Classification of Twitter Follow Links Based on the Followers' Intention[*]

Hikaru Takemura[†] takemura@dl.kuis.kyoto-u.ac.jp
Atsushi Tanaka[‡] a.tanaka@mbs.co.jp
Keishi Tajima tajima@i.kyoto-u.ac.jp
Graduate School of Informatics, Kyoto University, Sakyo, Kyoto 606-8501 Japan

ABSTRACT

In Twitter, a user follows other users for various purposes, e.g., for information gathering, for personal communication, and for reading a chat by celebrities. As a result, the intention of followers behind follow links is different from case to case. We classify follow links in Twitter based on the followers' intention along with three classification axes: user-orientation, content-orientation, and mutuality. The combination of these three axes covers most major types of followers' intention found in Twitter. We collected 1760 Twitter follow links through a questionnaire and we found that (1) user-orientation and content-orientation have weak positive correlation, and user-orientation also has weak positive correlation with mutuality, but content-orientation has no correlation with mutuality, (2) content-oriented follows are more frequent than user-oriented follows even among communication-oriented users, and (3) the users have no clear intention for more than 20% of their follow links. We then constructed classifiers with various features of the followee, the follower, and their relationship. We also developed a method of classifying "lists" in Twitter into information lists and community lists, and used the types of lists including the followee. Our experiments show that (1) no single property is a prominent discriminator, and (2) classification accuracy for follow links by information-gathering users are higher than that for communication-oriented users.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: Information Search and Retrieval

Keywords

social network; link classification; Twitter lists

1. INTRODUCTION

Twitter is a micro-blogging service where users can post short messages, which are called tweets. The most distinctive feature of Twitter is its mechanism of *"follow"*. If some user follows other users, all tweets by them are retrieved in real time, and are shown to the follower user in a list sorted in the reverse chronological order. This list is called the *"timeline"* of the follower user. In this paper, we call those who follow some user *followeRs* of the user, and we call those whom some user follows *followeEs* of the user. (We avoid using the word "friends" to mean followeEs because we use it in the ordinary meaning, i.e., to mean "personal friends.")

Unlike user-linking functions in other social network services (SNSs), following to a user does not require the permission by the followeE, and does not necessarily imply reciprocal relationship. The mechanism of follow makes Twitter more flexible than ordinary blogs or discussion forums. Each Twitter user can organize tweets by others into one's timeline, while posts to blogs and discussion forums are shown with fixed structure only on the corresponding Web pages.

The mechanism of follow is, to some extent, similar to RSS (RDF Site Summary) subscription. Many RSS feed readers allow a user to organize feeds from various RSS sources into a timeline as in Twitter. RSS is, however, cannot be used as a SNS. In Twitter, a user can see others' followeEs, and add them to his followeEs to extend his network, but RSS users cannot see what RSS sources other users are subscribing.

Although Twitter is different from these existing services as explained above, one can use Twitter as light-weighted substitutes for them [9]. By using Twitter, one can follow news media accounts for gathering information as with RSS readers, can follow celebrities for reading their chats as in blog services, can follow personal friends for daily conversation as in SNSs, and can follow technical experts for participating discussions led by them as in discussion forums.

1.1 Why is Link Classification Important?

As Twitter is used in various purposes as explained above, the intention of followers behind follow links differs from case to case. Information on their intention is useful in various applications, e.g., user recommendation [6, 28], link prediction [26, 7, 3], and information diffusion analysis [18, 10].

For example, some studies proposed Twitter user recommendation methods based on topic similarity between users [6]. If the target user follows some users because he is interested in the information in their tweets, we should recommend other users whose tweets are similar. On the other hand, if the target user follows some users because he is a fan

[*] This work was supported by JSPS Kakenhi 26540163.

[†] Currently at Yahoo Japan Corporation

[‡] Currently at Mainichi Broadcasting System, Inc.

SAC'15, April 13-17, 2015, Salamanca, Spain
Copyright is held by the owner/authors. Publication rights licensed to ACM.
ACM 978-1-4503-3196-8/15/04...$15.00
http://dx.doi.org/10.1145/2695664.2695940

of them, or if it is because they are friends, then it does not make sense to recommend other users with similar tweets.

On the other hand, some studies proposed Twitter user recommendation methods based on collaborative filtering [6, 2]. In collaborative filtering, we recommend users followed by users who are similar to the target user. Two users are similar iff they share many followeEs. For example, suppose the target user follows many users who are his personal friends, and there is a user who also follows many of them. That user must be in the same circle of friends as the target user, and the collaborative filtering method recommends users followed by that similar user. In this case, if the recommended user is a friend of the similar user, this recommendation is reasonable. However, if the recommended user is a famous actress, and the similar user follows her because he is a fan, this recommendation may not be reasonable. The target user and the similar user are similar in their friend links but may not be similar in their fan links.

Recently, Backstrom and Leskovec [3] proposed a method of predicting and recommending links in social networks based on supervised random walks. In order to estimate the probabilities that the target user will follow other users, they first assign various transition probabilities to existing follow links, run a random walk starting from the target user, and compute the probability that the random walk visits each user. For example, suppose the target user A follows A's close friend B, and B follows B's close friend C. We then assign high transition probabilities to the links A-B and B-C, and the probability that a random walk starting from A visits C is high. In this case, A is, in fact, likely to follow C. On the other hand, if A follows B because B is an authoritative source on some topic, and C is a close friend of B, then the probability that A will follow C is not as high as in the previous case. Models based on a graph with a single type of edges, however, cannot distinguish these cases.

1.2 Our Classification Scheme and Method

As shown in the examples above, classification of follow links according to the followers' intention is important in various applications. There are many ways to classify them, but we adopt a classification scheme we recently proposed in [22]. It consists of the following three classification axes:

1. *user-orientation*,
2. *content-orientation*, and
3. *mutuality*.

User-orientation means whether the follower is interested in the followeE user itself, and content-orientation means whether the follower is interested in specific information in tweets. Notice that these two factors are not exclusive. Mutuality means whether the follower expects to have mutual communication with the followeE. The meanings of these axes are explained in Section 3 in more detail. By combining these three axes, we can classify links into 8 types, which can represent most major types of follow links found in real Twitter data and the literature [11, 9, 13, 27, 25].

We conducted experiments on classification of Twitter follow links by using SVMs (support vector machines) and decision trees. We used properties of followeEs, properties of followeRs, and properties of their relationship. We also used information on "lists," which is a function of Twitter for grouping one's followeEs. This paper reports our several findings in the experiments.

2. RELATED WORK

Recently, Aiello et al. [1] has proposed a classification scheme for social interactions. They classify social interactions into three types: status exchange, knowledge exchange, and social support. Although these three types and our three axes do not perfectly coincide, they share some ideas. For example, their concept of knowledge exchange is related to our concept of content-orientation. They, however, classify each social interaction, while we classify each follow link, through which various interactions are exchanged over time. Moreover, their three types are exclusive categories, while our three axes are independent classification axes.

There have also been studies on Twitter user classification based on their purposes. Kwak et al. [13] reported that Twitter is used not only as SNSs but also as a media for disseminating/gathering information. Java et al. [9] proposed a method of determining purposes of Twitter users, i.e., why each user uses Twitter. There are also studies on the detection of a specific type of users, e.g., detecting spam users [15, 21], finding topic-specific influential users [24], or classifying users into human users, bots, and cyborgs [5]. These studies classify users, while we classify each follow link. In Twitter, two follow links by a single user, or two follow links to a single user, may have different link types.

There are also studies on tweet classification. Sankaranarayanan et al. [19] proposed the classification into those including news and others, and Sriram et al. [20] classify tweets into the following five categories: news, events, opinions, deals, and private messages. Tweet classification is more related to the purpose of the users posting the tweets, rather than to the intention of the followers.

There have been some studies on classification of links in SNSs, e.g., classification into positive links and negative links [16, 17, 12]. Their target is, however, Wikipedia and message boards like Slashdot, where negative links are frequent. In our survey, we found no negative links in Twitter.

Chen et al. [4] and Hopcroft et al. [8] studied the problem of predicting reciprocity between Twitter users. In [4], reciprocity is defined based on the number of exchanged messages, and in [8], it simply means if they follow each other. On the other hand, mutuality in this research is defined based on the subjective intention of the follower. In our definition, even if two users follow each other, if their intention is not to have mutual communication, the follow does not have mutuality. This is often the case because there are many Twitter accounts that automatically follow back to all the followers. On the contrary, even if two users do not exchange many messages, if the purpose of the follower is to have mutual communication, that follow link has mutuality.

3. THREE CLASSIFICATION AXES

In this section, we explain our three axes in more detail.

The two most common purposes of "ordinary" Twitter users are conversation with friends and information gathering from information sources [9, 13]. [9] also reported that users join communities composed of reciprocal follow links either for sharing daily experience or for discussing specific topics with users sharing some common interests.

In Twitter, there are also "elite" users. Wu et al. [25] classifies elite users into celebrities, media, organizations, and prominent bloggers. Followers of celebrities, such as pop idols, usually follow them not because they expect specific

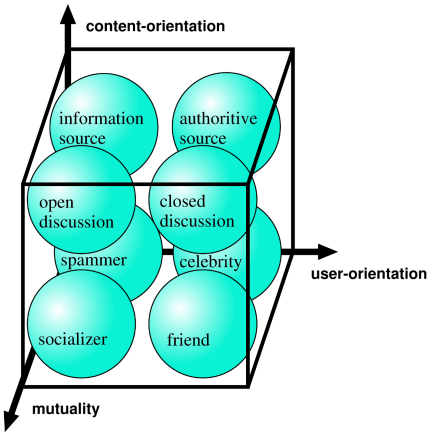

Figure 1: Major link types plotted in our classification space [22].

useful information, but simply because they are fans of them. On the other hand, users who follow news media are usually interested in information in their tweets. Recently, many organizations also have Twitter accounts. Users follow them because they expect information on some specific topics on which the organizations are the authorities. There are also many prominent bloggers who have many followers because their tweets include useful information on specific topics.

There are also special kinds of users. "Spammers" randomly follow many users just for obtaining "follow-backs" and distributing advertisement [15, 21]. "Socializers" also follow many users no matter who they are, just to extend their network. Some information source users also "follow-back" to all the followers in order to keep them. According to our questionnaire to Twitter users, which will be explained later, there are also many follow links without clear reasons.

They are not an exhaustive list, but as long as these major link types are concerned, we can distinguish them by using combination of our three classification axes, i.e., (1) user-orientation, (2) content-orientation, and (3) mutuality.

User-orientation means whether the follower is interested in the followeE user itself. *If a user follows some user and it cannot be replaced with other users whose tweets are very similar, that follow link is user-oriented.* For example, when a fan follows some pop idol, the fan probably does not want to follow another user whose tweets are very similar to the idol. Follow links to friends are also user-oriented.

On the other hand, content-orientation means whether the follower is interested in specific information in the tweets. *If a user follows some user but the original reason of the follow would be lost once the followeE stopped tweeting about a specific topic, that follow link is content-oriented.* For example, if a user follows an expert on Web technologies who often tweets about latest Web technologies, and if the user would have no reason to follow the expert once he stopped tweeting about Web technologies, that follow link is content-oriented. Follow links to news media are also usually content-oriented.

Notice that user-orientation and content-orientation are not exclusive. For example, suppose a user follows `bbcbusiness` because he thinks it is the most reliable. He does not want to switch to other similar news media, and will not follow `bbcbusiness` if it stops tweeting about business news. That link is both user-oriented and content-oriented.

Mutuality means *whether the follower expects to have mutual communication.* Reciprocal follow links do not necessarily mean mutuality as explained in Section 2.

Table 1: Features Used for Follow Link Classification

category	features
(A) followeE	# of followeEs, followeRs, reciprocal follows, lists, reciprocal followeR ratio, reciprocal followeE ratio, proportion of information lists to all lists including it
(B) followeR	# of followeEs, followeRs, reciprocal follows, lists, reciprocal followeR ratio, reciprocal followeE ratio
(C) relationship	reciprocity, number of common lists, frequency of replies by "@", frequency of RT of followeE's tweets by followeR

Figure 1 shows how some link types are plotted in the classification space generated by our three axes. Links to friends are user-oriented and have mutuality, but not content-oriented. Links to celebrities from fans are user-oriented but not content-oriented, and do not have mutuality. Links by spammers or socializers are neither user-oriented nor content-oriented. The purpose of spammers is to send advertisements, so their links do not have mutuality, while socializers expect to have mutual communication. Follow-back links by information sources or links without clear reasons are also neither user-oriented nor content-oriented, and do not have mutuality.

Links to authoritative information sources are both content-oriented and user-oriented, but do not have mutuality. Links to non-authoritative information sources, e.g. a bot tweeting weather reports, are content-oriented but not user-oriented, and do not have mutuality. If a user follows some users in order to join a discussion or a QA forum on a specific topic, these follow links are content-oriented and have mutuality. If the discussion is closed to a specific community, or the QA forum is answered by authoritative people, the links are also user-oriented, but if it is an open discussion or non-authoritative forum, the links are not user-oriented.

Our three axes can classify most major link types as shown above. Note that Figure 1 just shows how some of the major link types are typically plotted, and we do not mean that all follow links are classified into these eight types.

4. FEATURES FOR OUR CLASSIFIERS

We classify links by SVMs and decision trees as explained before. We use three kinds of features listed in Table 1: (A) properties of the followeE, (B) properties of the followeR, and (C) properties related to the relationship between them.

The most basic features extracted from both the followeE and the followeR of the link are: the number of followeEs, the number of followeRs, the number of reciprocal follows, and the number of lists including the user. We also compute *reciprocal followeE ratio* and *reciprocal followeR ratio*, which are defined as below, for both the followeE and the followeR.

$$\begin{aligned}\text{reciprocal followeE ratio} &= \frac{\text{\# of reciprocal follows}}{\text{\# of one's followeE}} \quad \text{and}\\ \text{reciprocal followeR ratio} &= \frac{\text{\# of reciprocal follows}}{\text{\# of one's followeR}}\end{aligned}$$

Notice that the denominator of the former is "followeE" and that of the latter is "followeR." When the denominator is 0, we let the ratio be 0. These two values, together with some other features, can represent the following four cases:

(1) If both ratios are high, there are two cases. If the user has many followeRs, then the user is probably an in-

formation source that follow-backs to all the followeRs. If the user does not have many followeRs, the user is probably communication-oriented user, and most of the followeRs and the followeEs must be friends of the user.

(2) If the mutual followeE ratio is low and only the mutual followeR ratio is high, there are mainly two cases. If the user does not have many followeRs, then the user is probably following some celebrities or information sources, but does not have many followeRs except for his friends. If there are many followeRs, the user may be an information source who follows back to all the followeRs, or a socializer or a spammer who obtains followeRs only by follow-backs.

(3) In contrast, if the mutual followeR ratio is low, and only the mutual followeE ratio is high, the user is probably a celebrity or an information source, who has many followers but who is following other users only for communication.

(4) If both ratios are low, the user is probably a heavy user active in both gathering and publishing information.

Another important information for classifying a follow link is how the followeE of the link is typically followed by others. A useful clue to that information is what kind of lists most often includes the followeE. List is a function of Twitter for grouping followeEs and creating separate timelines for each group. Because users usually group the followeEs based on the intention of following them, lists in Twitter can be regarded as a kind of social tagging to the followeEs.

According to our survey of Twitter data, lists can be classified into *information lists* and *community lists.* Information lists are used for grouping information sources of related topics, and community lists are used for grouping users that belong to a specific community, e.g., the classmates.

Given a follow link to classify, we first collect at most 20 lists including the followeE of the link, and classify them into information lists and community lists. We then compute the proportion of information lists among them and use it as a property of the followeE as shown in (A) in Table 1. If the followeE is mainly included in information lists, the link to the followeE is probably for gathering information, and if it is mainly included in community lists, the link is probably for communication. If a followeE has no list including it, we set it to 1/2, but there was no such a case in our data set.

We developed a method of automatically classifying lists into these two types. According to [13], celebrities or information sources who have more followeRs than followeEs account for only 20% of users. As a result, the ratio of members who have more followeRs than followeEs in an information list is usually far bigger than that in a community list. Based on this observation, for each list, we compute $\overline{ff}(l)$, the average ratio of the number of followeRs and followeEs of members in a list l. We define $\overline{ff}(l)$ as follows:

$$\overline{ff}(l) = \frac{1}{|member(l)|} \sum_{u \in member(l)} ff(u)$$

$$\text{where } \; ff(u) = \begin{cases} 0 & \text{if } |followeR(u)| < |followeE(u)| \\ |followeR(u)| & \text{if } |followeE(u)| = 0 \\ |followeR(u)|/|followeE(u)| & \text{otherwise} \end{cases}$$

where $member(l)$ is the members in l, $followeE(u)$ is the followeEs of the user u, and $followeR(u)$ is the followeRs of u. This value is expected to be large when l is an information list, and is expected to be small if l is a community list.

In order to evaluate the accuracy of this method, we randomly collected lists from Twitter and manually classified them until we obtained 100 information lists and 100 community lists. We excluded 2 information lists and 4 community lists for which we could not obtain necessary information, and computed $\overline{ff}(l)$ for the remaining 98 and 96 lists. The results are shown in Table 2. $\overline{ff}(l)$ is 0 for most community lists, and its median is 0. On the other hand, $\overline{ff}(l)$ for information lists are large, and even the median, 571, is larger than the maximum value for community lists, 471.

Table 2: $\overline{ff}(l)$ of Information and Community Lists

$\overline{ff}(l)$	information	community
average	2,694.8	25.3
standard deviation	5,057.1	79.5
median	571	0
min	0	0
max	21,534	471

Table 3: Precision/Recall for Classification of Lists

	information	community
precision	84.6%	88.8%
recall	89.7%	83.3%

Because the ranges of the values for information lists and community lists overlap, we cannot perfectly separate these two types of lists only by $\overline{ff}(l)$, but by using a threshold value of 4, we can classify lists into the two types with the precision and recall summarized in Table 3.

While types of lists including the followeE is useful, types of lists the followeR owns are not very useful. Many users group only a small fraction of their followeEs into lists, and therefore, even if most lists owned by the followeR are information lists, it does not necessarily imply that most follows by the followeR is for gathering information.

We also use the features related to the relationship between the followeE and the followeR, listed in (C) in Table 1. The most important one is a boolean value representing whether they follow each other. We also use the number of common lists including both of them. If there are many common lists, the followeE and the followeR may be celebrities or authorities related to the same topic. On the other hand, if there are only a few common lists, but the ratio of the common lists to all lists including either of them is high, it means whenever a list includes one of them, it also includes the other. In such cases, these two users probably belong to the same community. The frequency of replies using "@" is also a good indicator of mutuality. We also use the frequency of the followeR's "retweeting" (RT) of tweets by the followeE. According to [23], that frequency is very useful for estimating the similarity between their interests.

5. EXPERIMENTS

5.1 Data Set

Because our classification is based on subjective intention of followeRs, we cannot create a data set by crawling Twitter data. Instead, we used a questionnaire on a crowdsourcing service.[1] We sought Twitter users who have Twitter accounts that are open to public and have more than 40 followeEs, and we obtained 44 applicants. For each of them, we randomly chose 40 followeEs of the user, and ask the user to answer whether each follow link has user-orientation, content-orientation, and mutuality. Table 4 (1) shows the breakdown of the data set, where we see the following facts.

[1] `http://www.lancers.jp/`

Table 4: Breakdown of Data along Each Axes and Breakdown of Data into Eight Classes

(1) the original data set consisting of 1760 follow links by 44 users:

	yes	no
user-oriented	935	825
content-oriented	1120	640
mutuality	263	1497

mutuality		user-oriented		
	= yes	yes	no	total
content-oriented	yes	155	26	181
	no	48	34	82
	total	203	60	263

mutuality		user-oriented		
	= no	yes	no	total
content-oriented	yes	553	386	939
	no	179	379	558
	total	732	765	1497

(2) the data set D_1 including 840 follow links by 21 communication-oriented users:

	yes	no
user-oriented	395	445
content-oriented	431	409
mutuality	165	675

mutuality		user-oriented		
	= yes	yes	no	total
content-oriented	yes	75	19	94
	no	39	32	71
	total	114	51	165

mutuality		user-oriented		
	= no	yes	no	total
content-oriented	yes	224	113	337
	no	57	281	338
	total	281	394	675

(3) the data set D_2 including 920 follow links by 23 information-gathering users:

	yes	no
user-oriented	540	380
content-oriented	689	231
mutuality	98	822

mutuality		user-oriented		
	= yes	yes	no	total
content-oriented	yes	80	7	87
	no	9	2	11
	total	89	9	98

mutuality		user-oriented		
	= no	yes	no	total
content-oriented	yes	329	273	602
	no	122	98	220
	total	451	371	822

(4) the final data set, which was used in the experiments, including 1253 follow links with 507 links removed:

	yes	no
user-oriented	725	528
content-oriented	803	450
mutuality	218	1035

mutuality		user-oriented		
	= yes	yes	no	total
content-oriented	yes	126	23	149
	no	40	29	69
	total	166	52	218

mutuality		user-oriented		
	= no	yes	no	total
content-oriented	yes	410	244	654
	no	149	232	381
	total	559	476	1035

(1) Content-oriented follows are slightly more frequent than user-oriented follows. It is because of the 386 follows with content-orientation, without user-orientation, and without mutuality (typically links to information sources).

(2) Follows with user-orientation and content-orientation but without mutuality are the most frequent among 8 types.

(3) There are unexpected number (379, which is 21.5% of 1760) of links without user-orientation, content-orientation, and mutuality. According to the questionnaire, they are links "without any clear reasons." The users have no clear intention for more than 20% of their follow links.

(4) User-orientation and content-orientation are not exclusive, rather they have weak positive correlation ($r = 0.267$). User-orientation also has weak positive correlation with mutuality ($r = 0.202$), but content-orientation has no correlation with mutuality ($p = 0.058$, Pearson's chi-square test). Note that transitivity does not hold for correlation [14].

In order to see more details, we divided the data set into two data sets. We first divided the 44 applicants into two groups U_1 and U_2. U_1 includes users whose reciprocal followeE ratio is larger or equal to 1/2, and U_2 includes the others. We regard U_1 as communication-oriented users, and U_2 as information-gathering users. Among 44 users, 21 were classified into U_1 and 23 were classified into U_2. We then divided the data set into D_1 consisting of follows by users in U_1, and D_2 consisting of follows by users in U_2.

Table 4 (2) and (3) show the breakdown of the data set D_1 and D_2, respectively, where we see the following facts.

(1) The ratio of content-oriented follows in information-gathering users is higher than that in communication-oriented users. However, content-oriented follows are still more frequent than user-oriented follows even in communication-oriented users. It means that higher ratio of content-oriented follows in the whole data set is not because of a small number of users that follow huge number of information sources.

(2) The ratio of user-oriented follows without content-orientation and with mutuality (typically friend links) is higher in communication-oriented users. On the other hand, the ratio of user-oriented follows without content-orientation and without mutuality (typically celebrity links), and the ratio of content-oriented follows without user-orientation nor mutuality (typically links to information sources) is higher in information-gathering users.

5.2 Accuracy of Classification

Next, we explain our experiments on the link classification. Among the 1,760 follow links in the data set, we excluded 507 links for which we could not obtain all information listed in Table 1, and created a data set for the classification experiments consisting of 1,253 follow links. Table 4 (4) shows the breakdown of this data set.

We tested classification by SVMs (LIBSVM[2] with RBF kernel, which uses one-against-one method for multi-class classification) and by decision trees (scikit-learn[3]). For each of them, we also compared two methods: classification combining the results of three binary classifiers corresponding to the three axes, and classification by a single 8-class classifier. For each method, we run the classifiers with various combinations of feature sets (A), (B), and (C) in Table 1.

All the feature values were normalized to values within [0,1]. For the number of followeEs and followeRs, we first took a logarithm of the value, and then normalized it. The result was evaluated with 10-fold cross validation.

The upper half of Table 5 shows the accuracy of classification into eight classes. It also includes the accuracy by the majority class method for comparison. For three out of four

[2] `http://www.csie.ntu.edu.tw/~cjlin/libsvm/`

[3] `http://scikit-learn.org/stable/modules/tree.html`

Table 5: Classification Accuracy with Various Feature Sets (and Majority Class Method)

features used		A	B	C	A+B	A+C	B+C	A+B+C	majority
three binary SVMs combined		30.97	36.95	33.52	36.95	34.08	**50.20**	42.70	32.72
three binary decision trees combined		25.30	48.36	33.52	43.58	27.29	**54.91**	43.26	32.72
single 8-class SVM		38.07	43.81	32.72	46.93	37.51	43.26	**50.28**	32.72
single 8-class decision tree		28.01	51.32	35.67	49.16	29.93	**55.87**	50.12	32.72
binary SVMs	user-oriented	59.38	66.80	57.38	67.67	60.34	**68.56**	67.52	57.86
	content-oriented	64.41	68.16	64.49	67.60	64.33	**69.27**	**69.27**	64.09
	mutuality	82.28	83.96	85.63	84.12	85.16	**88.87**	87.63	82.60
binary decision trees	user-oriented	57.62	72.79	57.38	67.52	55.95	**73.34**	69.03	57.86
	content-oriented	55.95	74.62	63.93	70.15	56.58	**76.78**	70.31	64.09
	mutuality	76.70	83.96	85.40	81.88	78.93	**87.79**	82.68	82.60

Table 6: Accuracy of Decision Trees for Two User Groups

		A	B	C	A+B	A+C	B+C	A+B+C	majority
three binary decision trees combined	D_1	20.60	36.80	20.42	35.04	21.83	47.36	35.04	28.35
	D_2	30.80	57.96	42.77	49.64	32.41	**61.31**	51.68	38.98
single 8-class decision tree	D_1	20.77	42.25	26.76	41.73	24.47	43.31	43.13	28.35
	D_2	32.85	58.83	42.04	55.77	34.60	**61.46**	57.37	38.98

Table 7: Accuracy of Decision Trees without Each Feature

removed feature	3 binary	8-class
with all	43.26	50.12
(A) followeE		
# of followeEs	42.14	49.24
# of followeRs	42.00	49.40
# of reciprocal follows	42.06	*50.68
# of lists	41.74	*50.44
reciprocal followeR ratio	42.54	49.40
reciprocal followeE ratio	42.38	*51.32
ratio of info-list/comm-list	*44.05	49.24
(B) followeR		
# of followeEs	42.86	47.65
# of followeRs	*43.66	49.80
# of reciprocal follows	43.10	49.40
# of lists	42.78	49.00
reciprocal followeR ratio	42.54	49.24
reciprocal followeE ratio	**40.30**	*51.16
(C) relationship		
reciprocity	42.38	49.24
# of common lists	42.06	*50.20
frequency of @	40.94	**48.76**
frequency of RT	42.94	49.40

Table 8: Accuracy of Decision Trees without Each Feature (without Features of (A) FolloweE)

removed feature	3 binary	8-class
with all	54.91	55.87
(B) followeR		
# of followeEs	54.59	*56.03
# of followeRs	54.59	*55.95
# of reciprocal follows	54.91	*56.03
# of lists	54.75	*55.95
reciprocal followeR ratio	*55.07	*55.95
reciprocal followeE ratio	54.83	*56.26
(C) relationship		
reciprocity	**51.86**	**52.83**
# of common lists	53.23	*56.03
frequency of @	*55.23	*56.34
frequency of RT	*55.23	55.79

methods, the feature set B+C achieves the highest accuracy. It suggests that the type of a link does not depend solely on the type of the followeE, and the type of the followeR is more important. For both SVMs and decision trees, the difference between the accuracy of three binary classifiers and a single 8-class classifier is not significant, which suggests that the classification along the three axes can be done independently. For both the binary approach and the 8-class approach, decision trees works slightly better than SVMs.

The lower half of Table 5 shows the accuracy of binary classifiers for each axis. In all cases, B+C achieves the highest accuracy. The accuracy for mutuality is the highest, and that for user-orientation and content-orientation are similar.

To identify which features are important, we trained and ran decision trees with each properties removed. The result is shown in Table 7. The result shows that the accuracy becomes lowest when we remove reciprocal followeE ratio for the three binary classifier method, and when we remove frequency of @ for the 8-class classifier method. On the other hand, when we remove properties marked with "*" in Table 7, the accuracy was improved. However, the effect of each property is not largely different from each other.

Because the accuracy with properties B+C is better than that with A+B+C, we also ran experiments with removing each properties starting from B+C. The result is shown in Table 8. Also in this result, the effect of each property is not largely different from each other. These results show that no single property is prominent, and these many properties collectively achieve the accuracy of the classification.

We also trained and ran decision trees on the data set D_1 and D_2 separately. Table 6 shows the result. This result shows that the accuracy is far higher with D_2 for both the binary classifier method and the 8-class classifier method. It is probably because of data skewness in D_2 since the accuracy of the majority class method is also higher for D_2.

6. CONCLUSION

In this paper, we report the results of our experiments of Twitter follow link classification into 8 types along the following three axes: user-orientation, content-orientation, and mutuality. Through a questionnaire to 44 Twitter users on a crowdsourcing service, we collected a data set consisting of 1760 follow links (40 links from each of the 44 users). In this data set, we found the following facts.

(1) Content-oriented follows are slightly more frequent than user-oriented follows (even in communication-oriented users) because of many content-oriented follows without user-orientation nor mutuality (typically links to information sources).

(2) Follows with user-orientation and content-orientation but without mutuality are the most frequent among 8 types.

(3) More than 20% of follows are without clear intention.

(4) User-orientation and content-orientation are not exclusive, rather have weak positive correlation. User-orientation also has weak positive correlation with mutuality, while content-orientation has no correlation with mutuality.

(5) User-oriented follows without content-orientation and with mutuality (typically friend links) are more frequent in communication-oriented users. On the contrary, user-oriented follows without content-orientation nor mutuality (typically celebrity links), and content-oriented follows without user-orientation nor mutuality (typically information source links) are more frequent in information-gathering users.

We also tested link classification by SVMs and decision trees on this data set. We used various features of followeEs, followeRs, and their relationship. We also proposed a method of classifying lists into information lists and community lists, and we used the ratio of these two types of lists including the followeE. The results show the following facts.

(1) Link types do not solely depend on the followeEs.

(2) No single property can be a prominent discriminator.

(3) Classification accuracy for follows by information-gathering users is higher probably because of data skewness.

7. REFERENCES

[1] L. M. Aiello, R. Schifanella, and B. State. Reading the source code of social ties. In *Proc. of ACM WebSci*, pages 139–148, 2014.

[2] M. Armentano, D. Godoy, and A. Amandi. Topology-based recommendation of users in micro-blogging communities. *J. Comput. Sci. Technol.*, 27(3):624–634, 2012.

[3] L. Backstrom and J. Leskovec. Supervised random walks: predicting and recommending links in social networks. In *Proc. of WSDM*, pages 635–644, 2011.

[4] J. Cheng, D. M. Romero, B. Meeder, and J. M. Kleinberg. Predicting reciprocity in social networks. In *Proc. of IEEE SocialCom*, pages 49–56, Oct. 2011.

[5] Z. Chu, S. Gianvecchio, H. Wang, and S. Jajodia. Who is tweeting on twitter: human, bot, or cyborg? In *Proc. of ACSAC*, pages 21–30, 2010.

[6] J. Hannon, M. Bennett, and B. Smyth. Recommending twitter users to follow using content and collaborative filtering approaches. In *Proc. of ACM RecSys*, pages 199–206, 2010.

[7] M. Hasan and M. Zaki. A survey of link prediction in social networks. In *Social Network Data Analytics*, pages 243–275. Springer, 2011.

[8] J. E. Hopcroft, T. Lou, and J. Tang. Who will follow you back?: Reciprocal relationship prediction. In *Proc. of CIKM*, pages 1137–1146, Oct. 2011.

[9] A. Java, X. Song, T. Finin, and B. Tseng. Why we Twitter: Understanding microblogging usage and communities. *Proc. of KDD*, pages 56–65, 2007.

[10] M. Kim, D. Newth, and P. Christen. Modeling dynamics of meta-populations with a probabilistic approach: Global diffusion in social media. In *Proc. of CIKM*, pages 489–498, 2013.

[11] B. Krishnamurthy, P. Gill, and M. Arlitt. A few chirps about twitter. In *Proc. of WOSP*, pages 19–24, 2008.

[12] J. Kunegis, A. Lommatzsch, and C. Bauckhage. The slashdot zoo: mining a social network with negative edges. In *Proc. of WWW*, pages 741–750, 2009.

[13] H. Kwak, C. Lee, H. Park, and S. Moon. What is twitter, a social network or a news media? In *Proc. of WWW*, pages 591–600, 2010.

[14] E. Langford, N. Schwertman, and M. Owens. Is the property of being positively correlated transitive? *The American Statistician*, 55(4):322–325, Jan. 2012.

[15] K. Lee, J. Caverlee, and S. Webb. The social honeypot project: protecting online communities from spammers. In *Proc. of WWW*, pages 1139–1140, 2010.

[16] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Predicting positive and negative links in online social networks. In *Proc. of WWW*, pages 641–650, 2010.

[17] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Signed networks in social media. In *Proc. of CHI*, pages 1361–1370, 2010.

[18] D. M. Romero, B. Meeder, and J. M. Kleinberg. Differences in the mechanics of information diffusion across topics: idioms, political hashtags, and complex contagion on twitter. In *Proc. of WWW*, pages 695–704, 2011.

[19] J. Sankaranarayanan, H. Samet, B. E. Teitler, M. D. Lieberman, and J. Sperling. TwitterStand: News in tweets. In *Proc. of ACM GIS*, pages 42–51, Nov. 2009.

[20] B. Sriram, D. Fuhry, E. Demir, H. Ferhatosmanoglu, and M. Demirbas. Short text classification in Twitter to improve information filtering. In *Proc. of SIGIR*, pages 841–842, Jul. 2010.

[21] E. Tan, L. Guo, S. Chen, X. Zhang, and Y. Zhao. Unik: Unsupervised social network spam detection. In *Proc. of CIKM*, pages 479–488, 2013.

[22] A. Tanaka, H. Takemura, and K. Tajima. Why you follow: A classification scheme for Twitter follow links. In *Proc. of ACM Hypertext*, pages 324–326, 2014.

[23] M. J. Welch, U. Schonfeld, D. He, and J. Cho. Topical semantics of twitter links. In *Proc. of WSDM*, pages 327–336, 2011.

[24] J. Weng, E.-P. Lim, J. Jiang, and Q. He. TwitterRank: Finding topic-sensitive influential twitterers. In *Proc. of WSDM*, pages 261–270, 2010.

[25] S. Wu, J. M. Hofman, W. A. Mason, and D. J. Watts. Who says what to whom on twitter. In *Proc. of WWW*, pages 705–714, 2011.

[26] D. Yin, L. Hong, X. Xiong, and B. D. Davison. Link formation analysis in microblogs. In *Proc. of SIGIR*, pages 1235–1236, 2011.

[27] D. Zhao and M. B. Rosson. How and why people twitter: the role that micro-blogging plays in informal communication at work. In *Proc. of ACM GROUP*, pages 243–252, 2009.

[28] G. Zhao, M. L. Lee, W. Hsu, W. Chen, and H. Hu. Community-based user recommendation in uni-directional social networks. In *Proc. of CIKM*, pages 189–198, 2013.