Detection of Users Suspected of Pretending to Be Other Users in a Community Site by Using Messages Submitted to Non-Target Categories *

Naoki Ishikawa[a], Ryo Nishimura[a], Yasuhiko Watanabe[a], Masaki Murata[b], and Yoshihiro Okada[a]

[a]Ryukoku University, Dep. of Media Informatics, Seta, Otsu, Shiga, 520-2194, Japan
t10m096@mail.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp,
{watanabe, okada}@rins.ryukoku.ac.jp
[b]Tottori University, Koyama-Minami, Tottori, 680-8552, Japan
murata@ike.tottori-u.ac.jp

Abstract. Some users abuse the anonymity and disrupt communications in a community site. Authorship identification based on analyzing stylistic features of messages is effective in detecting these inadequate users. However, in this method, the scope of target users was often limited because the criteria for selecting learning examples were strict. To relax the criteria and extend the scope of target users, we propose a method of detecting users suspected of pretending to be other users in a certain category of a community site by using their messages submitted to other (non-target) categories. Also, we show the accuracy of user identification when we relax the criteria of selecting learning examples and extend the scope of target users.

Keywords: user identification, extension of scope of target users, messages in other categories, stylistic feature, community site

1 Introduction

In these days, many people use community sites, such as Q&A sites and social network services, where users share their information and knowledge. One of the essential factors in community sites is anonymous submission. It is important to submit messages anonymously to a community site. This is because anonymity gives users chances to submit messages without regard to shame and reputation. However, some users abuse the anonymity and disrupt communications in a community site. For example, some users pretend to be other users by using multiple user accounts and attempt to manipulate communications in the community site. Manipulated communications discourage other submitters, keep users from retrieving good communication records, and decrease the credibility of the community site. As a result, it is important to detect users suspected of pretending to be other users to manipulate communications in a community site. In this case, identity tracing based on user accounts is not effective because these suspicious users often attempt to hide their true identity to avoid detection. A possible solution is authorship identification based on analyzing stylistic features of messages (Craig, 1999) (de Vel *et al.*, 2001) (Koppel *et al.*, 2002) (Argamon *et al.*, 2003) (Zheng *et al.*, 2006). However, most of previous studies selected their learning examples carefully and, consequently, limited the scope of target users. For example, if target users are set to submitters in a certain category of a community site, learning examples are generally selected from their messages submitted to the category, not other categories. Actually, in

* This research has been supported partly by the Grant-in-Aid for Scientific Research under Grant No.20500106 and 22650021.

(Ishikawa *et al.*, 2010), we developed learning examples by using target user's messages submitted to the target category. However, the scope of target users was limited because there is a limited number of users who submitted more messages to the target category than the minimum needed to develop their learning examples. In order to extend the scope of target users, it is necessary to relax the criteria for selecting learning examples. The point is extension's effects on the accuracy of user identification. In other words, when we use learning examples which were not used in cases of previous studies and extend the scope of target users,

- how much the accuracy of identification may decrease, or
- how many learning examples should be increased to maintain the accuracy.

In this study, we propose a method of detecting users suspected of pretending to be other users in a certain category of a community site by using their answers submitted to other (non-target) categories and show the accuracy of user identification when we relax the criteria of selecting learning examples and extend the scope of target users. In this method, we use user identifiers which are developed by learning stylistic features of submitter's messages and determine by whom a series of input messages are submitted.

2 Submission Types of Spoofing Users

In this study, we used messages in the data of Yahoo! chiebukuro [1], a widely-used Japanese Q&A site, for observation, data training, and examination. The data of Yahoo! chiebukuro was published by Yahoo! JAPAN via National Institute of Informatics in 2007 [2]. This data consists of about 3.11 million questions and 13.47 million answers which were posted on Yahoo! chiebukuro from April/2004 to October/2005.

In Yahoo! chiebukuro, users need not reveal their real names to submit their messages. However, their messages are traceable because their user accounts are attached to them. Because of this traceability, we can collect any user's messages and some of them include clues of identifying individuals. As a result, to avoid identifying individuals, it is reasonable and proper that users change their user accounts or use multiple user accounts. However, the following types of message submissions using multiple user accounts are neither reasonable nor proper.

TYPE I one user submits a question and its answer by using multiple user accounts (Figure 1 (a)).

We think that the user intended to manipulate the message evaluation. For example, in Yahoo! chiebukuro, each questioner is requested to determine which answer is best and give a *best answer* label to it. These message evaluations encourage message submitters to submit new messages and increase the credibility of the community site. We think, the user aimed to get best answer labels and repeated this type of submissions.

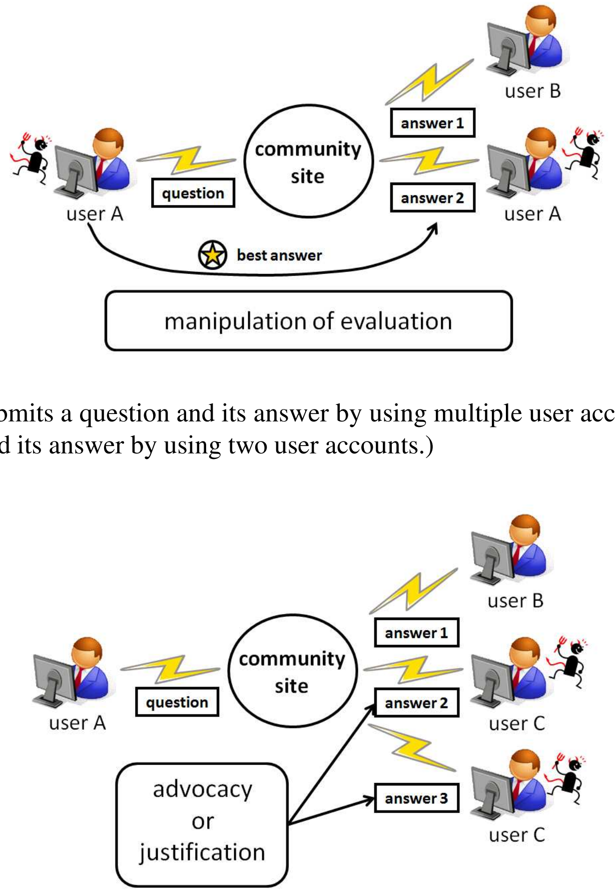
TYPE II one user submits two or more answers to the same question by using multiple user accounts (Figure 1 (b)).

We think that the user intended to dominate or disrupt communications in the community site. To be more precise, the user intended to

- control communications by advocating or justifying his/her opinions, or
- disrupt communications by submitting two or more inappropriate messages.

[1] http://chiebukuro.yahoo.co.jp

[2] http://research.nii.ac.jp/tdc/chiebukuro.html



(a) TYPE I: one user submits a question and its answer by using multiple user accounts. (In this case, user A submits a question and its answer by using two user accounts.)

(b) TYPE II: one user submits two or more answers to the same question by using multiple user accounts. (In this case, user C submits two answers by using two user accounts.)

Figure 1: Two types of abnormal submissions: TYPE I and TYPE II.

These kinds of submissions discourage other submitters, keep users from retrieving good communication records, and decrease the credibility of the community site. As a result, it is important to detect users suspected of pretending to be someone else to manipulate communications in a community site.

TYPE I submissions are sometimes obscurer than TYPE II submissions because the standards of best answer selection differ with each questioner. In other words, it is more possible to disrupt communications by TYPE II submissions than TYPE I. As a result, in this study, we intend to investigate a method of detecting users who have repeated TYPE II submissions.

3 Detection of Submitters Suspected of Pretending to Be Someone Else

In order to detect users who repeated TYPE II submissions, we intend to detect users who

- have similar styles of writing, and
- submitted answers to the same questions.

It is easy to detect users who submitted answers to the same questions by using their submission records. As a result, in this section, we explain a method of detecting users who have similar styles

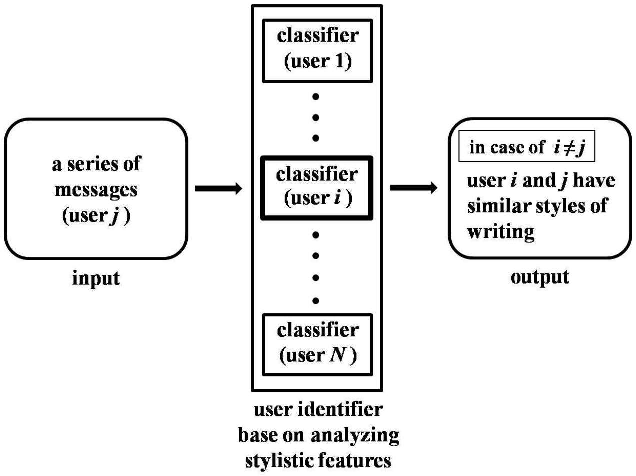

classifier (user 1)
classifier (user i)
classifier (user N)
a series of messages (user j)
input
in case of i≠j
user i and j have similar styles of writing
output
user identifier base on analyzing stylistic features

Figure 2: The outline of our method of detecting users who have similar styles of writing

of writing. Figure 2 shows the outline of our method of detecting users who have similar styles of writing.

In our method, we use an user identifier which is based on analyzing stylistic features and determines by whom a series of input messages are submitted. As shown in Figure 2, the user identifier consists of N classifiers developed by learning users' stylistic features. Each classifier has a target user and calculates the probability that a series of input messages were submitted by the target user. Then, the user identifier determines that a series of input messages were submitted by the user with the highest probability. When the user with the highest probability differs from the user submitted a series of input messages, our method determines that these users have similar styles of writing. In this case, our method determines that user i and j have similar styles of writing. In this way, the key to detecting users of similar writing styles is the user classifiers. As a result, we explain below how to develop these user classifiers.

Suppose that user i, ranked i-th place in the ranking of frequent answer submitters in a target category of a community site, submitted l answers which meet the criteria of selecting learning examples, and is the target user of classifier (user i). When a series of m answers, which were submitted by user j to the target category, are given to classifier (user i), probability score $score(i,j)$, which means that user i and j were one and the same user and user i submitted the series of m answers, is calculated as follows:

$$score(i,j) = \begin{cases} \displaystyle\prod_{k=1}^{m} P_{ijk} & \text{(in case of } \displaystyle\prod_{k=1}^{m} P_{ijk} > \prod_{k=1}^{m}(1-P_{ijk})) \\ 0 & \text{(in case of } \displaystyle\prod_{k=1}^{m} P_{ijk} \le \prod_{k=1}^{m}(1-P_{ijk})) \end{cases}$$

where P_{ijk} is the probability that user i submitted answers k ($1 \le k \le m$) in the series of m answers of user j. P_{ijk} is calculated by classifier (user i), which was developed by learning stylistic features of user i. We use two types of training data for learning stylistic features of user i:

training data (target category) of user i consists of

- n answers which were selected randomly from l answers submitted by user i to the

$s1$	the results of morphological analysis on sentences in the target message
$s2$	the results of morphological analysis on the sentence and sentence No.
$s3$	character 3-gram extracted from sentences in the target message
$s4$	character 3-gram extracted from the sentence and its sentence No.
$s5$	1 $\sim$ 10 characters at the head of each sentence
$s6$	1 $\sim$ 10 characters at the end of each sentence
$s7$	sequential patterns extracted by PrefixSpan (frequency is 5+, item number is 3+, maximum gap number is 1, and maximum gap length is 1)

Figure 3: Features used in maximum entropy (ME) method for learning stylistic features of submitters. PrefixSpan (http://prefixspan-rel.sourceforge.jp/) is a method of mining sequential patterns efficiently.

target category and

- n answers which are extracted randomly from answers submitted by other users to the target category.

training data (non-target category) of user i consists of

- n answers which were selected randomly from l answers submitted by user i to non-target categories and
- n answers which are extracted randomly from answers submitted by other users to non-target category.

This is because we intend to measure and compare the accuracy of user identifiers (Figure 2) developed by using these two kinds of training data. In this study, we use the maximum entropy (ME) method for data training. Figure 3 shows feature $s1 \sim s7$ used in machine learning on experimental data.

4 Experimental Results

To measure and compare the accuracy of user identification, we developed user identifiers for the following user groups:

group A 100 users who submitted over 200 answers to the target category and over 150 answers to non-target categories

group B 100 users who submitted 50 $\sim$ 199 answers to the target category and over 150 answers to non-target categories

In the experiments, the target category was set to social issues category of Yahoo! chiebukuro. In the social issues category, 78777 questions and 403306 answers were submitted by 13256 questioner and 25766 answer submitters, respectively.

We developed three kinds of training data,

- training data (target category) of group A
- training data (non-target categories) of group A
- training data (non-target categories) of group B

and examination data in the next way. First, we extracted 50, 100, and 150 answers from each user's answers submitted to non-target categories and, as mentioned in section 3, developed three different sizes (100, 200, and 300 answers) of training data (non-target categories) of group A and B. Secondly, in order to develop examination data of group A and B, we extracted 50 answers from each user's answers submitted to the target category. Finally, from the other answers of each user in group A, we extracted 50, 100, and 150 answers and developed three different sizes (100, 200, and 300 answers) of training data (target category) of group A.

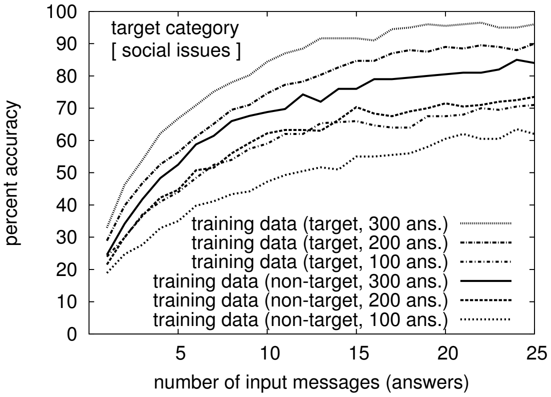


Figure 4: The accuracy of the identifiers developed by using various sizes of training data (target category and non-target categories) of group A.

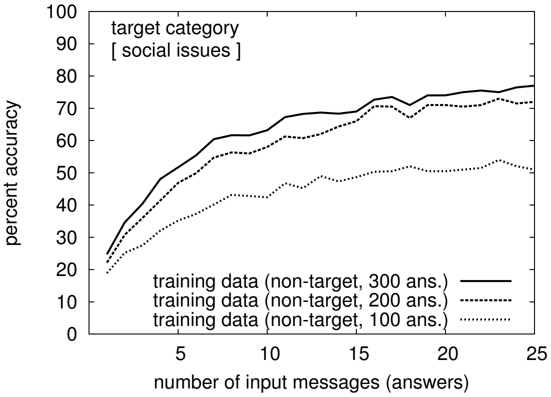


Figure 5: The accuracy of the identifiers developed by using various sizes of training data (non-target categories) of group B.

Next, we developed user classifiers and identifiers by applying maximum entropy (ME) method to the training data. In the experiments, we used a package for maximum entropy method, maxent [3], for data training. We also used a Japanese morphological analyzer, Mecab[4], for word segmentation of messages. Then, we varied the numbers of input messages to the classifiers and measured the accuracy of the identifiers. Input messages were extracted from the examination data. Figure 4 shows the accuracy of the identifier under the various numbers (1 $\sim$ 25) of input messages and the various sizes (100, 200, and 300 answers) of training data (target category and non-target categories) of group A. As shown in Figure 4, the identifier developed by using training data (target category, 100 answers) had about the same accuracy as the identifier developed by using training data (non-target categories, 200 answers). In other words, in this case, training data (non-target categories, 200 answers) can replace training data (target category, 100 answers). Furthermore, we found feature $s6$ in Figure 3 (1 $\sim$ 10 characters at the end of sentences) is effective to this experiment.

[3] http://mastarpj.nict.go.jp/ mutiyama/software/maxent

[4] http://mecab.sourceforge.net/

Table 1: The numbers of user pairs who have similar styles of writing and submitted answers to the same questions in the target (social issues) category.

	frequency of submissions to the same questions	
target users	one or more	ten or more
over 200 (312 users)	109	22
50 $\sim$ 199 (808 users)	451	25

Figure 5 shows the accuracy of the identifiers developed by using the training data (non-target categories) of group B. In social issues category of Yahoo! chiebukuro, the target users were only 312 users when we developed training data (300 answers) and examination data (50 answers) only by using answers submitted to the target category. On the other hand, there were another 808 users when we relaxed the criteria of selecting learning examples and used answers submitted to non-target categories.

Finally, we tried to detect users who have similar styles of writing and submitted answers to the same questions. The target users are

- 312 users who submitted over 200 answers to the target category in Yahoo! chiebukuro, and
- 808 users who submitted 50 $\sim$ 199 answers to the target category and over 150 answers to non-target categories in Yahoo! chiebukuro.

In the former case, we used training data (target category, 300 answers) for developing the identifier. In the latter case, we used training data (non-target categories, 300 answers). In both cases, we set the number of input messages to be 16. As a result, the accuracy of these identifiers were 91 and 73 %, respectively. Table 1 shows the numbers of user pairs who have similar styles of writing and submitted answers to the same questions. In this experiment, from 22 and 24 user pairs submitted answers ten times or more to the same questions, we found two and four user pairs suspected of pretending to be someone else, respectively. We intend to examine whether these user pairs are multiple account users from various perspectives and make criteria of suspicious user determination.

5 Conclusion

In this paper, we showed our method detected some users suspected of pretending to be other users in a certain category of a community site by using their answers submitted to other (non-target) categories. Furthermore, we examined the effects on the accuracy of user identification when we relaxed the criteria of selecting learning examples. For example, the identifier developed by using training data (target category, 100 answers) had about the same accuracy as the identifier developed by using training data (non-target categories, 200 answers).

References

Argamon, S., M. Saric, and S.-S. Stein. 2003. Style mining of electronic messages for multiple authorship discrimination: first results. *Proceedings of the Ninth ACM Conference on Knowledge Discovery and Data Mining*, 475-480.

Craig, H. 1999. Authorial attribution and computational stylistics: if you can tell authors apart, have you learned anything about them?. *Literary and Linguistic Computing*, 14(1), 103-113.

de Vel, O., A. Anderson, M. Corney, and G. Mohay. 2001. Mining e-mail content for author identification forensics, *Record of ACM Special Interest Group on Management Of Data*, 30(4), 55-64.

Ishikawa, N., R. Nishimura, Y. Watanabe, M. Murata, and Y. Okada. 2010. Detection of submitters suspected of pretending to be someone else in a community site. *Proceedings of the Seventh Language Resources and Evaluation Conference*, 3097-3100.

Koppel, M., S. Argamon, and R. Shimoni. 2002. Automatically Categorizing Written Text by Author Gender. *Literary Linguistic and Computing*, 17(4), 401-412.

Zheng, R.-T., J. Li, H. Chen, and Z. Huang. 2006. A framework for authorship identification of online messages: Writing style features and classification techniques. *Journal of the American Society for Information Science and Technology*, 57(3), 378-393.