**Master Thesis**

# Twitter User Classification Based on Specificity of their Information Dissemination Target

Supervisor Professor Keishi Tajima

Department of Social Informatics
Graduate School of Informatics
Kyoto University

Hikaru TAKEMURA

February 7, 2014

# Twitter User Classification Based on Specificity of their Information Dissemination Target

Hikaru TAKEMURA

**Abstract**

This guide gives instructions for writing your B.E. or M.E. theses following the standard of the Department of Information Science. The standard includes the structure and format which you must obey on writing your theses.

This guide also explains how to use a LaTeX style file for theses, named `kuisthesis`, with which you can easily produce well-formatted results. Since this guide itself is produced with the style file, it will help you to refer its source file `eguide.tex` as an example.

Note for graduate students: This document is written for students of old graduate school of information science, not for graudate school of informatics. Writers of master thesis belonging to graduate school of informatics must obey rules given by each department.

# 情報発信の対象限定性に基づく Twitter ユーザの分類

竹村 光

**内容梗概**

この手引では，特別研究報告書および修士論文をどのような構成とするか，またどのような形式で作成するかを説明したものである。また，当教室で定めた形式に則った論文を日本語 LaTeX を用いて作成するためのスタイル・ファイルである，kuisthesis の使い方についても説明している。なお，この手引自体も kuisthesis を用い，定められた形式に従って作成されているので，必要に応じてソース・ファイル eguide.tex を参照されたい。

**Twitter User Classification Based on Specificity of their Information Dissemination Target**

# Contents

# Chapter 1 Introduction

With widespread use of social medias, such as blogging services or social network services (SNS), today we have become able to publish information on the Web more easily than previously. Especially recently, microblogging services have been recently growing explosively.

Microblogging services are a new type of services which have both characteristics of blogging services and SNS. In microblogging services, users can post short messages more easily and rapidly than in conventional blogging services or SNS. Microblogging services are not necessarily regarded as a medium for publishing useful information to the public, and many microblog users post messages more casually than in conventional blogging services or SNS. Because of these characteristics of microblogging services, a large number of messages are posted on microblogging services every day, and the messages contain various types of contents, from personal notes or life logs to useful information or discussion on specific topics. Furthermore, among these various types of messages, those describing the current situations of the users posting the messages especially characterize microblogging services. This type of message is far more frequent than in conventional blogging services or SNS, on this account, a large number of messages posted on microblogging services include information on real-time events.

Among many microblogging services, Twitter[1)] is especially growing rapidly. As of 2012 December, Twitter has over 200 million active users in the world[1], and as of June, more than 400 million messages are posted on it per day[2]. In Twitter, users can post short messages with at most 140 characters, which are called tweets. By this limitation, Twitter makes information publishing more easily and rapidly than conventional blogging services or SNSs. The most distinctive feature of Twitter is its mechanism of *"follow"*. In Twitter, if a user follow other users, all tweets by these followee users are retrieved in real time, and are shown in a list sorted in the reverse chronological order, as shown in Figure 1. This list is called the *"timeline"* of the follower users. The mechanism

[1)] `http://twitter.com/`

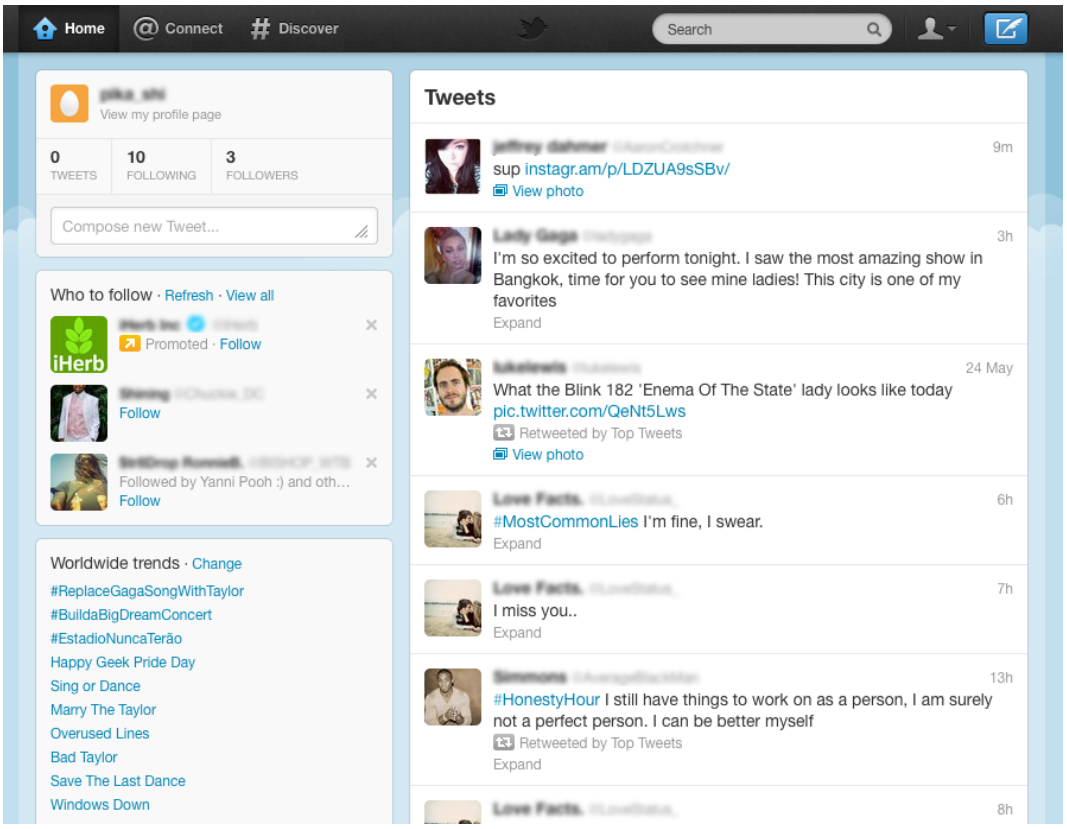


Figure 1: An example of a user's timeline in Twitter

of follow is more casual than user-linking functions in ordinary SNS; it does not require the permission by the followee, and does not necessarily imply reciprocal relationship. Another important function in Twitter is the *"reply"* function, by which a user can post a message as a reply to another user. By using this function, users can use Twitter for conversation, as in instant messaging services.

Twitter has many characteristics of conventional social medias, and because of this, it is used for many purposes. Some users publish information to the public widely like world news, some publish information specified for certain topics, and some communicate with their friends or others. Because of this characteristic of Twitter, it has attracted great attention as a new social media for information publishing.

As explained above, Twitter is used for various purposes. As a result, the wideness of target scope of information publishing varies greatly among users.

Therefore, in this study, we propose a method to classify Twitter users from the point of view of how widely the target scope of their information publishing is, i.e., whether they publish information to the public widely or publish information specified in certain users. In this study, we call the former *"the target specificity is low"*, the latter *"the target specificity is high"*. In this method, we focus on the followers of the user, and we classify him whether his/her followers are consistent in some noticeable characters or not. If his/her followers are consistent in some noticeable characters, it may be estimated that he/she publish information where certain or particular users are interested. On the contrary, if his/her followers are not consistent in any noticeable characters, there is high probability that he/she is followed by a wide variety of users and it may be estimated that he/she publish information where the public, almost all users, is interested.

In addition, in this study, we focus on the Twitter user classified into "the target specificity is high" by the above method, and we propose a method to determine what causes his/her target specificity, i.e., why his/her target scope of information publishing is specified to certain users. In a large number of Twitter users, their target scopes of information publishing are specified to certain users, and the causes of their target specificities vary from user to user. For example, a user who publishes technical information about programming may suppose that he/she publishes information toward unspecified users, but his/her target specificity is considered high because the topic of his/her publishing information is specialized to certain users who are interested in programming. And also, a user who communicates with his/her friends or a user who announces to members of a certain club may suppose that they publish information to the users specified extensionally. So their target specificities are considered high, regardless of contents of their publishing information. In this method, first, we roughly classify causes of the target specificity into two categories: (1) because they publish information specified for certain topics, and (2) because they publish information to the users specified extensionally. And we construct a three-class classifier which determine whether users only belong to category (1), only belong to category (2), or belong to both category (1) and (2). We then determine reasons

that the target specificity is high based on various features which characterize each category for the classifier.

On the Web, it is hard to know what kinds of users each Web page targets to. But in Twitter, we can know what kinds of users each user targets to by access to his/her follow relationships. By exploiting this relationships, we can estimate what kinds of users each user targets to, thus we can classify users based on the target specificity of their information publishing.

Twitter user classification of this study is considered to apply to Twitter search. In current Twitter search, we input a search in the search box and receive messages including the keyword. But with this method, messages in search results has various target scope of information publishing. Thus, it frequency happens that messages of certain target scope I need are buried in many other messages. For example, when a user performs Twitter search with the word "MacBook Air", what kinds of messages he/she needs depends on the situation of that time, e.g., he/she may need public news about MacBook Air, he/she may need technical knowledge about MacBook Air, or he/she may need users' reviews of MacBook Air to refer when he/she buys new one. But in current Twitter search, these information is mixed up in search results. At that time, by using Twitter user classification of this study, we can search messages based on what kinds of users they target to. In this way, we can prevent messages users need from buried in many other messages and help users to find messages they need easily.

The contribution of this paper is summarized as follows.

- We propose a new classification scheme of the target specificity of Twitter users' information publishing.
- We show a method of classifying Twitter users based on above scheme.
- As for users classified into "the target specificity is high", we show a method of determining the reason that the target specificity is high.

The rest of this paper is organized as follows. Next section explains some related work and makes the position of this study clear. Then we define Twitter user's target specificity of information publishing and formulate our problem, and we discuss why target scope of some messages in Twitter are confined to

certain users in Chapter 3. In Chapter 4, we explain the method of classifying Twitter users based on how high the target specificity of their information publishing. In addition, in regards to Twitter users classified into "the target specificity is high" by the above method, we explain the method of determining why their target specificities are high. Then in Chapter 5, we presents the results obtained from experiments we conducted to evaluate the precision of our methods. Chapter 6 concludes the paper.

# Chapter 2 Related Work

With explosive widespread use of microblogging services, Studies about them have recently become frequency performed. There are wide range of contents of studies of microblogging services, e.g., studies of the classification microblogging messages from various point of view[3], studies of ranking microblogging messages by the content relevance and so on[10], or studies focusing on real-time nature of migroblogging services[26, 22].

In this study, we focus on the fact that Twitter is used for a variety of purposes. We attempt to classify Twitter users based on the wideness of target scope of their information publishing, and apply this classification scheme on Twitter search and so on. We explain the following previous studies: studies about the purpose of use of microblogging services in 2.1, studies of classifying Twitter users and measuring their influence in 2.2, studies useful for finding microblogging messages related to only certain Twitter users in 2.3, and studies about Twitter search in 2.4. We make the position of this study clear by introducing these studies.

## 2.1 Purpose of Use of Microblogs

There has been many studies about the purpose of use of microblogging services. Java et al.[15] analyzed the topological and geographical structure of Twitter's social network and attempted to understand the user intentions and community structure in microblogging services. As a result, they found that the main types of user intentions are daily chatter, conversation, sharing information and reporting news. Kwak et al.[17] reported that Twitter is used both as a social network service and as a media for disseminating or gathering information, and in its follower-following topology analysis they have found a non-power-law follower distribution, a short effective diameter, and low reciprocity, which all mark a deviation from known characteristics of human social networks. There are many studies about the purpose of use of microblogging services other than these[29, 31].

And also, Ehrlich et al.[11] conducted a content analysis and examined the

use of public microblogging services (Twitter) for public and private use by comparing internal and external microblogging services (in the workspace). As a result, there were significant differences in content. The internal microblogging services were generally used to solicit technical assistance or as part of a conservation. The external microblogging services were used for status updates and to share general information.

In recent years, Twitter, one of public microblogging services, is often used for not only publishing information to the public but also having a relationship to only a certain community. It is able to be said that this study focuses on the fact that we use Twitter for a variety of purposes like this.

## 2.2 Classification of Twitter Users and Measuring their Influence

There are many other studies focusing on Twitter users, e.g., studies of classification them from various point of view, and studies which measure their influence.

Studies focusing on classification of Twitter users are performed frequently and they have a wide variety of classification schemes, e.g., classification them based on their attributes such as political orientation or ethnicity by leveraging observable information such as the user behavior, network structure, and linguistic content of their posting messages[25], classification them into spam users or not by extracting observable features from the collected candidate spam profiles, e.g., number of friends, text on the profile, age, and so on[18], and classification them into human users, bots, and cyborgs using entropy measures, machine learning, and so on[9]. Bots refer to automated programs posting on Twitter, and cyborgs refer to either bot-assisted humans or human-assisted bots, i.e., interweave characteristics of both humans and bots.

The classification scheme proposed by Yan et al.[30] deeply relates to ours. They proposed methods to classify Twitter users into open accounts. An account is the account with a purpose for advertising or spreading information such as a shop, a singer, a news agency, and so on. And a closed account is the account with a purpose for making friends or communication such as a

user who publishes messages about daily log, feeling show, and so on. This classification scheme is close to ours, but does not coincide with ours because open accounts don't often publish information to the public widely, e.g., an account who publishes very technical information about programming toward unspecified users.

In microblogging services like Twitter or other social network services like Facebook, users correspond to nodes in social network graphs. As well as the classification of users, i.e., nodes in the graphs, the classification of edges, i.e., relationship between a user and his followers, is related to our study. Leskovec et al.[19, 20] classify edges in SNS to positive edges such as friendship, and negative edges such as antagonism. Kunegis et al.[16] also use positive edges and negative edges in Slashdot, a message board service, in order to rank the users. Cheng et al.[8] and Hopcroft et al.[13] studied the problem of predicting reciprocity between two given Twitter users.

There are also studies focusing on measuring influence of Twitter users. Cha et al.[7] analyzed the influence of them by employing three measures that capture different perspectives: indegree, retweets, and mentions. Then they measured the dynamics of influence across topic and time. If target specificity of Twitter users defined in this study is high, there is a high probability that they have a big influence on Twitter, but how target specificity of a user is high does not necessarily coincide with how big his influence is. And Jianshu et al.[28] focused on the problem of identifying influential users of microblogging services.

## 2.3 Find Messages Related to Some Twitter Users

There are also studies useful for finding of microblogging messages related to only a part of Twitter users.

Sakaki et al.[12] proposed a method of monitoring messages in Twitter and detecting occurrences of a specific kind of event in the real world, such as earthquakes or typhoons. They produced a probabilistic spatiotemporal model for the target event that can find the center and the trajectory of the event location. Ikawa et al.[14] attempt to discover the location where a message

was generated by using its textual content. They learned associations between a location and relevant keywords from past messages, and guess where a new message came from. It is able to be said that these studies are useful for finding messages in Twitter related to a certain geographical area.

Sriram et al.[6] proposed approach effectively classifies the message to a predefined set of generic classes such as News, Events, Opinions, Deals, and Private Messages. They proposed to use a small set of domain-specific features extracted from the user's profile and text. Nishida et al.[24] proposed a method that uses data compression for classifying an unseen tweet as being related to an interesting topic or not. It is able to be said that these studies are useful for finding messages in Twitter specified to certain topics.

As mentioned above, there are studies useful for finding microblogging messages related to only a part of Twitter users in various points of view. But these points of view exist in great number, so it is not an effective approach to find these messages from each point of view. Thus in this study, we attempted to measure wideness of target scope of information publishing of a Twitter user in an integrated way. And also, we roughly classify various reasons that target specificity is high into two categories.

## 2.4 Twitter Search

The characteristic of search on microblogging services is different from that of Web search[4] in that search on microblogging services can get information in real time[5] and not only information published by the mass media but also much casual information published by individuals[15]. Thus, a purpose of use of search on microblogging services often becomes a subject of study.

Teevan et al.[27] observed that people use Twitter search to find temporally relevant information, e.g., breaking news, real-time content, and popular trends, and information related to people, e.g., content directed at the searcher, information about people of interest, and general sentiment and opinion. Furthermore, they compared Twitter search with Web search and found that search results on Twitter included more social chatter and social events, and those on the Web included more basic fact and navigation content. Massoudi et al.[21]

proposed a retrieval model for searching messages on microblogging services for a given topic of interest and a dynamic query expansion model for messages retrieval. And Nagamoti et al.[23] described several strategies for ranking messages of microblogging services in a real-time search engine.

As mentioned above, there are many studies about search on microblogging services, and contents of them are greatly various. In this study, we focus on the purpose of use of microblogging services, and attempt to apply it to Twitter Search. It is able to be said that this study also focuses on search on microblogging services just like studies explained above, but there has not been studies based on wideness of target scope of information publishing of a user so far.

# Chapter 3 Target Specificity of Twitter Users

In this chapter, we discuss the target specificity of Twitter users, the measure of to what degree target scope of their information publishing is specified, and define it. Then we also discuss why target scope of their information publishing is specified.

## 3.1 Definition of Target Specificity

In this study, we consider the target specificity of Twitter users, as the measure of how much target scope of their information publishing is specified. More formally, we define the *target specificity* of a Twitter user as to what degree the user set considered to be included in target scope of his/her information publishing is inclining toward a part of all Twitter users, i.e., to what extent this user set deviates from the user set randomly sampled from all Twitter users. In this paper, we express the target specificity of the Twitter user $u$ as $\mathit{TargetSpecificity}(u)$. This formula takes a range of $[0, 1]$.

For example, a user who mainly publishes technical information about programming is supposed to publish information toward programmers. So users who are interested in this information are inclining toward a part of Twitter users. Thus, it may be considered that target specificity of the user is high.

On the contrary, a user who publishes information about world news publishes information useful for the public widely. So the public is supposed to be interested in this information, and the deviation between users who are interested in this information and users randomly sampled from all Twitter users may be very small. Thus, it may be considered that target specificity of the user is low.

As mentioned above, the target specificity of Twitter users is defined as to what extent this user set deviates from the user set randomly sampled from all Twitter users. Thus, the fact that the target specificity of a user is high does not necessarily coincide with the fact that there is high similarity between users who are considered to be included in target scope of his/her information publishing each other. For example, users considered to be included in target scope of

information publishing of a user who publishes information about earthquake in a certain area are consistent in the area they live in, and so his/her target specificity is supposed to be high. But their other characteristics such as age, sex, interests, communities they belong to, and so on vary from user to user. Thus it is not be able to be said that they have high similarity each other. In other words, even if there are various types of users in target scope of his/her information publishing, we consider that the target specificity of the user is high as long as the majority of users in the target scope are consistent in at least one attribute.

In this paper, we determine a threshold $\delta$. If the target specificity of a user is higher than $\delta$, we call him/her a *target user*, and if lower, we call him/her a *non target user*. More formally, we define them as follows:

$$\begin{cases} u \text{ is a } target\ user, & \text{if } \mathit{TargetSpecificity}(u) > \delta \\ u \text{ is a } non\ target\ user, & \text{otherwise.} \end{cases}$$

## 3.2 Why Target Specificity is High

In this subchapter, we discuss what causes the target specificity of a Twitter user, i.e., why target scope of their information publishing is specified. As a result of our analysis, this is roughly classified into two causes: (1) topics of information and (2) target users. We discuss their two causes of the target specificity in the follow.

**(1) Specified topics of information extensionally**

The first cause of the target specificity of a Twitter user is because he/she publishes information specified to a few topics extensionally whether or not he/she specifies target users of his/her information publishing. For example, a user who mainly publishes technical information about programming is supposed to publish information to unspecified people, but it may be considered that target scope of his/her information publishing is specified because he/she specifies the topic of information: programming. Furthermore, a user who mainly publishes information about a certain conference is supposed to pub-

lish information to the users who attend the conference or are interested in it. Because of this, it may be considered that target scope of his/her information publishing is specified.

The way to specify topics of information is roughly classified into two cases. In the first case, a user specifies topics based on demographic data such as age, settled areas, sex, occupation, career, and so on. It is able to be said that a user who publishes weather information in a certain area specifies topics based on demographic data. In the second case, a user specifies topics based on psychographic data such as taste, hobby, values, and so on. It is able to be said that a user who publishes information about cooking specifies topics based on psychographic data.

**(2) Specified target users extensionally**

The second cause of the target specificity of a Twitter user is because he/she publishes information specified to some users extensionally whether or not he/she specifies topics of his/her publishing information. For example, a user who communicates with his/her friends publishes various contents of information, but it may be considered that target scope of his/her information publishing is specified because he/she specifies target users extensionally, i.e., he/she publishes information to the closed users: his/her friends. Furthermore, a user who mainly gets in touch with members of a certain club publishes information to the closed users specified extensionally: the club members. Because of this, it may be considered that target scope of his/her information publishing is specified.

Sometimes, both (1) and (2) cause the target specificity of a Twitter user simultaneously. For example, a user who publishes information to the members of the artist's fan club publishes information specified not only target users of his/her information publishing: the members of artist's fan club, but also the topic of information: the latest news about the artist and so on. Furthermore, it is also true in case of the user who notifies students of a certain university of the news toward them because he/she publishes information specified target users of his/her information publishing: students of the university, and the topics of information the news toward them. And also, some users use Twitter for the

both purpose of publication information of a certain topic and communication with their friends. It is able to be said that such users are also an example of the case that both (1) and (2) cause the target specificity of a Twitter user simultaneously.

# Chapter 4 Classification Methods

In this chapter, we explain the method of classifying Twitter users based on the target specificity of their information publishing defined in Chapter 3. In addition, in regards to Twitter users classified into "the target specificity is high" by the above method, we explain the method of determining why their target specificities are high based on the result of our analysis mentioned in 3.2.

Figure 2 shows the overall flow of our methods. First, we apply the method explained in 4.1 to Twitter users and classify them into two categories: (a) the target specificity is high, and (b) the target specificity is low. Second, we determine why their target specificities are high and apply the method explained in 4.2 to a set of users classified into "the target specificity is high" by the above method. Then we classify them into three categories: (1) they publish information specified for certain topics, (2) they publish information to the users specified extensionally, and (3) in the cause of both (1) and (2).

## 4.1 Classifying Users Based on Target Specificity

In this subchapter, we explain the method of classifying Twitter users based on the target specificity of their information publishing.

### 4.1.1 Assumptions and Classification Algorithm

In this study, we assume that a follower set of a Twitter user is the user set randomly sampled from users included in target scope of his/her information publishing. Thus, we focus on the follower set of a user we intend to classify.

A user who publishes information to the public widely, such as a user who publishes information about world news, is supposed to be followed by various types of users. On the contrary, followers of a user who publishes information specified in certain users are supposed to be consistent in a certain noticeable character. For example, a user who publishes technical information about programming is supposed to be mainly followed by programmers, and a user who communicates with his/her friends is supposed to be mainly followed by his/her friends. Based on the above, we classify a user whether his/her followers are consistent in a certain noticeable character and are difficult to suppose to be

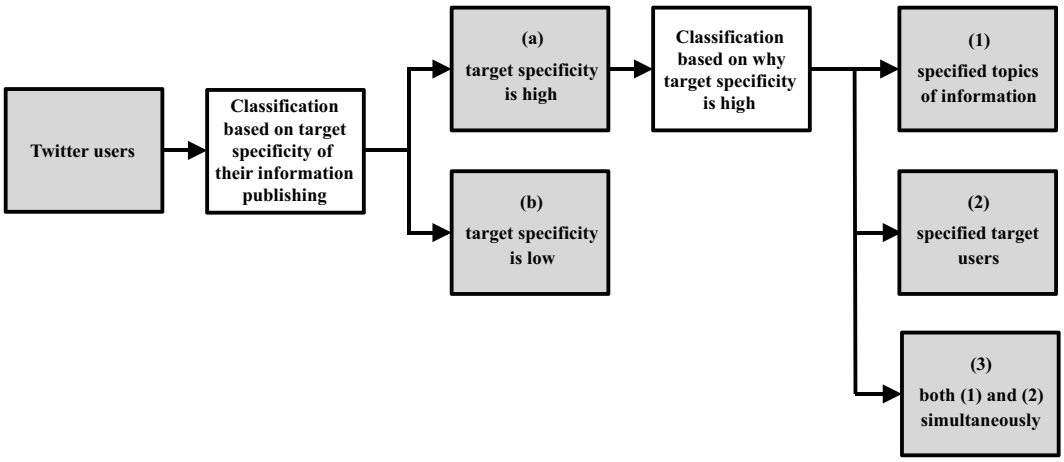


Figure 2: Overall flow of our methods

randomly sampled from all Twitter users, or his/her followers are not consistent in any noticeable character. Figure 3 (a) shows the case of followers of a user being consistent in the character noticeable $A$. In such a case, his/her followers are supposed to incline toward a part of all Twitter users, thus we consider that the more consistent his/her followers are in a certain noticeable character, the higher his/her target specificity is.

In addition to this parameter: whether followers of a user are consistent in a certain noticeable character or not, we consider whether his/her follower set are covered with consistency subsets which cover intermediately widely. Here, *a consistency subset* denotes a subset which have consistency in a certain noticeable character. As show in Figure 3 (b), in regard to followers of a user, when half of them are consistent in the noticeable character $A$ and the others are consistent in the character noticeable $B$, it cannot be said that they are consistent in one noticeable character, but his/her follower set are covered with two consistency subsets which cover intermediately widely. In such a case, we consider that his/her target specificity is high.

These two parameters are summarized as follows.

**(a)** The more consistent followers of a user are in a certain noticeable character, the higher his/her target specificity is.

**(b)** The more covered his/her follower set is with consistency subsets which cover intermediately widely, the higher his/her target specificity is.

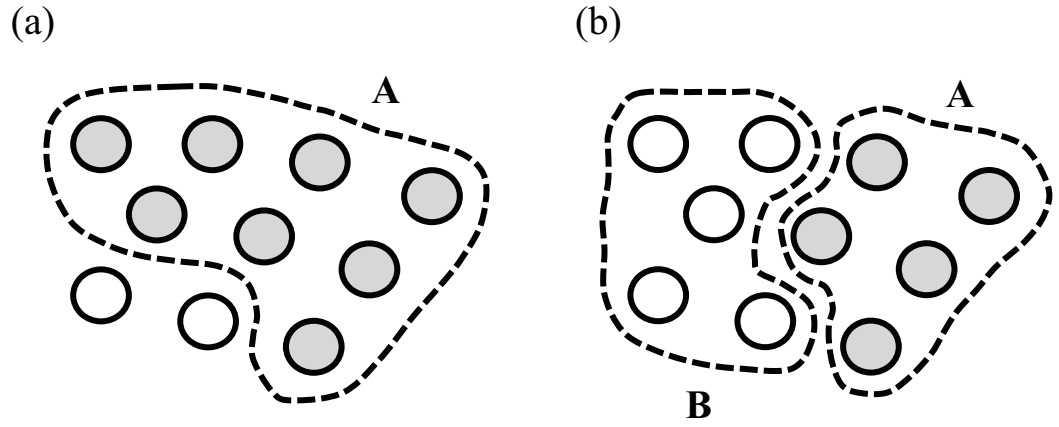


Figure 3: Two examples of high target specificity

The target specificity of a user is quite high when his/her followers are consistent in one noticeable character, and it becomes lower as they do not become consistent in any noticeable character.

Next, we explain the algorithm of computing a score of the target specificity of a user $u$.

First, we collect all consistency subsets included in $u$'s follower set $F_u$. Then, in regard to each subset $S_{F_uc}$ which is consistent in the character $c$ included in $F_u$, we compute $SubsetScore(S_{F_uc})$ which denotes to what degree users in $S_{F_uc}$ are consistent in $c$. We will propose two models which compute $SubsetScore(S_{F_uc})$ in 4.1.2.

Second, in descending order of $SubsetScore(S_{F_uc})$, we give this score to each follower $f$ in $S_{F_uc}$ as $UserScore_{att}(f)$, where $att$ means a attribute measuring consistency, and we explain two attributes in 4.1.3. Here, we do not give a score to $f$ if he/she already has a score. Then, we repeat this step over all $SubsetScore(S_{F_uc})$. In regard to users who are not given a score after the above repeat, we set 0 to them. In other words, in regard to each $f$ of $u$, we give $UserScore_{att}(f)$ to the largest $SubsetScore(S_{F_uc})$ of $S_{F_uc}$ in which $f$ is included as follows:

$$UserScore_{att}(f) = \max_{S_{F_uc} \in F_u} \{SubsetScore(S_{F_uc}) |\ f \in S_{F_uc}\}.$$

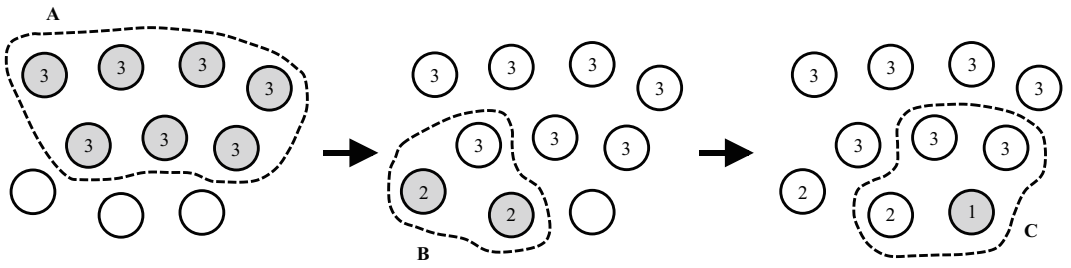


Figure 4: A case of the algorithm flow

Then, we take average of all $UserScore_{att}(f)$ for $SpecificityScore_{att}(u)$, a score of the target specificity of $u$, as follows:

$$SpecificityScore_{att}(u) = \frac{1}{|F_u|} \sum_{f \in F_u} UserScore_{att}(f).$$

For example, suppose the case shown in Figure 4. We assume that a follower set of a user $u$ have three consistency subsets: $A$, $B$, and $C$, and $SubsetScore(A)$, $SubsetScore(B)$, and $SubsetScore(C)$ are 3, 2, and 1 respectively. In descending order of $SubsetScore(S_{F_u c})$, i.e., in order of $A$, $B$, and $C$, we give scores to each follower $f$ as $UserScore_{att}(f)$ as shown in Figure 4. Finally, we compute average of $UserScore_{att}(f)$ and take 2.6 for $SpecificityScore_{att}(u)$.

This is how we compute a score of the target specificity of $u$. Below is a summary of the algorithm:

**1.** we collect all consistency subsets including the follower set,

**2.** we score each consistency subset based on to what degree users in it are consistent in a certain noticeable character

**3.** in descending order of the above score, we repeatedly give it to each user in the subset, and

**4.** we take average of scores for a score of the target specificity.

### 4.1.2 Scoring Models of Consistency Subsets

In this subsubchapter, we explain a couple of models: (i) the probabilistic model and (ii) the subtracting model which compute $SubsetScore(S_{F_u c})$ mantioned in 4.1.1. We explain these two models in the follow.

**(i) Probabilistic Model**

In this model, in regard to the consistency subset $S_{F_u c}$ which is consistent in the character $c$ in the follower set $F_u$, we consider that how low the probability that the user set of the same size as $F_u$ randomly sampled from the set of all Twitter users includes the subset which is consistent in $c$ and whose size is $|S_{F_u c}|$ and over. The lowness of the probability means that $S_{F_u}$ is inclining toward a part of the set of all Twitter users, so it is able to be said that $S_{F_u c}$ has consistency in a noticeable character if the probability is low. Thus, the lower the probability is, the higher score we give. On the contrary, if the probability is not so low, the deviation between $S_{F_u c}$ and the user set randomly sampled from all Twitter users may be small, and it is difficult to say that $S_{F_u c}$ has consistency in a noticeable character. Thus, we give a low score in this case.

In addition to this parameter: how low the probability mentioned above is, we consider that at what rate $S_{F_u c}$ covers with $F_u$. The more covered $F_u$ is with $S_{F_u c}$, the higher score we give. Based on these two parameters, we compute $\mathit{SubsetScore}(S_{F_u c})$.

These two parameters are summarized as follows.

- The lower the probability that the user set of the same size as $F_u$ randomly sampled from the set of all Twitter users includes the subset which is consistent in the same character as the character of $S_{F_u c}$ and whose size is $|S_{F_u c}|$ and over, the higher score we give.
- The more covered $F_u$ is with $S_{F_u c}$, the higher score we give.

Then, we define this model more formally. First, we compute $P(u)$, the probability that the user set of the size of $|F_u|$ randomly sampled from the set of all Twitter users includes the subset which is consistent in $c$ and whose size is $|S_{F_u c}|$, by the formula below:

$$P(S_{F_u c}) = \int_{|S_{F_u c}|}^{n} \binom{n}{x} p^x (1-p)^{n-x} dx, \qquad n = |F_u|, \qquad p = \frac{|S_A|}{|A|}$$

where $A$ is the set of all Twitter users, and $S_A$ is the subset which is consistent in the character included in $A$. The parameter $x$ in the above formula is based on a binomial distribution.

In addition, by adding the covering rate of $S_{F_uc}$ to $F_u$ to the above parameter, we compute $SubsetScore(S_{F_uc})$. However, there is some possibility that $P(S_{F_uc})$ is excessively low because $S_{F_uc}$ has a remarkable character, e.g., only a few users of all users have the character, in spite of the size of $S_{F_uc}$ is very small. In this model, we expect that the consistency subset covers the follower set at least intermediately widely. So in such a case, we do not give any scores to $S_{F_uc}$. More specifically, we determine a threshold $\gamma$ which cuts down the above case.

Then, we compute $SubsetScore(S_{F_uc})$ by the formula below:

$$SubsetScore(S_{F_uc}) = \begin{cases} \dfrac{|S_{F_uc}|}{|F_u|}\log(1 - P(S_{F_uc})) & \text{if } \dfrac{|S_{F_uc}|}{|F_u|} > \gamma, \\ 0 & \text{otherwise.} \end{cases}$$

This is how we compute $SubsetScore(S_{F_uc})$ by using probabilistic technique.

**(ii) Subtracting Model**

In this model, in regard to the consistency subset $S_{F_uc}$, we consider a covering rate of $S_{F_uc}$ to $F_u$ in comparison to a covering rate of the subset which is consistent in the $c$ to the set of all Twitter users. We call the former *a local rate* and the latter *a global rate*. The fact that a local rate is high and a global rate is low means $S_{F_uc}$ has consistency in a noticeable character and is including toward a part of the set of all Twitter users. If a local rate is low or a global rate is high, it is difficult to say that $S_{F_uc}$ has consistency in a noticeable character. Based on the above, we compute $SubsetScore(S_{F_uc})$ by the formula below:

$$SubsetScore(S_{F_uc}) = \max\{\frac{|S_{F_uc}|}{|F_u|} - \frac{|S_Ac|}{|A|}, 0\}.$$

In summary, we give a high score in the case of having the two features simultaneously as follows:

- a covering rate of $S_{F_uc}$ to $F_u$ is high, and
- a covering rate of the subset which is consistent in the same character as the character of $S_{F_uc}$ to the set of all Twitter users is low.

This is how we compute $SubsetScore(S_{F_uc})$by using subtracting technique.

### 4.1.3 Attributes for Extracting Consistency Subsets

In this subsubchapter, we explain a couple of attributes measuring consistency: (i) common terms in profiles and location information, and (ii) common followees. By using these attributes, we extract consistency subsets from the follower set of a user. We explain these two attributes in the follow.

**(i) Common Terms in Profiles and Location Information**

As the first attribute measuring consistency, we consider common terms included in profiles and local information of users in the follower set.

There is a high possibility that users belonging to the same community or having the same interest have the same term in their profiles or location information in common. Thus, we extract such terms for measuring consistency. Here, we extract only noun phrases, which characterize their profiles or location information strongly.

Based on the above, we define the method of extracting consistency subsets more formally. We compute the consistency subset $S_{F_u t}$ which is consistent in the term $t$ in $F_u$ by the formula below:

$$S_{F_u t} = \sum_{f \in F_u} \{f |\ t \in \mathit{Demography}(f)\}$$

where $\mathit{Demography}(f)$ is the profile and location information of $f$. This is how we extract consistency subsets by using common terms in profiles and location information.

**(ii) Common Followees**

As the second attribute measuring consistency, we consider common followees of users in the consistency subset.

Users who are consistent in a certain noticeable character often have the common tendency of the follow. Users belonging to the same community are dense on the social graph, so there is a high possibility that they follow common users in the community. And also, followers of a user publishing technical information about programming is considered to follow another user publishing useful information about programming in common. Thus, we focus on the tendency of the follow and extract such followees for measuring consistency.

Based on the above, we define the method of extracting consistency subsets more formally. We compute the consistency subset $S_{F_u e}$ which is consistent in the followee $e$ in $F_u$ by the formula below:

$$S_{F_u e} = \sum_{f \in F_u} \{f |\ e \in E_f\}$$

where $E_f$ is the followee set of $f$. This is how we extract consistency subsets by using common followees.

### 4.1.4 Final Classification Based on Target Specificity

In this subsubchapter, we explain the method of computing the target specificity of a Twitter user by using scores computed up to this point. In addition, based on the target specificity, we explain the method of classifying users.

The $\mathit{SpecificityScore}_{att}(u)$ computed in 4.1.1 is higher in the cases that

- the more consistent followers of a user are in a noticeable character, or
- the more covered his/her follower set are with consistency subsets which cover intermediately widely.

Then, we compute a score of the target specificity of $u$. We first compute $\mathit{SpecificityScore}_{term}(u)$, which is using common terms in profiles and location information as an attribute measuring consistency mentioned 4.1.3 (i), and $\mathit{SpecificityScore}_{followee}(u)$, which is using common followees mentioned 4.1.3 (ii). Then, we take the larger one of these two scores for the target specificity of $u$, because the larger one characterizes the target specificity more strongly than the other. More formally, we compute $\mathit{TargetSpecificity}(u)$ by the formula below:

$$\mathit{TargetSpecificity}(u) = \max\{\mathit{SpecificityScore}_{term}(u), \mathit{SpecificityScore}_{followee}(u)\}$$

Then, we determine a threshold $\delta$ which can classifies target users and non target users accurately the most, and we classify them by $\delta$. More formally, we classify them as follows:

$$\begin{cases} u \text{ is a } \textit{target user}, & \text{if } \mathit{TargetSpecificity}(u) > \delta \\ u \text{ is a } \textit{non target user}, & \text{otherwise.} \end{cases}$$

This is how we the target specificity of a Twitter user.

## 4.2 Classifying Users of High Target Specificity

In this subsection, in regards to Twitter users classified into "the target specificity is high" by the method mentioned in 4.1, we explain the method of determining why their target specificities are high based on the result of our analysis mentioned in 3.2.

Here, we focus on the two causes of the high target specificity mentioned in 3.2 as follows:

**(1)** because they publish information specified for certain topics, and

**(2)** because they publish information to the users specified extensionally,

and we determine whether a user we intend to classify correlates with each cause mentioned above.

We first determine various features of the user which potentially correlate with each cause. Then, based on these features, we construct 3-class classifiers which classify users into three categories: (1) their target specificity is high because they publish information specified for certain topics, (2) because they publish information to the users specified extensionally, and (3) in the cause of both (1) and (2). By classifying users with the above classifiers, we determine why their target specificities are high.

We adopted SVM and decision tree as a classifier. Next, we explain what features of users we used for the classification. All the feature values shown below were normalized to a value between 0 and 1.

**(i) numbers of followees and followers, and their ratio**

If the user publishes information to unspecified users, there is high probability that a number of his/her followers is quite large or a number of his/her followees is quite small. So in such a case, a ratio of a number of his/her followees to a number of his/her followers is supposed to be very small. Furthermore, if the user publishes information to the closed users, i.e., his/her friends, his/her club members, and so on, there is high probability that numbers of his/her followers and followees are very close because the user is supposed to have a

reciprocal connection with them. Thus, numbers of followees and followers, and their ratio are useful for determining why their target specificities are high.

**(ii) mutual follow ratio**

There is high probability that the user publishing information to the closed users has a large mutual follow ratio, i.e., a number of users with whom one follows one another is large, because the user is supposed to have a reciprocal connection with them. Therefore, a mutual follow ratio is expected to be useful for determining why their target specificities are high.

**(iii) frequency of replies by "@"**

There is high probability that the user publishing information to the closed users has a high frequency of replies by "@", i.e., the user replies to his/her followers frequently. In regard to a mutual follow ratio mentioned in (ii), there are some users publishing information to unspecified users in spite of a large mutual follow ratio, but in regard to a frequency of replies by "@", there is high probability that the user publish information to the closed users. This is because a high frequency of replies by "@" demonstrates that the user is supposed to have a reciprocal connection with them. Therefore, a frequency of replies by "@" is expected to be useful for determining why their target specificities are high.

**(iv) partialness of topics in messages**

In regard to a user publishing information specified for certain topics, topics in his/her messages are often partial. Thus, we use the partialness of topics in his/her messages as a feature for the classification.

Then, we explain how to compute a partialness of topics in messages of $u$. We first collect up to 200 messages in order of newness. Second, we determine the topic of each message by using Latent Dirichlet Allocation (LDA), which is a generative probabilistic model for collections of discrete data such as text corpora. Then, we compute partialness of topics in messages $\mathit{partialness}(u)$ as follows:

$$\mathit{partialness}(u) = -\sum_{t \in T_u} p_t \log p_t, \qquad p_t = \frac{|\{m|\ m \in M_u,\ \mathit{topic}(m) = t\}|}{|M_u|}$$

where $M_u$ is a message set of $u$, $T_u$ is the topic set we use for computing this feature, and $topic(m)$ is the topic of a message $m$. The partialness of topics $partialness(u)$ is the entropy of $M_u$ on the topic. This is how we compute a partialness of topics in messages.

# Chapter 5 Experiments and Discussions

In this chapter, we conduct experiments to evaluate our methods proposed in Chapter 4, and present the results obtained from them. In addition, we discuss our methods based on the results.

## 5.1 Data Set

We collected the data set from the real Twitter data by using Twitter API.

We first randomly selected 1,000 Twitter users whose timezone is Japan. At this time, we omitted users who are followed from nobody and who post no tweet in order to select only active users. Then, we divided them in two sets equally, i.e., each of which include 500 users.

Second, we had 6 experienced Twitter users as participants, all of whom are male graduate students in engineering, from 23 to 25 years old. We assigned each set to 3 participants, and we asked each participant to determine one of the following categories each user in the assigned set is supposed to be in:

**(i)** the user publishes information to the public widely,

**(ii)** the user publishes information specified for certain topics,

**(iii)** the user publishes information to the users specified extensionally, and

**(iv)** the user publishes information (ii) specified for certain topics (iii) to the users specified extensionally.

These categories correspond to the category (b), (1), (2), (3) in Figure 2 respectively.

Then, we selected users whose category at least 2 out of 3 participants coincide with, and as a result, we were able to collect 93, 320, 375, and 30 users in the category (i), (ii), (iii), and (iv). We randomly selected 90 users from the category (i), and 30 users from (ii), (iii), and (iv) respectively. We collected these 180 users in total, and we used them as the data set. Table 3 shows the breakdown of the data set: average and standard deviation of numbers of followers, followees, and tweets in each category.

Then, for each user, we collected at most 1,000 followers of the user, and in regard to the followers who follow at most 1,000 users, we also collected their

Table 1: Average and standard deviation of numbers of followers, followees, and tweets in each category

| | | i | ii | iii | iv |
|---|---|---|---|---|---|
| follower | average | 475,679 | 58,142 | 573 | 82,942 |
| | standard deviation | 535,894 | 171,784 | 1,389 | 262,161 |
| followee | average | 11,274 | 3,353 | 598 | 1,568 |
| | standard deviation | 37,906 | 7,218 | 1,545 | 3,594 |
| tweet | average | 9,763 | 9,992 | 8,829 | 5,677 |
| | standard deviation | 14,607 | 23,572 | 29,505 | 6,600 |

Table 2: Average and standard deviation of numbers of followers, followees, and tweets in each category

| Removed Feature | with all | i | ii | iii | iv |
|---|---|---|---|---|---|
| 3-class SVM | 65.6 | 64.4 | 62.2 | 63.3 | **56.7** |
| 2 binary SVMs | 67.8 | 65.6 | 68.9 | **73.3** | **63.3** |
| 3-class decision tree | 54.4 | 54.4 | 46.7 | 48.9 | 50.0 |
| 2 binary decision trees | 50.0 | 52.2 | 46.7 | 45.6 | 51.1 |

Table 3: Average and standard deviation of numbers of followers, followees, and tweets in each category

| Removed Feature | | with all | i | ii | iii | iv |
|---|---|---|---|---|---|---|
| SVM | topic | 82.2 | 81.1 | 84.4 | **86.7** | 81.1 |
| | user | 85.6 | 84.4 | 84.4 | **86.7** | 82.2 |
| decision tree | topic | 68.9 | 68.9 | 68.9 | 70.0 | 74.4 |
| | user | 75.6 | 76.7 | 74.4 | 71.1 | 70.0 |

followees. We used them in order to evaluate our methods.

## 5.2 Experimental Settings and Libraries

First, we conducted the experiments evaluating the method of classifying Twitter users based on the target specificity of their information publishing. We

used 90 users classified into the category (i) as target users, and 90 users classified into the category (ii), (iii), and (iv) as non target users. We first computed $SpecificityScore_{term}(u)$ and $SpecificityScore_{followee}(u)$ for each user $u$, and computed $TargetSpecificity(u)$ based on the above scores. Then, we determined a threshold $\delta$ which can classifies target users and non target users accurately the most, and evaluated the classification results with $\delta$. When computing $TargetSpecificity(u)$, we used two models: the probabilistic model and the subtracting model mentioned in 4.1.2, and compared them.

Second, we conducted the experiments evaluating the method of determing why the target specificity of the user is high in regard to the target user. We extracted users from the category (ii), (iii), and (iv) by 30 users, and used 90 users in total. We first extracted features mentioned in 4.2 from the user. Then, based on these features, we constructed 3-class classifiers which classify users into three categories: (ii), (iii), and (iv), and evaluated the classification results using 10-fold cross validation. We used two learning algorithms: SVM and the decision tree as classifiers, and compared them. For SVM, we used LIBSVM[1)], which is a popular SVM library, with the Gaussian kernel. For the decision tree, we used scikit-learn [2)].

We used twpro search API[3)] in order to get a number of users who have a certain term in their profiles. We also used MeCab[4)] for morphological analysis of Japanese sentences in profiles, local information, and tweets of $u$. Furthermore, we used gensim[5)] for using Latent Dirichlet Allocation (LDA).

実験結果を図??，図??，図??に示す．図??は，プロフィール・位置情報を用いて対象限定性を表した手法，図??は，フォローの傾向を用いて対象限定性を表した手法を表しており，それぞれスコアの高いものからソートしている．図??は，上記の二つの手法で求めたスコアをそれぞれ降順にソートし，ターゲット型ユーザを正解とした場合の，適合率・再現率曲線を表している．また，適合率・再現率曲線のベースラインとして，20 件のユーザそれぞれのフォロワー数

[1)] http://www.csie.ntu.edu.tw/~cjlin/libsvm/

[2)] http://scikit-learn.org/stable/modules/tree.html

[3)] http://twpro.jp/doc/api/search

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

[5)] http://radimrehurek.com/gensim/

を昇順にソートし，同じくターゲット型ユーザを正解としたものを用いている．

フォロワー内でのプロフィール・位置情報を用いた手法に関しては，図??に示すように，ターゲット型ユーザの方が，非ターゲット型ユーザよりも，対象限定性を表すスコアが高くなっており，これらを明確に分類することができているといえる．また，図??に示すように，適合率・再現率曲線も，ベースラインを上回っている．これらの結果より，プロフィール・位置情報を用いて対象限定性を表す手法は，ターゲット型ユーザと非ターゲット型ユーザを分類するのに有効であるといえる．

フォロワー内でのフォローの傾向を用いた手法に関しても，図??に示すように，ターゲット型ユーザの方が非ターゲット型ユーザよりも，対象限定性を表すスコアが平均的に高くなった．しかし，これら二つの間の差は大きくなく，適合率・再現率曲線に関しては，図??に示すように，ベースラインを下回る結果となった．これらの原因としては，以下のようなものが考えられる．

**(1)** 本実験では，フォロワーのフォローの傾向を用いた手法の中で，Twitter の日本人ユーザ全体でのフォローの頻度 $uf$ を求めているが，実際には，Twitter に登録しているものの，ほとんど利用していないユーザも多く存在する．そのようなユーザの影響で，$uf$ が必要以上に小さくなってしまっており，最終的なスコアに十分反映されていないと考えられる．今後は，Twitter ユーザ全体ではなく，Twitter のアクティブユーザのみの中からフォローの頻度を求めることで，本問題の改善を図る予定である．

**(2)** フォロワー内に強く一貫した傾向がなくても，フォローに共通の傾向が見られる場合があると考えられる．本実験において，社会のニュースを発信するユーザのフォロワーの多くが，社会のニュースを発信する別のユーザを同時にフォローしているというケースが見られたが，実際には，フォロワー内に社会のニュースに興味があるという弱い一貫性はあるものの，強く一貫した傾向は見られなかった．

## 5.3 実験結果の詳細と分析

本節では，実験結果から，正解データとして用いたユーザの分析を行った結果を示す．表??に，正解データ内のターゲット型ユーザ・非ターゲット型ユーザのうち，各 2 ユーザの詳細を表している．各ユーザそれぞれに関して，??節で

定義した *tf* が高い 3 語と，*ff* が高い 3 ユーザを示し，実験結果におけるそれらのスコアを示している．

ターゲット型ユーザである@minnanomachi(写真に関するプロジェクトのアカウント) に関しては,「写真」や「カメラ」といった語のスコアや，カメラ関連のアカウントのスコアが高くなっており，そのアカウントを特徴付けるような結果が得られているといえる．同様にターゲット型ユーザである@USJ_BGM(USJ に関するアカウント) に関しても,「USJ」といった語のスコアや，USJ 関連のアカウントのスコアが高くなっている．また,「好き」といった一般に出現頻度の高い語や，@masason(孫正義) といった広く一般に知られているユーザは，*tf* や *ff* が高くなっているが，*df* や *uf* によりそれらのスコアは下がっており，最終的なスコアは非常に小さくなっていることが分かる．

非ターゲット型ユーザに関しては,「大好き」や「フォロー」といった，一般に出現頻度の高い語や，@pamyurin(きゃりーぱみゅぱみゅ) や@TDR_PR(東京ディズニーリゾートに関するアカウント) といった，広く一般に知られているユーザが抽出されているが，*df* や *uf* によりこれらのスコアは下がっている．特に，プロフィール・位置情報を用いた手法のスコアは全て 0 となっている．

このように，これらの例に関しては，フォロワー集合の中で特徴のある語やフォローの傾向を抽出することができており，提案手法が有効であるといえる．ただし，現在の手法では,「写真」と「カメラ」というような同じトピックに関する語を，完全に独立なものとして扱っている．このような語を関連付けて考え，対象限定性を求める有効性を向上させることが今後の課題である．

## 5.4 アプリケーションの実装

本節では，本研究の提案手法を応用した，実装予定のアプリケーションについて述べる．

本アプリケーションは，提案手法に基づき，Twitter の検索結果を 3 つのタブに分類することで，ユーザの Twitter 検索を支援するものである．以下に，表示する 3 つのタブついてそれぞれ示している．

**(1)** 対象範囲が広いユーザの記事を表示する．ある内容に関する一般的なニュースのように，広く一般のユーザが興味を示すような記事を得ることができる．

**(2)** あるトピックに特化された情報を発信するユーザの記事を表示する．あるイベントなどに関する具体的な情報や，あるトピックに関する専門的な情報を得ることができる．

**(3)** クローズドなユーザ集合に情報を発信するユーザの記事を表示する．ある内容に関する一般ユーザの反応などといった，個人のつぶやきなどを得ることができる．

このように分類することで，自分の求めている情報を効率よく取得することができると考えている．また，**??**節で，あるトピックに特化されており，かつクローズドなユーザ集合に情報を発信するユーザが存在する旨を述べたが，そのような記事は，上記の (2) と (3) の両方のタブに表示する予定である．

# Chapter 6 Conclusion

In this study, we focus on the fact that the wideness of target scope of information publishing varies greatly among users bacause Twitter is used for various purpose, we proposed the method to classify Twitter users from the point of view of how widely the target scope of their information publishing is, i.e., whether they publish information to the public widely or publish information specified in certain users.

First, we defined the target specificity of the Twitter user, as the measure of how much target scope of their information publishing is specified. Second, based on this definition, we proposed the algorithm of computing a score of the target specificity. In this algorithm, we focused on the two parameters: (a) whether followers of the user are consistent in a certain noticeable character or not, and (b) whether the follower set of the user is covered with consistency subsets which cover intermediately widely or not. For computing the score, we proposed a couple of models computing a score of the consistency subset: the probabilistic model and the subtracting model, and we compared the above two models. Furthermore, we proposed a couple of attributes for extracting consistency subsets from the follower set of the user: common terms in profiles and location information and common followees. Then, we finally proposed the method of classifying Twitter users based on the score.

実験結果

In addition, in this study, in regard to Twitter users classified into “he target specificity is high ””y thehe above method, we proposed thy the above method, we proposed the method of determining why their target specificities are high. We analysed the causes of high target specificity, and based on them, we classified the user into three categories: (1) they publish information specified for certain topics, (2) they publish information to the users specified extensionally, and (3) in the cause of both (1) and (2). In this method, we constructed 3-class classifiers which classify a user into the above three categories based on various features of the user which potentially correlate with each cause, and we determined why the target specificity is high by using them. 実験結果

今後の課題としては，以下のものが挙げられる．

**(1)** 本研究の予備実験における正解データデータセットは，決して十分なデータ数とは言えず，信頼性も十分に担保されているとはいえない．今後，さらに正解データ数を増やし，かつその信頼性を担保することで，再実験を行うことを検討している．また，今回は，**??**節の対象限定性を求める手法に関する予備実験しか行っていないので，**??**節の対象限定性の高いユーザの分類手法に関する実験も行う予定である．

**(2)** **??**節や 5.3 節で示したような，提案手法の改善について検討を行うことで，提案手法の有効性をさらに向上させることを検討している．

**(3)** 5.4 節で示した，Twitter 検索の際に，Twitter ユーザの対象限定性を考慮して記事の分類を行うアプリケーションの実装を行う予定である．

# Acknowledgments

I would like to express my sincere gratitude to my supervisor, Professor Keishi Tajima for his valuable advises and helpful discussions. I deeply appreciate Proffessor Yasushi Sakurai and Associate Professor Yasuhito Asano for spending precious time for valuable discussions. I would also like to thank the other lab members for their supports for our research.

# References


[1] Brian, M.: Twitter now has 200 million monthly active users, up 60 million in 9 months, http://thenextweb.com/twitter/2012/12/18/twitter-now-has-200-million-monthly-active-users-up-60-million-in-9-months/ (2012).

[2] McGee, M.: With 400 Million Tweets Per Day, Twitter Spending "Inordinate Resources" On Improving Content Discovery, http://marketingland.com/twitter-400-million-tweets-daily-improving-content-discovery-13581 (2012).

[3] Irani, D., Webb, S., Pu, C. and Li, K.: Study of trend-stuffing on twitter through text classification, *Collaboration, Electronic messaging, Anti-Abuse and Spam Conference (CEAS)* (2010).

[4] Broder, A.: A taxonomy of web search, *ACM Sigir forum*, Vol. 36, No. 2, ACM, pp. 3–10 (2002).

[5] Busch, M., Gade, K., Larson, B., Lok, P., Luckenbill, S. and Lin, J.: Earlybird: Real-time search at twitter, *Data Engineering (ICDE), 2012 IEEE 28th International Conference on*, IEEE, pp. 1360–1369 (2012).

[6] Sriram, B., Fuhry, D., Demir, E., Ferhatosmanoglu, H. and Demirbas, M.: Short text classification in twitter to improve information filtering, *Proceeding of the 33rd international ACM SIGIR conference on research and development in information retrieval*, ACM, pp. 841–842 (2010).

[7] Cha, M., Haddadi, H., Benevenuto, F. and Gummadi, K.: Measuring user influence in twitter: The million follower fallacy, *4th international aaai conference on weblogs and social media (icwsm)*, Vol. 14, No. 1, p. 8 (2010).

[8] Cheng, J., Romero, D. M., Meeder, B. and Kleinberg, J.: Predicting reciprocity in social networks, *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, IEEE, pp. 49–56 (2011).

[9] Chu, Z., Gianvecchio, S., Wang, H. and Jajodia, S.: Who is tweeting on twitter: human, bot, or cyborg?, *Proceedings of the 26th Annual Computer Security Applications Conference*, ACM, pp. 21–30 (2010).

[10] Duan, Y., Jiang, L., Qin, T., Zhou, M. and Shum, H.: An empirical study on learning to rank of tweets, *Proceedings of the 23rd International Con-*

*ference on Computational Linguistics*, Association for Computational Linguistics, pp. 295–303 (2010).

[11] Ehrlich, K. and Shami, N.: Microblogging inside and outside the workplace, *Proc. ICWSM*, Vol. 10 (2010).

[12] Sakaki, T., Okazaki, M. and Matsuo, Y.: Earthquake shakes Twitter users: real-time event detection by social sensors, *Proceedings of the 19th international conference on World wide web*, ACM, pp. 851–860 (2010).

[13] Hopcroft, J., Lou, T. and Tang, J.: Who will follow you back?: reciprocal relationship prediction, *Proceedings of the 20th ACM international conference on Information and knowledge management*, ACM, pp. 1137–1146 (2011).

[14] Ikawa, Y., Enoki, M. and Tatsubori, M.: Location inference using microblog messages, *Proceedings of the 21st international conference companion on World Wide Web*, ACM, pp. 687–690 (2012).

[15] Java, A., Song, X., Finin, T. and Tseng, B.: Why we twitter: understanding microblogging usage and communities, *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 workshop on Web mining and social network analysis*, ACM, pp. 56–65 (2007).

[16] Kunegis, J., Lommatzsch, A. and Bauckhage, C.: The slashdot zoo: mining a social network with negative edges, *Proceedings of the 18th international conference on World wide web*, ACM, pp. 741–750 (2009).

[17] Kwak, H., Lee, C., Park, H. and Moon, S.: What is Twitter, a social network or a news media?, *Proceedings of the 19th international conference on World wide web*, ACM, pp. 591–600 (2010).

[18] Lee, K., Caverlee, J. and Webb, S.: The social honeypot project: protecting online communities from spammers, *Proceedings of the 19th international conference on World wide web*, ACM, pp. 1139–1140 (2010).

[19] Leskovec, J., Huttenlocher, D. and Kleinberg, J.: Predicting positive and negative links in online social networks, *Proceedings of the 19th international conference on World wide web*, ACM, pp. 641–650 (2010).

[20] Leskovec, J., Huttenlocher, D. and Kleinberg, J.: Signed networks in social media, *Proceedings of the SIGCHI Conference on Human Factors in*

*Computing Systems*, ACM, pp. 1361–1370 (2010).

[21] Massoudi, K., Tsagkias, M., de Rijke, M. and Weerkamp, W.: Incorporating query expansion and quality indicators in searching microblog posts, *Advances in Information Retrieval*, Springer, pp. 362–367 (2011).

[22] Mathioudakis, M. and Koudas, N.: Twittermonitor: trend detection over the twitter stream, *Proceedings of the 2010 ACM SIGMOD International Conference on Management of data*, ACM, pp. 1155–1158 (2010).

[23] Nagmoti, R., Teredesai, A., De Cock, M. et al.: Ranking approaches for microblog search, IEEE Computer Society (2010).

[24] Nishida, K., Banno, R., Fujimura, K. and Hoshide, T.: Tweet classification by data compression, *Proceedings of the 2011 international workshop on DETecting and Exploiting Cultural diversiTy on the social web*, ACM, pp. 29–34 (2011).

[25] Pennacchiotti, M. and Popescu, A.: A machine learning approach to twitter user classification, *Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)* (2011).

[26] Takemura, H. and Tajima, K.: Tweet Classification Based on Their Lifetime Duration, *Proceedings of the 21st International Conference on Information and Knowledge Management (CIKM)*, ACM, pp. 2367–2370 (2012).

[27] Teevan, J., Ramage, D. and Morris, M.: # TwitterSearch: a comparison of microblog search and web search, *Proceedings of the fourth ACM international conference on Web search and data mining*, ACM, pp. 35–44 (2011).

[28] Weng, J., Lim, E.-P., Jiang, J. and He, Q.: Twitterrank: finding topic-sensitive influential twitterers, *Proceedings of the third ACM international conference on Web search and data mining*, ACM, pp. 261–270 (2010).

[29] Wu, S., Hofman, J. M., Mason, W. A. and Watts, D. J.: Who says what to whom on twitter, *Proceedings of the 20th international conference on World wide web*, ACM, pp. 705–714 (2011).

[30] Yan, L., Ma, Q. and Yoshikawa, M.: Classifying Twitter Users Based on User Profile and Followers Distribution, *Database and Expert Systems Ap-*

*plications*, Springer, pp. 396–403 (2013).

[31] Zhao, D. and Rosson, M. B.: How and why people Twitter: the role that micro-blogging plays in informal communication at work, *Proceedings of the ACM 2009 international conference on Supporting group work*, ACM, pp. 243–252 (2009).