Detection of submitters suspected of pretending to be someone else in a community site

Naoki Ishikawa, Ryo Nishimura, Yasuhiko Watanabe, Yoshihiro Okada, Masaki Murata

Ryukoku University, Dep. of Media Informatics, Seta, Otsu, Shiga, 520-2194, Japan
t060528@mail.ryukoku.ac.jp, r_nishimura@afc.ryukoku.ac.jp, watanabe@rins.ryukoku.ac.jp, okada@rins.ryukoku.ac.jp
NICT, Seika-cho, Soraku-gun, Kyoto, 619-0289, Japan
murata@nict.go.jp

Abstract

One of the essential factors in community sites is anonymous submission. This is because anonymity gives users chances to submit messages (questions, problems, answers, opinions, etc.) without regard to shame and reputation. However, some users abuse the anonymity and disrupt communications in a community site. These users and their submissions discourage other users, keep them from retrieving good communication records, and decrease the credibility of the communication site. To solve this problem, we conducted an experimental study to detect submitters suspected of pretending to be someone else to manipulate communications in a community site by using machine learning techniques. In this study, we used messages in the data of Yahoo! chiebukuro for data training and examination.

1. Introduction

In these days, many people use community sites, such as Q&A sites and SNS, 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 do the following types of message submissions:

TYPE I a question and its answer are submitted by one and the same user. We think that the user intended to manipulate the message evaluation.

TYPE II two or more answers are submitted to the same question by one and the same user. 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.

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 submitters suspected of pretending to be other users to manipulate communications in a community site. In recent years, a large number of studies have been made on authorship identification based on analyzing stylistic features of messages (Craig, 1999) (de Vel et al., 2001) (Koppel et al., 2002) (Corney et al., 2002) (Argamon et al., 2003) (Zheng et al., 2006), however, few researchers addressed the identification issues of authors who submit messages in a community site. To solve this problem, in this study, we propose a method of detecting submitters suspected of pretending to be someone else to manipulate communications in a community site. In this method, in order to detect submitters suspected of pretending to be someone else, we used an submitter identifier which was developed by learning stylistic features of user's messages and determine by whom a series of input messages are submitted. In this study, we used messages in the data of Yahoo! chiebukuro, a widely-used Japanese Q&A site, for observation, data training, and examination. 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 [1].

2. Detection of submitters suspected of pretending to be someone else

TYPE I submissions, described in section 1., are sometimes obscurer than TYPE II submissions because the standards of message evaluation differ with each user. 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.
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 of writing. Figure 1 shows the outline of our method of detecting users who have similar styles of writing.
In our method, we used a submitter identifier which is based on analyzing stylistic features and determines by whom a

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

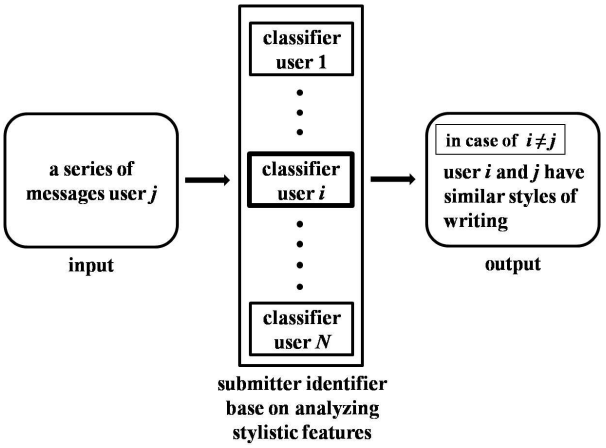


Figure 1: The outline of our method of detecting users who have similar styles of writing

series of input messages are submitted. As shown in Figure 1, the submitter identifier consists of N user 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 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. For example, in Figure 1, a series of input messages submitted by user j are given to the submitter identifier. Then, the identifier determines that the series of of input messages were submitted by user i. 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 submitted l answers to a communication site, ranked i-th place in the ranking of frequent answer submitters, and is the target user of classifier (user i). When a series of m answers of user j are given to classifier (user i), probability score $score(i, j)$ that user i and user j were one and the same user and user i submitted the series of m answers is calculated as follows:

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

where P_{ijk} is the probability that user i submitted message k ($1 \leq k \leq m$) in the series of m messages of user j. P_{ijk} is calculated by classifier (user i), which was developed by learning stylistic features of user i. Training data for learning stylistic features of user i consists of

- n messages which were selected randomly from l messages submitted by user i, and
- n messages which are extracted randomly from messages submitted by other users.

$S1$	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 2: 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.

In this study, we used the maximum entropy (ME) method for data training. Figure 2 shows feature $S1 \sim S7$ used in machine learning on experimental data. $S1$ and $S2$ were obtained by using the results of the morphological analysis on experimental data. $S3$ and $S4$ were obtained by extracting character 3-gram from experimental data. This is because Odaka et al. reported that character 3-gram is good for Japanese processing (Odaka et al., 2003). $S5$ and $S6$ were introduced because, we thought, clue expressions to the author identification are often found at the head and end of Japanese sentences. $S7$ was obtained by using PrefixSpan[2]. PrefixSpan is a method of mining sequential patterns efficiently and often used in document classification. By using PrefixSpan, Tsuboi et al. identified mail senders (Tsuboi and Matsumoto, 2002) and Matsumoto et al. classified reviews into positive and negative ones (Matsumoto et al., 2004).

3. Experimental results

To evaluate our method, we conducted the following experiments:

experiment 1 The accuracy measurement of the user classifiers.

experiment 2 The accuracy measurement of the submitter identifier.

experiment 3 The detection of users who have similar styles of writing and submitted answers to the same questions.

In this experiment, the target users were all submitters who submitted over 200 answer messages to PC, healthcare, or social issues category in Yahoo! chiebukuro. Table 1 shows the numbers of target submitters and their messages in each category.

[2]http://prefixspan-rel.sourceforge.jp/

Table 1: The number of target users in PC, healthcare, and social issues category.

	PC	healthcare	social issues
submitters	395	134	312
messages	260183	57406	180503

We developed experimental data for data training and examination in the next way. First, in order to develop experimental data of examination, we extracted 50 messages from each user's messages. Then, from the other messages of each user, we extracted 50, 100, and 150 messages and, as mentioned in section 2., developed three different sizes (100, 200, and 300 messages) of experimental data for data training. 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.

In experiment 1, we first developed user classifiers by applying maximum entropy (ME) method to the training data. Then, we varied the numbers of input messages to the classifiers and measured the accuracy of them. Input messages were extracted from the experimental data for examination. Figure 3 shows the accuracy of the classifiers under the various numbers ($1 \sim 5$) of input messages and the various sizes (100, 200, and 300 messages) of training data. As shown in Figure 3, we obtained more than 95% accuracy when we set the size of training data and the number of input messages to be 300 (including 150 target user's messages) and 4, respectively.

In experiment 2, we measured the accuracy of the identifier consisting of N classifiers, the accuracy of which are shown in Figure 3. 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 messages) of training data. As shown in Figure 4, we obtained approximately 85% accuracy when we set the size of training data and the number of input messages to be 300 (including 150 target user's messages) and 16, respectively.

In experiment 3, because we wanted to use the identifier with 85 % accuracy, we gave training data consisting of 300 messages (including 150 target user's messages) and set the number of input messages to be 16. Table 2 shows the numbers of user pairs who have similar styles of writing and submitted answers to the same questions. In this experiment, we found two user pairs suspected of pretending to be someone else to manipulate communications. Those user pairs submitted answers to the same questions in social issues category 43 and 17 times, respectively. We intend to examine whether these user pairs are multiple account users, from various perspectives.

Table 2: The numbers of user pairs who have similar styles of writing and submitted answers to the same questions.

	frequency of submissions to the same questions	
category	one or more	ten or more
PC	87	12
healthcare	17	0
social issues	109	22

4. References

Argamon, Saric, and Stein. 2003. Style mining of electronic messages for multiple authorship discrimination: first results. *9th ACM SIGKDD*, pages 475–480.

Corney, de Vel, Anderson, and Mohay. 2002. Gender-preferential text mining of e-mail discourse. *ACSAC 2002*, page 282.

Craig. 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, Anderson, Corney, and Mohay. 2001. Mining e-mail content for author identification forensics. *ACM SIGMOD Record*, 30(4):55–64.

Koppel, Argamon, and Shimoni. 2002. Automatically categorizing written text by author gender. *Literary Linguistic and Computing*, 17(4):401–412.

Matsumoto, Takamura, and Okumura. 2004. Sentiment classification using word sequences and dependency trees. *FIT2004*, pages 212–214.

Odaka, Murata, Gao, Suwa, Shirai, Takahashi, Kuroiwa, and Ogura. 2003. A proposal on student report scoring system using n-gram text analysis method. *IEICE Transactions*, J86-D-I(9):702–705.

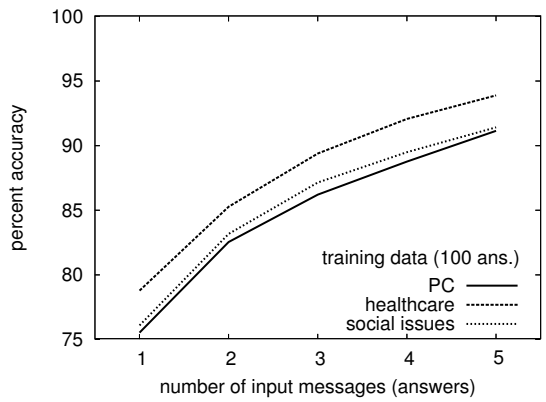
Tsuboi and Matsumoto. 2002. Authorship identification for heterogeneous documents. *IPSJ SIG NL*, 2002(20):17–24.

Zheng, Li, Chen, and 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.

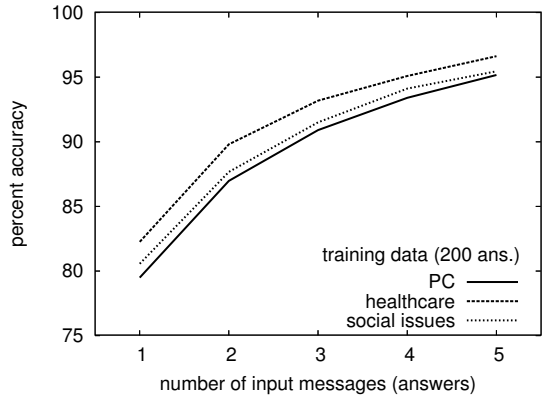
Acknowledgement We would like to thank Yahoo Japan Corporation and The National Institute of Informatics who provide us the data set of Yahoo! Chiebukuro. This research has been supported partly by the Grant-in-Aid for Scientific Research (C) under Grant No.20500106.

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

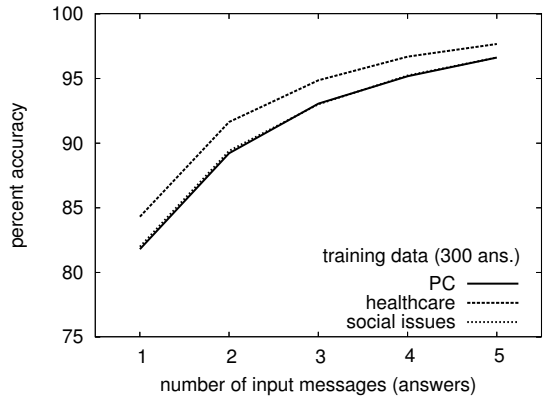
[4] http://mecab.sourceforge.net/



(a) training data (100 answers)

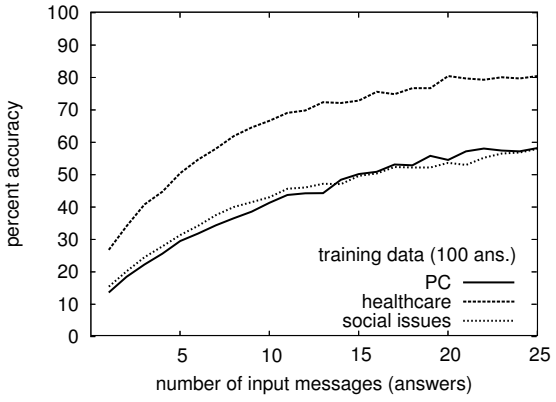


(b) training data (200 answers)

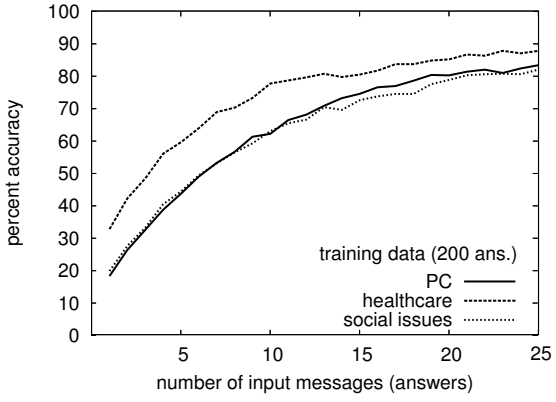


(c) training data (300 answers)

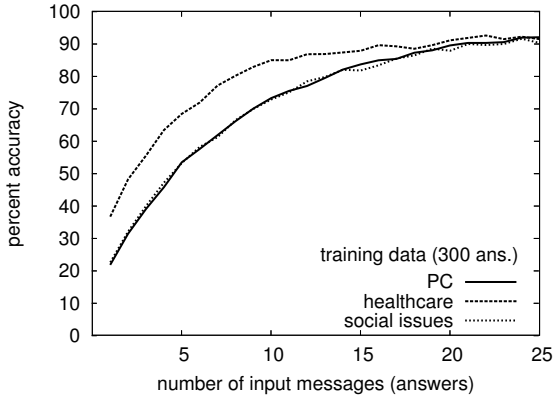
Figure 3: The accuracy of the classifiers which determine whether a series of messages were submitted by their target users, under the various number ($1 \sim 5$) of input messages and the various size (100, 200, and 300 messages) of training data. The target users were all submitters who submitted over 200 answer messages to PC, healthcare, and social issues category in Yahoo! chiebukuro.



(a) training data (100 answers)



(b) training data (200 answers)



(c) training data (300 answers)

Figure 4: The accuracy of the identifier which determines by whom a series of messages were submitted, under the various number ($1 \sim 25$) of input messages and the various size (100, 200, and 300 messages) of training data. The target users were all submitters who submitted over 200 answer messages to PC, healthcare, and social issues category in Yahoo! chiebukuro.