The 32nd Annual Conference of the Japanese Society for Artificial Intelligence, 2018

Filtering out improper accounts from Twitter user accounts for discovering individuals interested in certain topic

Chao Cai[*1] Shun Shiramatsu[*1]

[*1] Department of Computer Science, Graduate School of Engineering, Nagoya Institute of Technology

Nowadays, the continuously growing popularity of Web-based Debates and online survey are requiring participants from different places and backgrounds which raises the need of the application for topic-oriented participants recommendation based on social networking service(SNS). Among several kinds of SNS, Twitter holds over 300 million users located all over the world. Therefore, the aim of this paper is to propose a filter technique applying natural language processing skill, which filters the individual users from other official or robot users of twitter for Web-based Debate system and online survey participants recommendation. To keep the balance of user groups for training, the search was divided into two parts with different list of keywords. As result, we gained the F measure of 0.93 by processing the data with SVM.

1. Introduction

Recent years, the application of Web-based debate system is gradually growing. In 2013, the debate concerning Nagoya City Time Plan by Imi et al. [Imi 2014] involved 266 people by utilizing the Online Debate System COLLAGREE. Recommending and inviting participants online has become an essential issue when organizing a debate or survey. This paper is focused on twitter which holds over 300 million users from different locations. The research A Machine Learning Approach to Twitter User Classification by Marco Pennacchiotti et al. [Pennacchiotti 2010] could gather groups of twitter user who are interested in certain topic. However, there are unavoidably official users or robot users mixed in the user group. To discover individual users who can participate in debate or survey on specific topic, we describe a user filter which can filter out the improper users.

Since the robot or official user generally has different tweet behavior and user information, the main idea of this paper is to suppose a filter based on the concept. The system is designed to analyze the tweet part (created time and tweet text) and the user information part (onscreen name and description) of each data set with TFIDF and support vector machine (SVM).

2. Designing feature vector

2.1 Details of the data

We have collected 669 data sets in advance through twitter stream API and tweet search API, in which each data set involves 4 parts: (1) tweet text, (2) tweet created time, (3) user name, (4) user description. To keep the balance of number of individual users and improper user, the data was collected based on two different lists of keywords in two separated searches. The ratio of individuals and improper users is 452:217.

2.2 Features of user groups

As the main concept defined above, to decide which group one user is, two main parts are considered, which are the tweet contents and user information. For the improper users, we separated them into 3 groups: (1) official user, (2) inactive user, (3) robot user by analyzing the data.

- Official user: specific terms in user onscreen name or description (e.g. kousiki akkaunto (official account) in description or company name in both part).
- Inactive user: retweet only the campaign contents and without a user description with random characters terms in onscreen name.
- Robot user: specific terms (e.g. bot or common entities) in user onscreen name or description, involving advertisement or promotion either in description and tweet content.

While the individual users who generally have specific terms representing their real name in onscreen name and most likely with a description involving location or date of birth. Table 1 shows the samples contents of each user group.

2.3 Pre-processing the data

As we mentioned above, the specific terms in each tweet text or user information may help judge the group of users. Hence, for words embedding and extracting the classification information, we apply MeCab and term frequency-inverse document frequency(TFIDF) algorithm. In this work, all tweets and users have been collected under a list of Japanese keywords. Meanwhile, MeCab can segment Japanese sentences into words which can be the input of TFIDF later. TFIDF is to retrieve unique terms for document in a corpus. Since the same term will unavoidably appear in both tweet content part and user information part, we process them separately by TFIDF to generate two matrix-like data which both contain 669 rows with 5110 dimensions for tweet part, 5949 dimensions for user information part. Because the same row in two matrix data refers to one user-tweet data set, we combine the two vectors to generate a xxx multiply xxx matrix. For each vector, dimension from 1st to 5110th refers to tweet contents and the rest dimensions refer to user information.

2.4 Future works

For our future works, we are planning to take the user tweet behavior (the tweet frequency in a certain period, times of retweet

Contact: Chao Cai, Nagoya Institute of Technology, Gokiso-cho, Showa-ku, Nagoya-shi, Aichi 466-8555, Japan

Table 1: Samples for each user group

Group	Onscreen name	Description	Tweet text
Official user	@HuffPostJapan	ハフポスト日本版公式アカウントです。 ライフスタイル@HPJPLifestyle	みんなに支持されるのは、どちらの歌でしょうか？
Robot user	@warakosodateBOT	ちょっと知っておくと子育てで役に立つヒントや、笑ってしまう子育てエピソードをつぶやきます！	親だって人間だから、感情的に怒鳴っちゃうことだってあるよね～。そんな時は後で素直に謝ればいいんじゃないかな。
Inactive user	@vJX3qMICBftMeco	None	RT @UNITED_CINEMAS: ★赤ちゃんと一緒に映画を観られる「抱っこ de シネマ」 10 月・11 月の上映作品決定！
Individual user	@shochan710	大阪のはずれ、ひらパー村から愛を叫ぶ…(≧ω≦) 「ぶっさん大好き～！」 まだまだ強火剛担ジャニオタ高校生男子と親子で V6応援中～	へそ出し短パンの 20 代、ミニスカ・ネイルの 30 代(OL さん)、遠征・オールの 40 代・・

et al.), user relationship (followers and followee analysis) and analysis of other contents of user information (created time, location etc.) into consideration for design of feature vector to construct a more distinguishable representation of each user.

3. Filter with SVM

SVM is known as an efficient machine learning model with high accuracy by constructing a hyperplane that can represents the largest margin between two groups of data, which have been constantly showing reliance on binary classification. By applying the Kernel Function, SVM can perform a nonlinear classification. In this paper, we apply SVM with RBF Kernel, which is with 2 significant parameters, cost parameter (C) and Gamma.

3.1 C and Gamma

C is a parameter that determines the penalty of misclassification in case that the data cannot be separated linearly (or nonlinearly). Gamma parameter will change the complexity of detective range which unavoidably decrease the compatibility of the model. The Gaussian function can be defined as below:

$$K(x, x') = exp(-\gamma \|x - x'\|^2)$$

We can find out that the graph of K will become more complex as the value of gamma grows. Thus, the tuning of parameter will be important for this work.

3.2 Result of experiments

We simply run the filter to process the data that we have vectorized and fabricated for 60 times. For each 10 times, the parameters are changed, and the training-test data separation will be randomly determined to gain a cross-validation like result, of which we calculate the averages of precision score, recall score and F measure. We simply show the results with different parameter settings in Figure 1, 2 , 3. In addition, the default setting is C = 1.0, Gamma = 0/number of dimensions.

We can find out from Figure 1 and Figure 3 that the precision score and F measure can be only visually affected with large C and small gamma and the recall score barely change as the parameters change.

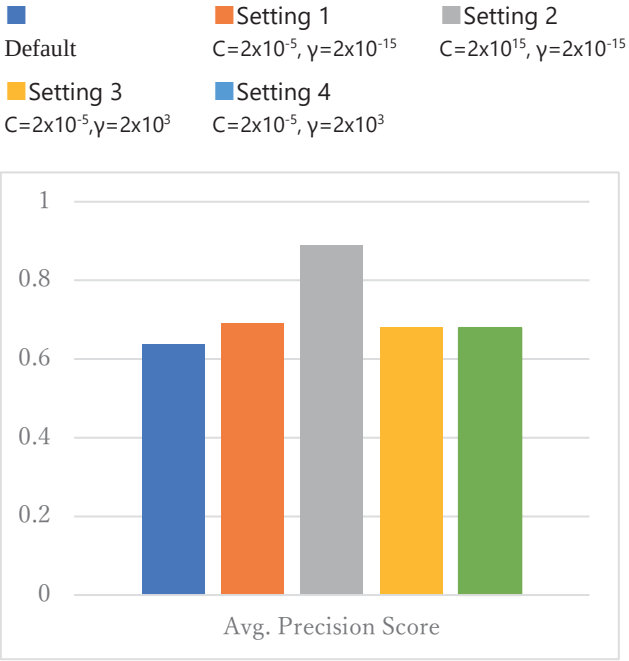

Figure 1: the average of precision score

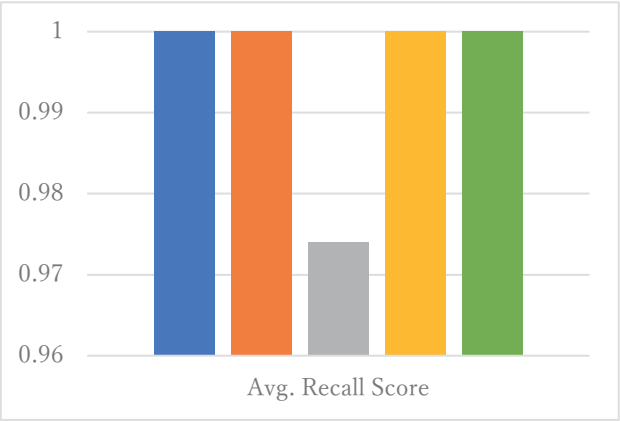

Figure 2: the average of recall score

4. Evaluation of result

Precision Score. Since the precision score is determined by the number of true positive samples and false positive samples, the miss recognition of improper user significantly affects the precision score. Because of the similarity of users group and the small size of data with less well balance, SVM runs with small C will simply misjudge the user group since there exist much user data without unique feature.

Recall score. Because of the balance of user group, the individual user prediction appears unavoidably more than improper one. Since the recall score decreases only when misjudgment of individual user occurs, the recall score stays stable as expected.

F measure. F measure is the harmonic mean of precision and recall. We suppose that F measure indicates the convincible accuracy of the model. As the result shown in Figure 3, much tolerance of misrecognition and simpler detected range by large C and small gamma significantly improve the performance of this model even though the recall score decreases.

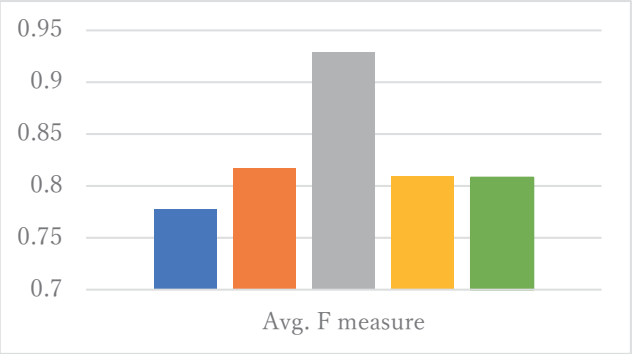

Figure 3: the average of F measure

5. Conclusions and future work

We presented a filter system that filter out the improper user from the raw data set collected by twitter API. We attempted to summarize the features of each user group with the tweet contents and user description. With the feature vector we designed, we have gained acceptable result and showed that the user filter is a feasible task.

However, the data size and balance have limited the reliance and the compatibility of this model. The design of feature vector is too simple especially when dealing with larger size of data. Hence, for future work:

- Larger size of data
- Introducing Deep Learning such as CNN into the model
- More well-thoughted design of feature vector

Acknowledgement

This work was partially supported by JST CREST (No. JPMJCR15E1) and JSPS KAKENHI (No. 17K00461).

References

[Imi 2014] 伊美祐麻 伊藤孝行 伊藤孝紀 秀島栄三, “大規模意見集約システム COLLAGREE の 開発と名古屋市時期総合計画に関する社会実験”, JSAI, 4M1-4, 2014

[Pennacchiotti 2011] Marco Pennacchiotti, Ana-Maria Popescu, ***A Machine Learning Approach to Twitter User Classification***, Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media, AAAI, 2011.