IR-based k-Nearest Neighbor Approach for Identifying Abnormal Chat Users

Notebook for PAN at CLEF 2012

In-Su Kang¹, Chul-Kyu Kim¹, Shin-Jae Kang², Seung-Hoon Na³

¹Kyungsung University, South Korea  
²Daegu University, South Korea  
³Electronics and Telecommunications Research Institute, South Korea  
¹dbaisk@ks.ac.kr, ¹ckkim@ks.ac.kr, ²sjkang@daegu.ac.kr, ³seunghoonna@gmail.com

Abstract. This paper addresses a task of automatically identifying abnormal chat users where training data is given as a collection of chat messages from both abnormal and normal users. We employ a k-NN classification based on an IR technique. A document is constructed in per-conversation for each user by concatenating his/her messages in a conversation. A query is constructed for a new user in the same way. A k-NN classification is then performed using top retrieved documents in response to the query.

1 Introduction

A chat user has his/her intended goals when taking part in chatting with others. This paper addresses a task of identifying such goals of a chat user. Our assumption is that strong clues for inferring their intended goals may commonly appear in chat messages of similar users. Based on this assumption, we represent a chat user as a document comprising his/her chat messages in a specific conversation and then identify chat users with abnormal goals by finding chat user documents with similar goals. We employ an information retrieval (IR) technique to discern such documents. In a training step, we prepare an IR system by indexing a collection of chat logs with chat-user goals marked either ‘abnormal’ or not. Given an unseen chat user, its chat messages are collected to formulate a query to be submitted to the IR system, and its chat goal is automatically classified using top-retrieved documents to which a k-NN approach is applied.

2 Method

A chat conversation can be viewed as a set of one or more user documents each of which consists of sentences from a particular user of the conversation. The training conversations are thus converted into a collection of user documents which is indexed using an information retrieval (IR) system. Given a test conversation, it is divided
similarly into documents \( \{ q_i \} \) each of which is then submitted as a query to the IR system to retrieve a set \( R = \{ d_1, \ldots, d_k \} \) of its highly related \( k \) training documents. For each \( q_i \), the following \( k \)-nearest neighbor classifier (Tan, 2005) is then used to determine whether \( q_i \) is uttered from a sexual predator (SP) or not:

\[
c^* = \arg \max_{c \in \{Y,N\}} \sum_{d \in R} \text{sim}(q_i, d) \delta(d, c)
\]

\[
\delta(d, c) = \begin{cases} 
1 & \text{if } d \in c \\
0 & \text{if } d \notin c 
\end{cases}
\]

where \( Y \) and \( N \) indicate SP class and non-SP class respectively, and \( \text{sim}(\cdot, \cdot) \) is a query-document similarity score from the IR system.

### 3 Evaluation Results and Discussion

To evaluate the performance of our IR-based \( k \)-NN classifier and to find the best parameter value for the number \( k \) of top-retrieved documents, 5-fold cross-validation was performed on the training set. Apache Lucene\(^1\) was employed for the IR system. Without stop-words removal and stemming, all 1-gram and 2-gram terms were used for index terms, where only rare terms with frequencies less than 3 were removed. For retrieval, the Lucene’s default retrieval formula was used.

![Figure 1. 5-fold cross-validation for training data.](http://lucene.apache.org/core/)
0.7960, and 0.3373, respectively for precision, recall, and F1. However, it was found that in our official runs, roughly a half of the test set was missed when preparing run submissions. So, we have fixed the error and have reiterated the same experiment. Figure 2 presents our revised result.

As Figures 1-2 show, the best $k$ values for $k$-NN classifier are significantly different between training and test data, and this is the main reason for the poor performance in this year’s SPI task. Using a robust value for $k$ was indeed important in our approach; when we used $k$ values which are optimal for both training and test data, the proposed method showed more than 60% and 70% in F1, respectively for training and test data.

Overall, our current use of $k$-NN classifier was not very successful in obtaining a good performance. We believe that this is because our current approach is not so matured with a lot of further explorations remaining. In the future, we will further examine the effect using document similarity on the same task by focusing on finding a robust range for $k$ and using more advanced IR similarity functions, and so on.

![Figure 2. Performance for test data.](image)

**References**