A Challenge of Authorship Identification for Ten-thousand-scale Microblog Users

Syunya Okuno  
Dept. of Computer Sci. and Eng.  
Waseda Univ.  
Tokyo, Japan  
o_syunya@yama.info.waseda.ac.jp

Hiroki Asai  
Dept. of Computer Sci. and Eng.  
Waseda Univ.  
Tokyo, Japan  
asai@yama.info.waseda.ac.jp

Hayato Yamana  
Faculty of Sci. and Eng., Waseda Univ.  
Tokyo, Japan  
yamana@yama.info.waseda.ac.jp

Abstract— Internet security issues require authorship identification for all kinds of internet contents; however, authorship identification for microblog users is much harder than other documents because microblog texts are too short. Moreover, when the number of candidates becomes large, i.e., big data, it will take long time to identify. Our proposed method solves these problems. The experimental results show that our method successfully identifies the authorship with 53.2% of precision out of 10,000 microblog users in the almost half execution time of previous method.

Keywords- authorship identification; authorship detection; authorship attribution; microblog; Twitter;

I. INTRODUCTION

In recent years, many user-created posts such as twitter posts are used in criminal investigations; however, most of them are anonymous besides some of them disguise an originator, which requires authorship identification.

Although there exist authorship identification researches for internet contents; however, most of them target blog texts that are well written in comparison with microblog texts. Narayanan et al. [1] achieved low precision@1 as 20% with 100,000 blog users. On the other hand, Ragel et al. [2] targeted short message service (SMS); however, they achieved low precision@1, less than 25%, with only 70 SMS users. Moreover, to the best of our knowledge, none of previous papers has focused on the execution time of authorship identification.

When we target microblog users’ texts, we must consider the characteristics of them, i.e., short texts containing wide-ranging topics, to keep high precision@1. Also we must shorten the execution time to handle large number of microblog users. In this paper, we propose a new method to identify authorship of microblog texts. Our contributions are 1) a combined selection technique for training dataset to handle wide ranging topics, 2) a biased weighting technique for n-gram to control the weight of n-gram features to handle short texts, and 3) POS-tag-combined-n-gram to shorten the execution time of identification by decreasing the number of features.

II. RELATED WORK

Originally, authorship identification has been studied in determining authorship of written text from a few candidate authors such as validating whether a given novel was written by Shakespeare among a few candidate set of contemporary novel writers inspired by Shakespeare.


Narayanan et al. [1] did the authorship identification for large scale blogs, i.e., 100,000 blog users; however, achieved low precision@1 as 20%. They adopt machine learning techniques to conclude that the combination of “nearest neighbor” and “regularized least squares classification” works well after comparison of many classifiers. Koppel et al. [3] also targeted blogs. They used 1,000 blog users to identify authorship by employing word frequency as features followed by calculating cosine-similarity. Since precision@1 is low as 35% by using 500 words for each blog, they propose how to calculate the confidence of each result to increase precision but decrease recall by applying support vector machine (SVM) to the results. Since blogs are well written in comparison with microblog texts, we must think about more sophisticate method to achieve high precision.

Ragel et al. [2] targeted SMS to identify authorship by adopting unigrams as features followed by calculating cosine-similarity; however, they achieved low precision@1, less than 25%, with only 70 SMS users. It shows that when targeting SMS, it becomes hard to achieve high precision even with small number of SMS users. Silva et al. [4] targeted tweets; however, they showed precision and recall only by employing 3 users to identify. They used 40 sets of tweets each of which includes 3 users. Here, 75 to 2000 tweets were used for each user. F-value, i.e., harmonic mean of precision and recall, is 0.54 when using 75 tweets and 0.73 when using 2,000 tweets. It shows that the difficulty of targeting microblog users to identify authorship.

As shown above, when targeting microblog users, authorship identification becomes much harder than targeting blog users, because a tweet is just a string of up to 140 characters and a set of tweets includes wide ranging topics. Moreover, to the best of our knowledge, none of previous papers has focused on the execution time of authorship identification; however, we must decrease the execution time when handling huge amount of users to identify.
III. METHODOLOGY

In this work, we aim to identify authorship from given examples of tweets written by a set of authors. Figure 1 shows the overall view of our proposed method.

As for the baseline method, we adopt a combination of character-n-gram frequency, where “n” is 1, 2 and 3, as features. Then we calculate the similarity between a target user and candidate users by using cosine-similarity followed by ranking candidate users based on the similarity. This method is commonly used as stylometric features.

Then, we extend the baseline method to adopt three new techniques.

A. Combined selection technique for training dataset

First, for the purpose of robustness to topics, i.e., handling wide ranging topics, we propose combined selection technique for training datasets.

Previous method [2] varies the number of tweets for training and testing, from 50 to 400 tweets for each, to confirm how the precision of authorship identification changes; however, it does not consider the time difference between training dataset and test dataset. Tweets tend to include various topics and the topics tend to change often even if the tweets are posted by a single user. Thus, if we prepare different datasets for training, i.e., a set of datasets whose posted times are different, the precision will vary depending on which dataset is used for training.

In our method, we prepare three kinds of training dataset called \(T_{\text{recent}}\), \(T_{\text{week}}\), and \(T_{\text{month}}\) as shown in Figure 2. \(T_{\text{recent}}\) consists of 30 tweets posted immediately before the test dataset called \(T_{\text{test}}\). \(T_{\text{week}}\) consists of another 30 tweets posted immediately one week before \(T_{\text{test}}\). \(T_{\text{month}}\) consists of another 30 tweets posted immediately one month before \(T_{\text{test}}\).

Our proposed combined selection technique for training dataset automatically selects the best matched training dataset for a given test dataset. We calculate three kinds of similarity with a given test dataset, i.e. cosine-similarity between \(T_{\text{recent}}\) and \(T_{\text{test}}\), that between \(T_{\text{week}}\) and \(T_{\text{test}}\), and that between \(T_{\text{month}}\) and \(T_{\text{test}}\), followed by selecting maximum similarity score among the three similarity scores. By adopting the technique, we may handle wide ranging topics.

Here, it is an open question how we should prepare these different datasets to increase precision@1 more than the above preparing technique such as \(T_{\text{recent}}\), \(T_{\text{week}}\), and \(T_{\text{month}}\).

B. Biased weighting technique for n-gram

Second, a biased weighting technique for n-gram is proposed to handle short texts. Since short texts usually include small number of features, we strengthen the weight of n-gram features in proportion to the number of “n”. Here, “n” represents “n” of n-gram. Specifically, we multiply “n” to each features of n-gram frequency.

C. POS-tag-combined-n-gram

Third, we propose POS-tag-combined-n-gram to shorten the execution time by decreasing the number of the features. We convert the morpheme related to a specific topic, into a part-of-speech (POS) tag before calculating the n-gram frequency. We assume that such morphemes have small

---

**Figure 1.** Overview of our proposed method

**Figure 2.** Three train-datasets and a test-dataset
priority to identify authorship, because a user usually tweets various kinds of topics. Based on our pre-investigation, morphemes except verb, adverb, adjective, sign, unknown word, interjection, or conjunction are converted. For example, “school” whose POS is noun is converted into [noun] instead of using “school” for n-gram.

IV. EXPERIMENT & RESULT

We collected 2,000 tweets per user from over 8 million Japanese users by using Twitter API during Jan. 2013 to Dec. 2013. Before the experiment, we preprocessed the collected tweets: 1) discarding mention tags such as “@username,” 2) discarding hashtags such as “#hashtag,” 3) discarding retweets, and 4) discarding all the users who are bots as possible as we can extract by checking their tweet-client-name.

Experiments were performed to randomly select 10,000 Japanese tweet users, who had enough number of tweets, from preprocessed dataset. Based on Figure 1, the newest 30 tweets were used for test dataset, and other 90 tweets were used for training datasets. Table 1 shows results of our experiment. Precision@1 shows the percentage of the correct result, i.e. 1st ranked estimated candidate user is correct. We executed the authorship identification by changing the target user followed by calculating how many authorship is correctly identified out of 10,000 users.

A. Combined selection technique for training dataset

First of all, the difference among baselines using T_recent, T_week, and T_month is obvious. Since the topics of tweets tend to change day by day, we could achieve higher precision@1 when we used T_recent as training dataset.

Moreover, our proposed combined selection technique for training dataset, indicated as a1) in Table 1, works well in comparison with baselines b1), b2), and b3). It shows the effectiveness of automatic selection of training dataset.

B. Biased weighting technique for n-gram

Comparing the proposed method with a3), proposed method achieves higher precision@1 than a3), where w/o biased weighting but w/ POS-tag-combined-n-gram. Moreover, comparing a2) and a1), a2) achieves higher precision@1 than a1), where w/o biased weighting. These results show that we have successfully strengthen important features by adopting biased weighting technique for n-gram, regardless of adopting POS-tag-combined-n-gram.

C. POS-tag-combined-n-gram

Comparing both proposed method and a3) with both a2) and a1), both proposed method and a3) that adopts POS-tag-combined-n-gram executes authorship identification faster than other two methods that are equipped with no POS-tag-combined-n-gram. These results show that we have successfully shorten the execution time, i.e., almost half, by adopting POS-tag-combined-n-gram with small degradation of precision@1.

V. CONCLUSION

In this paper, we proposed a new method to identify the authorship out of ten-thousand scale microblog users. For the purpose of automatic authorship identification for short texts, we employ both combined selection technique for training dataset and biased weighting technique for n-gram. Moreover, to shorten the execution time, we employ POS-tag-combined-n-gram to decrease the number of features.

Out results show that in 53.2% of cases, we can correctly identify an anonymous author in the almost half execution time of previous method.

Further research will target 100,000 Japanese tweet users besides tuning of parameters to achieve higher precision@1. Also, we will experiment with English tweet users.

ACKNOWLEDGMENT

This research has been supported by Grants-in-Aid for Scientific Research (B) (25280113) from JSPS.

REFERENCES


