Real Friendship and Virutal Friendship: Differences in Similarity of Contents/People and Proposal of Classification Models on SNS

Takafumi Komori*§, Yoshinori Hijikata†, Tomu Tominaga*, Shogoro Yoshida*, Nobuchika Sakata‡, Kensuke Harada*
*Graduate School of Enginnering Science, Osaka University, Japan
†School of Business Administration, Kwansei Gakuin University, Japan
‡Division of Information Science, Nara Institute of Science and Technology, Japan
Email: §komori@hlab.sys.es.osaka-u.ac.jp, †contact@soc-research.org

***Abstract*—When people anonymously use social network sites (SNSs) like Twitter, they may interact with not only real world friends but also strangers, who are not acquaintances in the real world. Therefore, both of real friendship (RF) and virtual friendship (VF) coexist in these SNSs. In this research, we investigated the differences in similarity of user pairs in Japan by their types of relationship, i.e. RF or VF. The primary results indicated that RF user pairs have more common follow users on SNSs than VF user pairs, and that contents posted by VF user pairs include more similar words than RF user pairs. It is implied that two users with RF have a similar interest in neighborhood users, and two users with VF are interested in similar topics.**

After that, we built the models to classify user pairs into RF or VF using the similarity measures. These models showed high performance to distinguish between RF and VF (their F-measures are larger than 0.80).

***Index Terms*—real friendship, virtual friendship, similarity, classification model, Twitter**

I. INTRODUCTION

The development of the Internet has made it possible for people to interact through online communication tools such as e-mail, online chat, or social network sites (SNSs), without meeting in the real world. Communication in the virtual world enables not only to complement and strengthen existing offline relationship [1], [2], but also to build friendship only online [3]. In particular, on SNSs, it has become much easier to responded content posted by people who have never met in the real world, or to form friendship with them.

A lot of researchers have examined the relationship formed only in the virtual world. Differences between the relationship in the virtual world and that in the real world have been studied in terms of the Internet addiction [4], the self-disclosure [5], and its ability of the information diffusion [6].

The authors are also interested in the differences between the two types of relationship and are especially interested in the user similarity between the people of the same type of relationship. We consider the following two aspects: 1) contents and 2) people might differ between the two types of relationship. Thus, we verify whether there exist differences in each aspect by the relationship types using users' action data in SNSs. We define the relationship between two users who recognize each other in the real world as "real friendship" (RF), and define the other relationship as "virtual friendship" (VF). In addition, we call a user pair with real friendship "RF user pair", and a user pair with virtual friendship "VF user pair".

We assume that the similarity of user pairs in SNSs differs by the relationship types, i.e., RF or VF. Thus, two users with RF are likely to share some of their real-world friends. Considering this, we think they are expected to follow same users on SNSs. For example, two persons working in a same office would usually follow same colleagues in their office. On the other hand, two users with VF are likely to have common hobbies or preferences, while they have extremely low possibility to share their real-world friends. Based on these assumptions, we believe that 1) RF user pairs are similar in users that they connected to and 2) VF user pairs are similar in contents that they post and respond.

We validated these assumptions for Japanese user pairs who follow each other in Twitter. To quantify the similarity of user pairs, we introduced two measures according to the two aspects, called "neighborhood similarity" and "content similarity" in this study. We then compared them by the relationship types. Neighborhood similarity is similarity between a user set one user follows and a user set the other user follows. Content similarity is similarity between a tweet set one user posted or liked and a tweet set the other user posted or liked. We attempt to investigate the differences between RF and VF by quantify these similarities from the action data of Twitter users.

After that, we built friendship classification models using several machine learning algorithms. As far as we know, there is no research that addresses classification of the two relationship types (RF and VF). As a results, these models showed high performance (F-measures are larger than 0.80). We expect that these classification models are applied to privacy-protection in which the system suggests or selects audience when users share contents, especially when they share their private information. The models can also be used for user recommendation. When RF occupies most of a user's friendship, the system will recommend him/her users who might have RF with the user. When VF occupies most of the user's friendship, the system will recommend him/her users who might have VF with the user.

The present paper is constructed as follows. Firstly, we introduce related work. We then describe the data collection method and the analysis method. Next, we show our results and the implications. Finally, we discuss the application of our models and limitations in this research.

II. RELATED WORK

In this section, we summarize previous research focusing on relationship formed only in the virtual world. We then clarify the position of this study.

We use terms with meaning similar to VF and RF (e.g., onlineoffline friends, onlineoffline relationship). Note that they are the terms defined in each cited paper.

A. Value of friendship in the virtual world

With the spread of the Internet, people have more chances to build relationship even though they do not meet in the real world. The value of the friendship formation in the virtual world has been one of the central research topics. Chan et al. showed that the value of online friends is less than that of offline friends [7]. However, they also found that connections to online friendship lasting more than a year have the value comparable to offline friendship. Chan et al. investigated the differences in friendship quality and intimacy between online and offline interactions [8]. As a result, the quality of offline relationship is slightly higher than that of online relationship, while online friendship displays a higher intimacy level than the offline relationship. Bulow et al. argued that relationship in the virtual world have unique value, because people can build equal relationship with others even though they were unequal in the real world [9]. As these studies showed, VF is not as valuable as RF, but VF may have independent value that RF does not have.

B. Differences between VF and RF

The differences between online friendship and offline friendship have been examined from various viewpoints (e.g., about their motives, demographic status). Wang et al. found nine motives of forming cyber relationship [10]. They found several motives peculiar to the virtual world such as "the opportunity to meet new people" and "escape from the real world," which are different from motives of forming real relationship. It is shown that offline friends are more similar in age, sex, and place of residence than online friends [11], [12]. Zinoviev et al. revealed that 25% of online relationship is recognized as true friendship, not simple acquaintanceship [13]. Antheunis et al. revealed the differences in cognitive similarity between offline friendship and online friendship [14]. Kim et al. focused on ego networks of Twitter users, and then investigated the difference in the power of information diffusion on online and offline relationship [6]. They investigated the differences in usage features of Twitter users between the two relationship types, including user pairs which are one-way following.

C. Influence of VF on the real world

It is also investigated how the formation of the relationship in the virtual world affects the personality and the health of the person. Interaction with strangers online has led to reduction of social anxiety [15], [16]. However it is also known that building more online friends is related to Internet addiction [4], [10]. On the other hand on the online social network that manages and shares users' health status, users' weight changes correlated positively with the number of their online friends [17].

D. Position of this research

In this paper, we investigated the differences in similarity of contents and follow between RF user pairs and VF user pairs by using the action data of Japanese Twitter users. There is no research that studies the differences between RF and VF from the viewpoint of the similarity of contents or follow as far as we know. Although Kim et al. also examined the differences between RF and VF by using the action data of Twitter users [6], our research focuses on two aspects (contents and follow) of users' similarity and consider the differences of user pairs' similarity by the relationship types.

III. DATA COLLECTION

In this section, we explain how to collect user pairs and obtain their action data on Twitter.

A. User pairs

We investigate differences between RF and VF in contents/follow similarity of user pairs on Twitter. Based on our definition of the two types of relationship, we target only user pairs which follow each other on Twitter. We do not target user pairs of one-way following because one user follows the other to read his/her posts. This means that they do not necessarily recognize each other in SNSs.

We built a survey system to obtain user pairs for our study. We recruited 96 participants who have Twitter accounts through a direct message in SNSs. We asked them to answer questionnaires using our online survey system. The system randomly selected up to 30 mutual-follow accounts of the participant's Twitter account, and asked him/her to answer whether he/she recognizes owners of the selected mutual-follow accounts in the real world. We classified the user pairs that the participant answered they knew each other as RF and classified the other user pairs as VF. As a result, 1388 user pairs, 1029 RF user pairs and 359 VF user pairs, were obtained from 96 Japanese participants (71 males, 25 females). The distribution of their ages was 10's (n = 8), 20's (n = 86) and 30's (n = 2).

B. Twitter data

We used Twitter REST API to obtain action data of the users (the users in all the user pairs) such as tweets, favorites, and information about followers and followees (users that a user follows). As a result, 4,372,128 tweets and 19,821 user profiles were obtained (January 19th - 25th, 2017).

IV. ANALYSIS

In this section, we describe the measures of similarity used for our analysis and the procedure of our analysis.

A. Measures

We presume that RF and VF differ in two aspects of similarity : 1) contents and 2) people. To verify these assumption, we introduced content similarity and neighborhood similarity and used several measures for them. These measures are calculated for each user pair. In this section, we explain the definition of the measures. Note that these measures are calculated for each user pair. In the following explanation, A and B are users in a user pair.

1) Content similarity: For calculation of content similarity, we make three measures which indicate the degree of similarity of contents that two users in a user pair posted or responded. Three measures are similarity of words used in tweets (*tweet content*), overlap of tweets liked by users (*like tweet*) and similarity of the period of times when posting tweets (*post time*).

The reason for including *post time* is that we hypothesize that VF user pairs frequently post tweets in the specific period of times about certain hobbies or preferences. For example, both users with VF who like common anime would post tweets more active when the anime is aired on TV.

The calculation methods of the three measures are described below.

- *tweet content*
  In *tweet content*, we measure similarity of the words in tweets they have posted. First, we created a document that unifies his/her all tweets for each user in our dataset and extracted only nouns from the document with the morphological analyzer; Mecab[1]. Here, let the set of nouns in the document of user i be D_i. Next, we removed nouns appearing in more than 25% and those appearing less than 1% from the documents instead of excluding stop words. We then calculated the value of TF-IDF for noun j in D_i. The value is defined as w_i^j. User i has a word vector u_i whose elements are from w_i^1 to w_i^n ($n = |D_i|$). We define *tweet content* as cosine similarity between u_A and u_B.

$$u_i = (w_i^1, w_i^2, \cdots, w_i^n)^T$$

$$\textit{tweet words} = \frac{u_A \cdot u_B}{|u_A||u_B|}$$

- *like tweet*
  Like tweet means similarity of tweets that both users in a user pair have given ”Like”. ”Like” is a function in Twitter to show appreciation for other users’ tweets. The similarity is calculated by Jaccard index between two sets of tweets each user in a user pair has sent likes. We calculated *like tweet* as below.

$$\textit{like tweet} = \frac{|Like_A \cap Like_B|}{|Like_A \cup Like_B|}$$

[1]MeCab:Yet Another Part-of-Speech and Morphological Analyzer. http://taku910.github.io/mecab/.

  Here, $Like_X$ means a set of tweets that user X has sent likes.

- *post time*
  Post time refers similarity of the period of times when users in a user pair post tweets. To measure the similarity, we introduced posting time vector. At first, we divide a week by 30 minutes and name each period t_i (t_1: Sun 0:00-0:29, t_2: Sun 0:30-0:59, $\cdots$, t_{336}: Sat 23:30-23:59). Next, using all the tweets of user u, we define $n_{t_i}^u$ as the number of tweets which is posted in t_i. User u has a posting time vector p_u whose elements are from nu_{t_1} to $n_{t_336}^u$. We define *post time* as cosine similarity between p_A and p_B as below.

$$p_u = (n_{t_1}^u, n_{t_2}^u, \cdots, n_{t_{336}}^u)^T$$

$$\textit{post time} = \frac{p_A \cdot p_B}{|p_A||p_B|}$$

2) Neighborhood similarity: For calculation of neighborhood similarity, we use four measures which indicate the degree of overlap of users who two users in a user pair commonly connect to. It is supposed that a RF user pair belongs to a common community in the real world. Four measures are overlap of follow users (*follow*), overlap of mutual follow users (*mutual follow*), ratio of authority (*authority*), and ratio of protected users (*protected*).

The calculation methods of the four measures are described below.

- *follow & mutual follow*
  In *follow* and *mutual follow*, we measure the similarity of following networks of user pairs. In order to calculate these measures, we adopt Jaccard index as below. $Follow_X$ means the set of users who user X follows.

$$\textit{follow} = \frac{|Follow_A \cap Follow_B|}{|Follow_A \cup Follow_B|}$$

  $Mutual\ follow_X$ means the set of users who user X follows mutually.

$$\textit{mutual follow} = \frac{|Mutual\ follow_A \cap Mutual\ follow_B|}{|Mutual\ follow_A \cup Mutual\ follow_B|}$$

- *authority*
  In this paper, we call a user who attract a lot of users’ attention on Twitter “authority”, and authorities are determined by the number of followers and the follower-followee ratio (calculated by $followers/(followees+1)$). We chose authorities from users who are followed by both users in all user pairs in our dataset according to the criteria that the number of follower is in top 10% and follower-followee ratio is in top 10%. In *authority*, we measure the ratio of authorities in their common followees. The calculation method is as below.

$$\textit{authority} = \frac{\#\ of\ authority\ users}{|Follow_A \cap Follow_B| + 1}$$

- *protected*
  Twitter has an account-protection function that enables

users to hide their tweets from those who do not follow the users and to avoid being followed by other users without the users' permission. In *protected*, we measure the ratio of protected users in the users who a user pair commonly follow each other.

$$protected = \frac{\#\, of\, protected\, users}{|Mutual\, follow_A \cap Mutual\, follow_B| + 1}$$

B. Procedure

We conducted three analyses described below.

1) Variance differences: We verify the differences between RF and VF in similarity of contents and people. At first, we calculated measures defined in IV-A for all user pairs in our dataset. We then used the Mann-Whitney U test to investigate whether there are statistical differences in variances of these measures between RF user pairs and VF user pairs.

2) Classification models: To verify whether the relationship types (RF and VF) of a user pair can be estimated from the measures, we build binary classification models using logistic regression, random forest, and support vector machine (SVM).

First, we selected positive and negative datasets. To align the number of RF user pairs with the number of VF user pairs, we randomly selected the same number of RF user pairs as VF user pairs. The reason behind the random user selection is avoiding imbalanced-data problem [18]. RF user pairs are defined as a positive dataset, and VF user pairs are defined as a negative dataset.

Second, we conducted a 10-fold cross validation to avoid over-fitting problems. Thus, the training set contains 90% of a positive and a negative dataset respectively. For the training dataset, we train the classification models. In the all machine learning techniques, the objective variable is positive or negative. Moreover, the predictor variables are 6 measures (*tweet content*, *like tweet*, *post time*, *mutual follow*, *authority* and *protected*). Considering multicollinearity, *follow* is excluded from the predictor variables because *follow* has the high correlation with *mutual follow* ($r = 0.94$). In each machine learning model, we tuned for the best performance (e.g., controlling internal parameters), as shown below.

- Logistic regression
  We select influential predictor variables based on Akaike's Information Criteria (AIC) [19], which represents adaptability of models (lower is better). The variables selection aims to minimize AIC. Here, we adopt backward elimination method for the selection.
- Random forest
  This method measures mean decrease Gini index of each predictor variable. The Gini index reveals the extent of deviation of a classification result. The smaller the deviation, the better the classification result. Therefore, variables having larger mean decrease Gini index are regarded as better predictors. The number of variables for selecting the best performance is based on the mean decrease Gini index.
- SVM
  Instead of selecting influential variables, SVM tunes two internal parameters: cost and gamma, using grid search. Cost determines the extent of wrongly classified instances, and gamma represents boundary simplicity. An RBF Gaussian kernel is used for base conversion.

Finally, we evaluated the performance of each model using the test set with F-measure. This index is a harmonic mean of precision and recall. In order to reduce bias due to sampling of RF user pairs, this procedure was repeated ten times. We then calculated the averages of ten evaluation values of three models.

3) Contributory features of the classification models: To inspect contributory variables of these classification models, we confirmed the partial regression coefficients of logistic regression and mean decrease Gini index in random forest. In this analysis, we made a new dataset that includes the same number of each type of user pairs and additionally conducted logistic regression analysis and random forest.

Using the two types of machine learning techniques, we built three kinds of models: Hybrid model, Content model, and Neighborhood model. The predictor variables of Hybrid model consist all measures in content similarity and neighborhood similarity. While the predictor variables of Content model are only measures in content similarity, those of Neighborhood model have only measures in neighborhood similarity. With the three models, we examine how influential the similarity variables are to discriminate RF from VF.

V. RESULTS & DISCUSSIONS

A. Variance difference

Table I shows the mean, variance, and median of all the measures for each relationship type. Using the Mann-Whitney U test for each measure, we found statistically significant differences between RF and VF in all measures ($p < .001$). The results and considerations are summarized below.

1) Content similarity: VF user pairs showed higher values than RF user pairs in *tweet content*. VF user pairs should have common interests to express in their tweets, because they follow each other in Twitter even though they have never contacted in the real world. This may lead words in their tweets to be similar. Kim et al. stated that tweets similarity extracted with LDA showed higher value in real friendship [6], which is inconsistent with our result. Unfortunately, they did not clearly describe how to collect data and who are the study participants (e.g., nationality, gender, or age). Thus, it is difficult to discuss why we obtained the different results. Note that we do not say our results are general, because we collected the data from Japanese Twitter users by SNSs.

Like tweet and *post time* indicated the higher value in RF user pairs than VF user pairs against our expectations. Like has roles of expressing not only interests in contents of tweets but also gratitude or approval to users [20]. Therefore, there is a possibility that personal interests are not reflected in tweets sent like by users as much as tweets posted by users. As a reason why RF user pairs showed higher values at *post time*, we think that they posted tweets on common events that occur in the real world. For another reason, RF users possibly

TABLE I
CALCULATION RESULT OF THE MEASURES DEFINED IN IV-A. MEDIAN, MEAN AND VARIANCE ARE CALCULATED FOR RF USER PAIRS ($n = 1029$) AND VF USER PAIRS ($n = 359$) RESPECTIVELY. THE RELATIONSHIP TYPE WHICH SHOWS HIGH MEDIAN IS WRITTEN IN THE ROW OF RFORVF.

			RF			VF		
	Measure	RForVF	Median	Mean	SD	Median	Mean	SD
Content similarity	*tweet content*	VF	5.46e-5	1.88e-4	7.67e-4	7.09e-5	3.42e-4	7.00e-4
	like tweet	RF	8.31e-4	3.94e-3	8.48e-3	3.25e-4	1.78e-3	4.01e-3
	post time	RF	0.720	0.705	0.105	0.654	0.631	0.122
Neighborhood similarity	*follow*	RF	5.28e-2	7.18e-2	6.48e-2	1.88e-2	2.87e-2	3.48e-2
	mutual follow	RF	6.17e-2	8.87e-2	8.49e-2	1.69e-2	2.79e-2	3.71e-2
	authority	VF	0.054	0.128	0.184	0.375	0.443	0.326
	protected	RF	0.484	0.459	0.236	0.125	0.179	0.202

TABLE II
THE PERFORMANCE OF THE CLASSIFICATION MODELS. TO REDUCE BIAS DUE TO SAMPLING OF USER PAIRS, WE CONDUCTED 10-FOLD CROSS VALIDATION TEN TIMES FOR EACH MODEL. MEAN AND SD OF PRECISION, RECALL AND F-MEASURE IN THIS TABLE ARE CALCULATED BY THE RESULT OF TEN CROSS VALIDATIONS.

	Precision		*Recall*		*F-measure*	
Models	mean	SD	mean	SD	mean	SD
Logistic regression	0.804	7.89e-3	0.807	1.14e-2	0.804	8.75e-3
Random Forest	0.805	1.35e-2	0.821	1.41e-2	0.811	1.18e-2
SVM	0.802	5.52e-3	0.812	1.32e-2	0.806	7.28e-3

exchange replies more than once in a short time. The survey by Kim et al. also showed that the frequency of reply is higher in RF user pairs [6], so this may have affected *post time*.

2) Neighborhood similarity: The measures of neighborhood similarity except *authority* showed higher values for RF user pairs than VF user pairs. Especially in *follow* and *mutual follow*, we found the comparatively large differences between RF user pairs and VF user pairs. Therefore, it means that RF user pairs usually follow same users and follow each other with same users. This might be because RF user pairs establish friendship on SNSs based on the community they belong to in the real world.

The result of *protected* indicated that RF user pairs have higher ratio of protected users in common mutually-following users. Protected users can reject following requests from strangers and it is thought that they mainly follow only close friends like real-world friends. Users who follow a number of protected users might establish their following networks for the purpose of connecting with friends in the real world.

Although most of the measures showed higher values for RF user pairs than VF user pairs, only *authority* did not. In fact, we manually confirmed that VF user pairs often follow the common official accounts of games or anime. They are likely to build relationship based on common hobbies or interests, and may follow common authorities related to them.

B. Classification models

As shown in Table II, F-measures of all models are larger than 0.80. Besides, there is little difference in average value of F-measure for each model. From the high performance of these models, it is supposed that the values of similarity measures of user pairs on Twitter differ by the relationship types: RF or VF. Furthermore, these models can perform stably regardless of sampling from the small standard deviation of the F-measures.

C. Contributory features of the classification models

The results are shown in Table III (a). We confirmed that the coefficients and its significant probabilities of the model are substantially consistent. When the absolute value of the partial regression coefficient or the value of the mean decrease Gini index represents high values, it shows that the variables have strong influences on the model.

As a result, we found that *mutual follow*, *authority* and *protected* have strong influence on discrimination between RF and VF in both logistic regression and random forest. It suggests that these variables about follow are more important than those about content for the classification between RF and VF.

tweet content and *post time* showed significant influence on these models, however, the influential powers were weaker than variables in neighborhood similarity.

The variables calculated by like that users in a user pair has sent (*like tweet*) have little impact on these models. It is considered that likes users have sent tend to be closely related to other users who they follow, because *like tweet* is correlated with *mutual follow* ($r = 0.38$). Therefore, this result implies that *mutual follow* is more dominant than *like tweet* for the classification models. Follow, which is connection that maintain constantly, is more likely to explain the relationship types than the temporary interests to contents.

Table III (b) and (c) show results of content model and neighborhood model. From the result of partial regression coefficients in Table III (b), all the variables represent the statistically significant values ($p < .001$). The reason why there are not the statistically significant values in *like tweet* in section V-C may be that *like tweet* is closely related to the variables included in neighborhood similarity. However, it turns out that *like tweet* does not have much influence on the classification models. Thus, the variables defined by like activity are not likely to be important for building the

TABLE III
THE INFLUENCE OF THE PREDICTOR VARIABLES ON THE CLASSIFICATION MODELS THAT ARE BUILT BY (A) SIX MEASURE, (B) THREE MEASURES IN CONTENT SIMILARITY AND (C) THREE MEASURES IN NEIGHBORHOOD SIMILARITY. "β" MEANS PARTIAL REGRESSION COEFFICIENTS OF LOGISTIC REGRESSION AND "MDG" MEANS MEAN DECREASE GINI INDEX IN RANDOM FOREST. AIC MEANS AKAIKE'S INFORMATION CRITERIA [19]. $* \ldots p < .05$, $*** \ldots p < .001$

		(a) Hybrid model		(b) Content model		(c) Neighborhood model	
	Feature	β	mdG	β	mdG	β	mdG
	(Intercept)	0.020		-0.029		-0.033	
Content similarity	*tweet content*	-0.24*	42.7	-0.50***	93.4	-	-
	like tweet	-0.073	27.3	0.46***	74.7	-	-
	post time	0.24*	49.3	0.68***	114.0	-	-
Neighborhood similarity	*mutual follow*	1.16***	77.1	-	-	0.88***	104.7
	authority	-0.61***	72.1	-	-	-0.87***	84.5
	protected	0.94***	89.6	-	-	0.95***	120.5
AIC		684.2		871.8		703.9	

classification models.

From the results in Table III (c), we found that all variables have equally strong influence to the classification models. Moreover, when comparing logistic regression performed by the variables in contents similarity and that performed by the variables in neighborhood similarity, the value of AIC of the latter is higher than the former. Therefore, important variables for building the classification models between RF and VF are those of neighborhood similarity.

VI. Design Implications and Limitations

A. Design implications

Here, we discuss the implication of the classification models created in this study.

First, the classification models created in this research will be applicable to limitation of information disclosure range. According to the specifications of Twitter, tweets are published to all followers regardless of the topics of tweets. As a result, tweets you want to share with only real-world friends may be released to those who have not been acquainted with the users. Therefore, there is a risk that personal information such as face photographs and real names included in tweets may be leaked to strangers. For this reason, some people may hesitate tweets. If the classification models can distinguish RF from VF accurately, it is possible to hide their personal information included in tweets from strangers.

Since Twitter allows a user to own multiple accounts, some people have several accounts and use them properly by the topics of tweets they want to post. In the online survey we conducted in this study, we asked the participants a question : Do you use multiple accounts properly on Twitter? In fact, there were 51 people out of 96 who answered "Yes" to this question. However, when we asked the reason why the 51 people use multiple accounts, some of them answered that "Because I want to use properly one account for my hobby and the other account for interaction with my friends." or "I own multiple accounts because the account with my real name has many followers who have different preferences or hobbies from mine." From these responses, it is supposed that there are users who use multiple accounts for the purpose of restricting audiences of their tweets. By incorporating the classification models into Twitter, we can offer the more comfortable service and eliminate the trouble of using accounts properly.

Second, the classification models that distinguish RF from VF can be applied to switching the recommendation strategies of the advertisement. The classification models can predict the relationship types (RF or VF) of all the mutual-follow users who a user has. While the users who have a lot of RF in their follow network are likely to have user-oriented motivation, the users who have a lot of VF in their follow network are likely to have content-oriented motivation to make following networks. It may be valid to use the different recommendation strategies dependent on their motivations. For example, the system can provide the former users with the advertisement that friends in the real world are interested in, and the system can give the latter users the advertisement based on their own interests. Like this, the classification models allow us to present more appropriate advertisement to all users by switching the recommendation strategies depending on the characteristic of users.

Third, we can monitor the state of users forming VF by our models. Although the formation of VF allows us to interact with strangers who cannot meet in the real world, it also has some risks. The previous research revealed that forming more online friends is related to Internet addiction [4]. Moreover, it was reported that strangers connect to the youth on SNSs and urge the youth to kill themselves in Japan[2]. Monitoring users by our models, we are likely to cover users who have these risks in advance.

B. Limitations & Future work

There were several limitations in this study. The questionnaire we conducted was a small scale, and the sex, occupation and age of the respondents were biased because they were gathered mainly by university students. Especially in terms of age, it is worth noting that the result in this paper was obtained from the dataset that consists of young people. In the future, we collect more diverse people at a large scale conducting questionnaires in crowd sourcing services.

[2]https://www.sankei.com/affairs/news/171122/afr1711220043-n1.html

Besides, in this survey, we found several results different from the previous study by Kim et al [6]. This suggests that the usage trend of SNS may be different depending on the nationality of the target users. Therefore, we will investigate differences in usage characteristics of SNS among several countries in our future survey.

There is possibility of the performance improvement in the classification models to distinguish between RF and VF. In this survey, we used only the measures based on the similarities of two users in a user pair. However, we believe the existence of other features that might improve the performance of the models. For example, in the previous research, Kim et al. reveals that the frequency of interactions between two users tends to differ between RF and VF [6]. The features based on the frequency of interactions (reply, like and retweet.) may improve the classification models. In the future, we consider adding other measures to improve the accuracy of the classification models.

VII. Conclusion

In this research, we investigated the differences between RF user pairs and VF user pairs in similarity of their action on SNSs. We assume that there are two aspects of similarity: 1) content and 2) people; and define two types of similarity according to these aspects: contents similarity and neighborhood similarity. We examined Japanese Twitter users to understand the differences of the similarities in these aspects between RF user pairs and VF user pairs. As a result, RF and VF showed remarkable differences in all of the similarity measures. It was found that RF user pairs tend to follow same users, and that VF user pairs tend to use same words in their tweets. These results implied that RF user pairs are likely to follow similar users and VF user pairs tend to be interested in the same contents.

After that, we built the models to classify user pairs into RF or VF using the similarity measures. These models showed high performance to distinguish between RF and VF (their F-measures are larger than 0.80). Moreover, we found that the measures in neighborhood similarity are more important for these models than those in content similarity. In our expectations, these classification models can be applied to privacy-protection techniques hiding users' personal information from strangers and alter strategies of recommendations of advertisement to users based on relationship types of users' connections. We believe that this study contributes to not only social science but also computer science, because our findings can be applied to the design of platforms of SNSs.

References

[1] S. Vallor, "Flourishing on facebook: virtue friendship & new social media," *Ethics and Information technology*, vol. 14, no. 3, pp. 185–199, 2012.

[2] K. Subrahmanyam and D. Smahel, *Digital youth: The role of media in development*. Springer Science & Business Media, 2010.

[3] N. K. Baym, Y. B. Zhang, and M.-C. Lin, "Social interactions across media: Interpersonal communication on the internet, telephone and face-to-face," *New Media & Society*, vol. 6, no. 3, pp. 299–318, 2004.

[4] D. Smahel, B. B. Brown, and L. Blinka, "Associations between online friendship and internet addiction among adolescents and emerging adults." *Developmental psychology*, vol. 48, no. 2, p. 381, 2012.

[5] J. H. Søraker, "How shall i compare thee? comparing the prudential value of actual virtual friendship," *Ethics and Information technology*, vol. 14, no. 3, pp. 209–219, 2012.

[6] Y. Kim, F. Natali, F. Zhu, and E. Lim, "Investigating the influence of offline friendship on twitter networking behaviors," in *System Sciences (HICSS), 2016 49th Hawaii International Conference on*. IEEE, 2016, pp. 736–745.

[7] D. K.-S. Chan and G. H.-L. Cheng, "A comparison of offline and online friendship qualities at different stages of relationship development," *Journal of Social and Personal Relationships*, vol. 21, no. 3, pp. 305–320, 2004.

[8] G. H.-Y. Chan and T. W. Lo, "Do friendship and intimacy in virtual communications exist? an investigation of online friendship and intimacy in the context of hidden youth in hong kong," *Revista de Cercetare si Interventie Sociala*, vol. 47, p. 117, 2014.

[9] W. Bülow and C. Felix, "On friendship between online equals," *Philosophy & Technology*, vol. 29, no. 1, pp. 21–34, 2016.

[10] H. Wang and B. Wellman, "Social connectivity in america: Changes in adult friendship network size from 2002 to 2007," *American Behavioral Scientist*, vol. 53, no. 8, pp. 1148–1169, 2010.

[11] G. S. Mesch and I. Talmud, "Similarity and the quality of online and offline social relationships among adolescents in israel," *Journal of Research on Adolescence*, vol. 17, no. 2, pp. 455–465, 2007.

[12] D. Lu, J. Sang, Z. Chen, M. Xu, and T. Mei, "Who are your "real" friends: Analyzing and distinguishing between offline and online friendships from social multimedia data," *IEEE Transactions on Multimedia*, vol. 19, no. 6, pp. 1299–1313, 2017.

[13] D. Zinoviev and V. Duong, "Toward understanding friendship in online social networks," *arXiv preprint arXiv:0902.4658*, 2009.

[14] M. L. Antheunis, P. M. Valkenburg, and J. Peter, "The quality of online, offline, and mixed-mode friendships among users of a social networking site," *Cyberpsychology: Journal of Psychosocial Research on Cyberspace*, vol. 6, no. 3, 2012.

[15] V. M. Buote, E. Wood, and M. Pratt, "Exploring similarities and differences between online and offline friendships: The role of attachment style," *Computers in Human Behavior*, vol. 25, no. 2, pp. 560–567, 2009.

[16] K. Subrahmanyam and P. Greenfield, "Online communication and adolescent relationships," *The future of children*, vol. 18, no. 1, pp. 119–146, 2008.

[17] X. Ma, G. Chen, and J. Xiao, "Analysis of an online health social network," in *Proceedings of the 1st ACM international health informatics symposium*. ACM, 2010, pp. 297–306.

[18] N. Japkowicz, "Learning from imbalanced data sets: a comparison of various strategies," in *AAAI workshop on learning from imbalanced data sets*, vol. 68. Menlo Park, CA, 2000, pp. 10–15.

[19] H. Akaike, "A new look at the statistical model identification," *IEEE transactions on automatic control*, vol. 19, no. 6, pp. 716–723, 1974.

[20] G. Gorrell and K. Bontcheva, "Classifying twitter favorites: like, bookmark, or thanks?" *Journal of the Association for Information Science and Technology*, vol. 67, no. 1, pp. 17–25, 2016.