.5$)	0.8382	1.0000	0.9120	0.8911

Table 4: Experimental results on two subsets

Method	Accuracy on Set1 (size=59)	Accuracy on Set2 (size=42)
Baseline	0.6441	0.1220
Model 1	0.8571	0.6441
Model 2 ($\mu = 0.8$)	0.9322	0.9048

We further examine the missing values problem over different n-gram probabilities of usernames. Figure 7 shows the percentage of missing values for each feature according to different n-gram probabilities of usernames. We can observe that the proportion of missing values for each feature becomes higher with the increment of the n-gram probability of a username. Recall the results of our survey (Section 3.2), 15.75% participants reported that a unique username would help them build their online reputation and make them easily identified by other people. Similarly, one possible reason behind the above phenomenon (Figure 7) is that the person a using unique username (low n-gram probability) prefers to maintain their online reputation via a unique avatar, signature etc. This would result in the different performances of our classifier on the instances with different n-gram probabilities. We will provide detailed results in Section 6.5.

6.5 Experimental Results

We first examine the possibility of applying the previous approaches to this task. Most previous methods [29, 25, 19] are heavily dependent on the accessability of user profile pages. However, user profile pages are private at many online communities. Among the 53 forums in our data set, 45.28% of forums have user profile pages that are private. Zafarani et al. [33] proposed a web search based approach, which is dependent on: (a) the URL of a user profile contains the corresponding username; (b) the user profile pages of usernames should be publicly available. However, only 24.53% of forums in our data set meet these two conditions.

We then empirically evaluated three methods. The simplest baseline is to treat all pairs of users sharing the same usernames as positive instances. We also examined our proposed **Model 1** and **Model 2**.

Table 3 shows the experimental results on the whole set of ground truth. The baseline of treating two identical usernames always belonging to a single natural person only achieves an accuracy of 56.44%. This tells us that the *alias-disambiguation* step is not a trivial task. Both **Model 1** and **Model 2** outperform the baseline. **Model 1** gives an accuracy of 73.27%. It shows that our classifier on the automatically training data is reasonably effective. **Model 2**, which additionally utilizes the priori knowledge of usernames learned from Yahoo! Answers, gives the best performance. It shows that the rareness or commonness of usernames can help user linking.

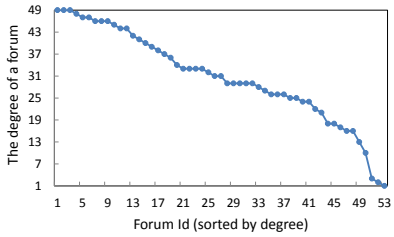

Figure 8: Degree distribution of the forum graph

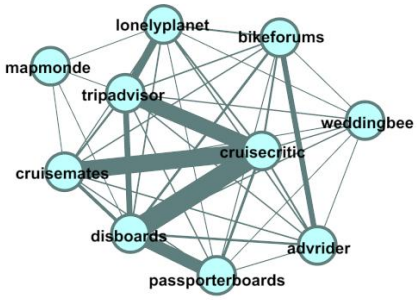

Figure 9: The graph of top 10 forums

Additionally, we found that **Model 2** gives higher recall and lower precision with the increment of the threshold μ.

We further investigated the cases where **Model 1** performs poorly. We separated the ground truth into two subsets: (a) the set of instances with n-gram probabilities lower than a threshold θ (named Set 1); (b) the set of instances with the n-gram probabilities higher than the threshold θ (named Set 2). In this paper, we set the threshold θ to be -7.42, since we learned from Yahoo! Answers (Figure 4) that when the n-gram probability of a username is higher than -7.42, the probability of the username used by only one person is lower than 50.0%. Table 4 shows the performances of the three methods on the two subsets. From Table 4, we can see that **Model 1** performs much better on Set 1 than on Set 2. As we described in Section 6.4, this is due to the fact that there is a larger proportion of missing values for each feature in Set 2 than in Set 1.

6.6 Analysis on the Graph of Forums

In this section, we examine the effectiveness of our proposed method (**Model 2**) by analyzing the graph of forums. A graph of forums can be constructed by using the linked user pairs: if there is one pair of users from two forums linked together, there is an undirected edge between these two forums (nodes). The weight of each edge is the number of linked user pairs between the corresponding two forums.

We first apply our proposed method (**Model 2**) on the 53 travel related forums. There are 68,349 user pairs (21.02%) linked according to the classification results of our method. Then, the graph of forums can be constructed. Figure 8 shows the degree distribution of the constructed forum graph. From Figure 8, we can observe that all 53 forums are connected to at least one other forum. Figure 9 shows the subgraph of the 10 biggest forums. The width of the edge is proportional to the weight of the edge. It is expected that two forums with similar topics share common users. Hence, the width of the edge between two forums with similar top-

ics should be wide. We have several interesting observations from Figure 9:

- There is a strong connection between Cruisecritic and Cruisemates, since the main discussion topics on both forums are cruise;
- There is a strong connection between Bikeforums and Advrider, since the topic of these two forums is cycling;
- There is a strong connection between Disboards and Passporterboards, since both of them focus on Disney World;
- The strong connection between Tripadvisor and Lonelyplanet is due to their similar board hierarchies for discussing world travel.

7. CONCLUSIONS

In this paper, we focus on the *alias-disambiguation* step of the user linking task. We quantitatively analyze the importance of this step by conducting a survey and an experiment. The analysis shows that the *alias-disambiguation* solution can address a major part of the user linking problem in terms of the coverage of true pairwise decisions (46.8%). Morever, 89.17% of participants in our survey reported that they prefer to use one main username across multiple communities. To the best of our knowledge, it is the first study on the human behavior of the usage of online usernames. We then propose an unsupervised approach for user linking, which utilizes the rareness and commonness of usernames measured by their n-gram probabilities to automatically label training instances. The empirical evaluation verifies the effectiveness of the classifiers with the automatically generated training data. It also shows that the rareness and commonness of usernames can help user linking. We further analyze the cases where our classifiers would fail.

Future work may follow two paths: (a) investigate *alias-conflation* step by using the SVM model (**Model 1**) that is learned from the *alias-disambiguation* step and does not rely on usernames; (b) employ blocking techniques with pairwise classification models to efficiently explore this problem on a larger scale data set.

8. ACKNOWLEDGEMENT

The authors thank Jie Cai for her helpful discussions and supports on this work.

9. REFERENCES

[1] F. Abel, N. Henze, E. Herder, and D. Krause. Interweaving public user profiles on the web. In *Proceedings of UMAP*, 2010.

[2] E. Acuna and C. Rodriguez. The treatment of missing values and its effect on classifier accuracy. *Classification, Clustering, and Data Mining Applications*, 2004.

[3] L. Backstrom, C. Dwork, and J. Kleinberg. Wherefore art thou r3579x?: anonymized social networks, hidden patterns, and structural steganography. In *Proceedings of WWW*, 2007.

[4] E. Bengtson and D. Roth. Understanding the value of features for coreference resolution. In *Proceedings of EMNLP*, 2008.

[5] I. Bhattacharya and L. Getoor. Collective entity resolution in relational data. *ACM Transactions on Knowledge Discovery from Data (TKDD)*, 1(1), 2007.

[6] J. Cai and M. Strube. End-to-end coreference resolution via hypergraph partitioning. In *Proceedings of COLING*, 2010.

[7] C. Chang and C. Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3), 2011.

[8] T. Cover, J. Thomas, J. Wiley, et al. *Elements of information theory*, volume 6. 1991.

[9] A. Elmagarmid, P. Ipeirotis, and V. Verykios. Duplicate record detection: A survey. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 19(1), 2007.

[10] D. Frankowski, D. Cosley, S. Sen, L. Terveen, and J. Riedl. You are what you say: privacy risks of public mentions. In *Proceedings of SIGIR*, 2006.

[11] R. Gonzalez and R. Woods. *Digital image processing*. 2002.

[12] T. Iofciu, P. Fankhauser, F. Abel, and K. Bischoff. Identifying users across social tagging systems. In *Proceedings of ICWSM*, 2011.

[13] D. Kalashnikov, Z. Chen, S. Mehrotra, and R. Nuray-Turan. Web people search via connection analysis. *IEEE Transactions on Knowledge and Data Engineering (TKDE)*, 20(11), 2008.

[14] S. Kumar, R. Zafarani, and H. Liu. Understanding user migration patterns in social media. In *Proceedings of AAAI*, 2011.

[15] S. Labitzke, I. Taranu, and H. Hartenstein. What your friends tell others about you: Low cost linkability of social network profiles. In *Proceedings of SNAKDD*, 2011.

[16] J. Liu, X. Song, J. Jiang, and C. Lin. An unsupervised method for author extraction from web pages containing user-generated content. In *Proceedings of CIKM*, 2012.

[17] J. Liu, Y. Song, and C. Lin. Competition-based user expertise score estimation. In *Proceedings of SIGIR*, 2011.

[18] K. Liu and E. Terzi. A framework for computing the privacy scores of users in online social networks. In *Proceedings of ICDM*, 2009.

[19] A. Malhotra, L. Totti, W. Meira, P. Kumaraguru, and V. Almeida. Studying user footprints in different online social networks. In *Proceedings of CSOSN*, 2012.

[20] A. Narayanan and V. Shmatikov. Robust de-anonymization of large sparse datasets. In *Proceedings of S&P*, 2008.

[21] A. Narayanan and V. Shmatikov. De-anonymizing social networks. In *Proceedings of S&P*, 2009.

[22] A. Narayanan and V. Shmatikov. Myths and fallacies of personally identifiable information. *Communications of the ACM*, 53(6), 2010.

[23] M. Newman. Communities, modules and large-scale structure in networks. *Nature Physics*, 8(1), 2011.

[24] J. Novak, P. Raghavan, and A. Tomkins. Anti-aliasing on the web. In *Proceedings of WWW*, 2004.

[25] A. Nunes, P. Calado, and B. Martins. Resolving user identities over social networks through supervised learning and rich similarity features. In *Proceedings of SAC*, 2012.

[26] Y. Qian, Y. Hu, J. Cui, Q. Zheng, and Z. Nie. Combining machine learning and human judgment in author disambiguation. In *Proceedings of CIKM*, 2011.

[27] J. Rao and P. Rohatgi. Can pseudonymity really guarantee privacy. In *Proceedings of USENIX*, 2000.

[28] W. Soon, H. Ng, and D. Lim. A machine learning approach to coreference resolution of noun phrases. *Computational Linguistics*, 27(4), 2001.

[29] J. Vosecky, D. Hong, and V. Shen. User identification across multiple social networks. In *Proceedings of NDT*, 2009.

[30] K. Wang, C. Thrasher, and B. Hsu. Web scale nlp: A case study on url word breaking. In *Proceedings of WWW*, 2011.

[31] K. Wang, C. Thrasher, E. Viegas, X. Li, and B. Hsu. An overview of microsoft web n-gram corpus and applications. In *Proceedings of NAACL*, 2010.

[32] J. Weng, E.-P. Lim, J. Jiang, and Q. He. Twitterrank: finding topic-sensitive influential twitterers. In *Proceedings of WSDM*, 2010.

[33] R. Zafarani and H. Liu. Connecting corresponding identities across communities. In *Proceedings of ICWSM*, 2009.