A Study on Personalized Influencer Estimation and Bookmarking Forecast Based on Social Data Analysis

その他（別言語等）のタイトル	ソーシャルデータの解析に基づく個々人のインフルエンサ推定およびブックマーク予測に関する研究
著者	荒澤 孔明
学位名	博士（工学）
学位の種別	課程博士
報告番号	甲第464号
研究科・専攻	工学専攻
学位授与年月日	2021-03-23
URL	http://doi.org/10.15118/00010389

国立大学法人室蘭工業大学
博士学位論文

A Study on Personalized Influencer Estimation and Bookmarking Forecast Based on Social Data Analysis

（ソーシャルデータの解析に基づく個々人のインフルエンサ推定およびブックマーク予測に関する研究）

荒澤孔明

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Recommender System Combined with SNS

Discovering a method for changing people's interests is an important task. For example, it helps for us to promote various items, to improve students' motivation to learn, and to care dependent patients.

In Informatics, a number of studies have discussed various technologies on Information Recommendation to change people's interests. These technologies can induce many users' interests and support their decision makings, by recommending information that would be useful for them. There are various methods [1, 2] for recommending information, and they can recommend various kinds of information. For instance, several systems recommend information on a music [3–5] and a recipe [6–9], they can induce a user's interest for a new "Hobby." In addition, several systems recommend information on a job listing [10–12] and a tourist spot [13–15], and they can induce a user's interests for a new "Job" and a new "Place." Furthermore, Information Recommendation is utilized in several services that support people's encounters [16–19], and they can induce a user's interest for a new "Person." Therefore, it appears that Information Recommendation is one of the important techniques for changing a user's interests for various things.

On the other hand, in Social Science, SNS (Social Networking Service) gets attention as a service for changing people's interests. A user's interests for various things are changing in SNS, because he/she interacts with various persons and is influenced by them. In recent years, several studies [20, 21] have proposed a strategy that advertises a company and its items using word of mouth in SNS, such as Viral Marketing. Therefore, various methods on Marketing with SNS would be studied rapidly [22–24], because there is a tendency that the kinds of SNS and the number of users increase.

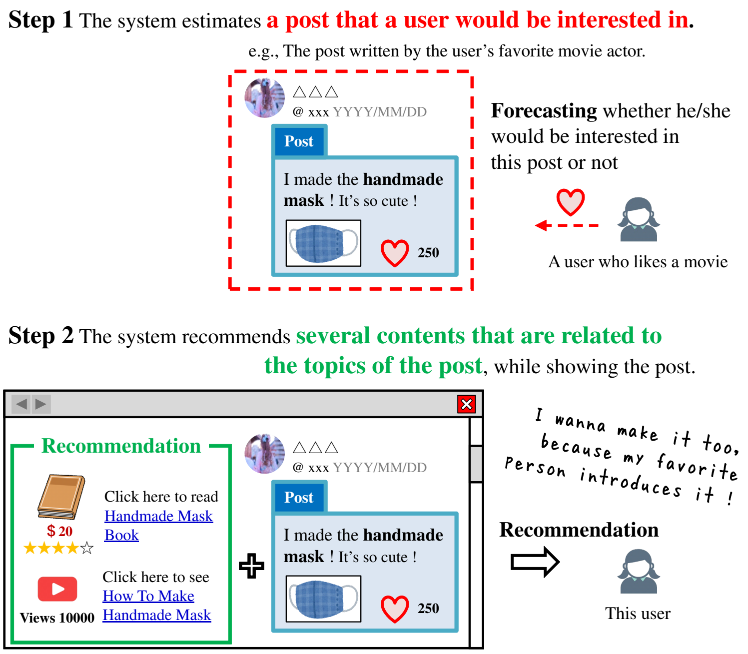


Figure 1.1: A recommender system utilizing a post in SNS

On these backgrounds, I have examined a recommender system combined with SNS. Figure 1.1 is one of the systems that I have examined as a vision for the future. The system would have the following processes. First, the system estimates a post that a user would be interested in, by analyzing this user's profile and the logs on this user's behaviors (Figure 1.1–Step1). Second, the system recommends several contents that are related to the topics of the post, while showing the post (Figure 1.1–Step2). When a user is interested in the topics of a post, the system advertises several contents that are related to the topics, and encourage this user to understand them and to behavior (click/purchase). Therefore, it appears that the system can change people's interests efficiently.

Incidentally, there are various reasons why a user is interested in a post (Table 1.1). In other words, even if a user was not interested in the topics of a post, this user's interest may change by the triggers like Table 1.1. That is to say, it appears that a post that the recommender system of Figure 1.1 shows to a user, has not only the topics that this user is already interested in but also various topics. It is expected that the recommender system in my vision improves diversity and serendipity on Information Recommendation because it is able to recommend various items.

Table 1.1: Various reasons why a user is interested in a post

	Reason
1	The post includes an expression recommending something.
2	The person who wrote the post has an influence on this user.
3	Many users react to the post positively.
4	The topics of the post match with this user's preferences.
5	The post (or the person who wrote it) has reliability.
6	Timing when this user looks at the post matches this user's contexts.

1.2 Purpose and Organization of This Paper

This paper discusses two phases that are the bases of the recommender system in my vision (Figure 1.2).

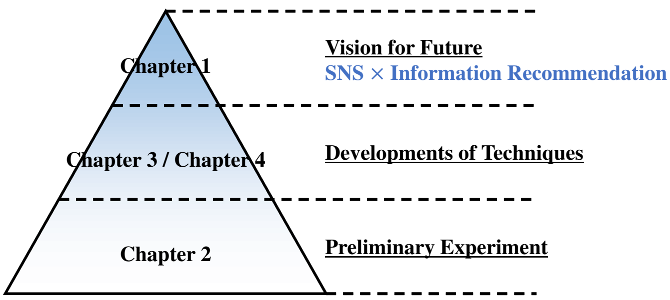


Figure 1.2: The grade of each chapter in this paper

1.2.1 Preliminary Experiment (Chapter 2)

The first phase conducts a preliminary experiment and discusses the effects of the assumed recommender system by using its prototype system. This prototype system recommends several movies. In this experiment, when the prototype system recommends a movie, it shows also the post where an influencer introduces the movie positively and the distribution of many users' opinions on the movie. In addition, I analyze the change of a user's interest for the movie between before the user looks at the advertisement and after he/she looks at it.

1.2.2 Developments of Techniques (Chapter 3/4)

The second phase develops several component technologies of the assumed recommender system. It is one of the important technologies of the assumed recommender system, to calculate how much a user would be interested in a post (i.e., to estimate a post that a user would be interested in). In other words, it is necessary for the societal implementation of the assumed recommender system to develop a model that can classify whether a user would be interested in a post or not (Figure 1.3–Bottom). However, the discussions of this technique cannot proceed.

The direct cause is that it is not good enough to analyze various behaviors after a user looks at a post. For instance, each behavior in Table 1.2 is one of the behaviors of a user who is interested in a post. Therefore, forecasting whether a user takes a behave in Table 1.2 or not, helps for the assumed recommender system to judge whether this user would be interested in the post or not.

Table 1.2: Various behaviors of a user who is interested in a post

	Behavior
A	This user bookmarks the post.
B	This user collects information on its topics (e.g., Web search).
C	This user posts a message including the same topics with the post.

In addition, the root cause (i.e., a reason why it is difficult to forecast the behaviors in Table 1.2) is that a method that calculates each of several independent variables for forecasting them, has not been established. For instance, each affector in Table 1.3 would be one of the independent variables for forecasting the behaviors in Table 1.2.

Table 1.3: The independent variables to forecast the behaviors in Table 1.2

	Affector
I	Whether the writer has an influence on the user or not
II	Whether a topic is introduced positively in the post or not
III	Whether the user is easily swayed by around or not
IV	The user's interests for the topics of the post
V	The contexts of the user
VI	Reliability of the post (or its contents)

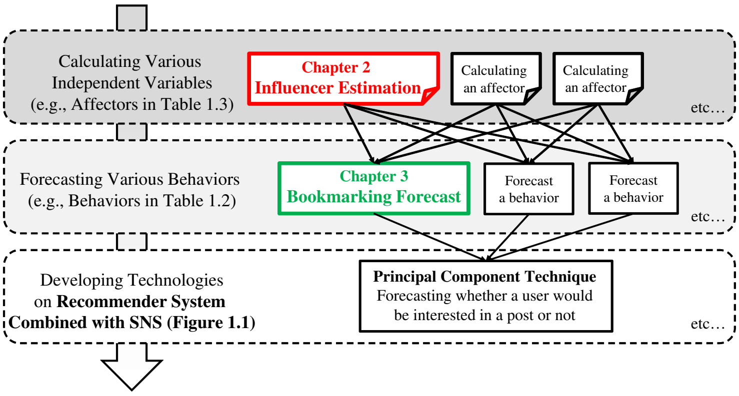


Figure 1.3: The relationships between each technology

Therefore, the assumed recommender system needs the techniques for forecasting various behaviors and calculating various independent variables for these forecasts, to estimate a post that a user would be interested in. This paper discusses two techniques among them, and their overviews are as follows.

Personalized Influencer Estimation (Overview of Chapter 3)

First, this paper discusses a method for judging whether the person who wrote a post has an influence on a user or not, which is included among the top of Figure 1.3. Specifically, I propose the methods that can estimate a person who have an influence on a user (i.e., the influencers for a user) by analyzing the user's reactions and interests that show he/she is influenced by other persons in SNS. In addition, I evaluate their estimation performances. The goal in this chapter is that to develop a method that can estimate not only a common influencer for all users (i.e., social influencer) but also the different influencers for each user (personalized influencers).

Bookmarking Forecast (Overview of Chapter 4)

Second, this paper develops a model that can forecast a user's behavior in SNS by calculating various independent variables for this forecast. This chapter especially focuses on forecasting a user's bookmark for a post. I consider three kinds of concepts on the affectors when a user bookmarks a post, 1) the goodness of the content and the format of the post, 2) the influence of the person who wrote the post, and 3) the social assurance of the post. In addition, I create a model that forecasts a user's bookmark for a post by using Random Forest that has 22 kinds of independent variables. Specifically, this chapter aims to propose a method that classifies whether or not a user would bookmark a post that he/she may look at.

1.3 Differences from Previous Studies

1.3.1 Novelty in Personalized Influencer Estimation

A number of previous studies define Influencer as a rare person who has a large influence on many users. For example, Merton et al. [25] define an influencer as a person who directly has an influence on at least four users. In addition, Watts et al. [26] calculate each person's influence, and define an influencer as a person who is included in the top 10%. There are various methods for discovering an influencer based on these definitions.

For example, there are the methods [27–29] for estimating an influencer of people by analyzing 1) the number of retweets that a person has received from others, 2) the number of replies that a person has received from others, and 3) the number of followers that a person has (i.e., the number of other users who want to look at the posts written by the person) in Twitter. In addition, Weng et al. [30] and Ding et al. [31] have proposed the methods for estimating an influencer of people based on PageRank. Furthermore, Katagiri et al. [32] have proposed the method for estimating an influencer of people based on the order of users to download a mobile application (e.g., an analysis of the trace that a user downloaded the mobile application that has been downloaded by a friend of the user).

Besides, there are various studies that discuss Information Diffusion and analyze a mathematical model called as Independent Cascade Model (ICM). For example, Several studies [33–35] have proposed the methods for discovering a person who became the starting point of diffusion of information (i.e., a seed that can widely spread information in a social network) by tracing the cascade (i.e., the flow of information). In addition, another study [36] has proposed the method for calculating the time-dependent influence of a node.

These previous studies have a common point. It is that they aim to estimate a person who has an influence on many users (i.e., social influencer). That is to say, an influencer who is estimated by these studies, is the common person for all users.

On the other hand, this paper has a different concept on Influencer, and a person who should be estimated in this paper is different from the persons who has been estimated in previous studies. In this paper, a person whom a system would like to estimate, does not necessarily have a social influence (i.e., an influence on many users). That means when a person has an influence on a specific user even if the person does not have an influence on many other users (i.e., society), the person can become the influencer for the user.

Therefore, a task in this paper is to discover not only a common influencer for all users but also the different influencers for each user. That is, a feature in this paper is to regard not only a social person (e.g., a celebrity) but also a user-specific familiar person (e.g., a friend) as an influencer for the user.

A number of studies have reported the importance of the influence of a user's familiar person. Reingen [37] has found that a recommendation by a person who has a strong connection with a user (e.g., a friend of the user in real world) is easier to move the users' feelings/behaviors than a recommendation by a person who has a weak connection with the user (e.g., a celebrity whom the user likes). In addition, Bither [38] has indicated that a remark of a person who has a strong connection with a user, has more reliable than a remark of a person who does not have a strong connection with the user. That is to say, several posts written by not only a social influencer but also a familiar influencer, would work effectively for the assumed recommender system (Figure 1.1). Therefore, it is important for the assumed recommender system to discover not only a common influencer for all users but also the different influencers for each user.

1.3.2 Novelty in Bookmarking Forecast

There are no previous studies that discuss the same task with this paper (i.e., forecasting a user's bookmark for a post). However, it appears that the task in this paper is similar to the research field on the recommendation of a Web content (e.g., a Web page) that a user would like, and there are various studies about it.

First, there are various recommendation methods that utilize the behaviors of other users. For example, Terveen et al. [40] and Shardanand et al. [41] have proposed the methods that recommend a Web page that many users have viewed (i.e., a Web page whose number of views is large). In addition, there are the methods that recommend a Web page that has been viewed by others who are similar to a user [42, 43]. Furthermore, Hotho et al. [44, 45] have proposed the recommendation algorithm based on PageRank.

Second, there are also various recommendation methods that analyze the contents of a Web page. For example, Niwa et al. [46] have proposed the method that calculates the similarity between the contents of a Web page and a user's preferences. In addition, Zhang et al. [47] have proposed the recommendation method based on Topic Model. Besides, there are the recommendation methods based on the co-occurrence between each Web page [48], Graph Theory [49], and Tensor Decomposition [50]. Therefore, a number of studies have reported that it helps for the recommendation of a Web page to analyze other persons' behaviors and the contents of a Web page.

On the other hand, it is not enough to forecast a user's bookmark for a post, by only analyzing other persons' behaviors and the contents of the post. A system also needs to analyze several relationships between the user and the person who wrote the post. For instance, it has to analyze the influence of the person who wrote the post, because it is an effect for a user to bookmark the post (e.g., it appears that a user is easy to bookmark a post written by a person who has an influence on the user). That is to say, forecasting a user's bookmark for a post, needs also the techniques for estimating the influencers of each user (Chapter 3).

This paper aims to propose a method for discovering a post that a user would be interested in, by analyzing not only other persons' behaviors and the contents of the post but also several relationships between the user and the person who wrote the post. This paper especially focuses on "post in SNS" among various Web contents, and few studies have discussed this task.

1.4 Other Related Studies

This section explains several proposed methods and their ways of thinking that are bases of the calculations of the affectors (II to VI in Table 1.3).

First, I introduce several techniques for calculating whether the topic of a post is introduced positively or not, from the viewpoint of Reputation Analysis for social data. For example, a number of studies [51–53] have proposed the methods that can classify a post as Positive/Negative/Neutral by supervised learning that analyzes text data of various posts. In addition, there are the methods that can calculate the reputations of a post by analyzing related information of the post [54] such as Reply and Retweet (a function spreading a post in Twitter) or focusing on several emphatic expressions [55].

Second, I introduce several techniques for calculating whether a user is easily swayed by around or not. That is methods estimating a user's reaction to a post that is popular around. These techniques would be discussed from the viewpoint of Information Diffusion Forecast, Trend Forecast, and Personality Estimation. For example, there are the methods [56–59] that analyze and forecast several future trends by using an information diffusion model that formulates the spread of information in a social network. In addition, Kleinberg et al. [60] and Bollen et al. [61] have proposed the methods that calculate the attention degree of a post (or a word in the post) by using the frequency of appearance of a word in the posts written by people. Furthermore, Asur [62] has proposed the method that calculates the population of a post by Sentiment Analysis of the users in a social network. Besides, Quercia et al. [63, 66] and Golbeck et al. [64, 65] have studied on Personality Estimation, which would be utilized when a system estimates how easy a user is swayed by around.

Third, I introduce several techniques for calculating how much a user is interested in the topics of a post, from the viewpoint of Attributes Extraction and Preference Extraction. A number of studies [73, 74] have proposed the methods that extract a user's attributes form SNS. For example, there are the methods [68–70] that can estimate a user's residence and his/her principal places of the activities by analyzing the posts written by the user. In addition, Burger et al. [71] and Rao et al. [72] have proposed the methods that can estimate the gender of a SNS user by a machine learning that analyzes the comments in the profile of his/her SNS account. On the other hand, there are various methods that extract a user's preferences from SNS. For example, there are the methods that can estimate a SNS user's preferences by using the frequency of appearance of a word in the posts written by the user [75, 76], its hypernyms [77–79], DBpedia [80], and LDA [81, 82].

Fourth, I introduce several techniques for recognizing the contexts of a user. A number of studies have discussed these methods since focusing on ubiquitous computing. In particular, there are many studies on a recommender system using location information [84–86]. In addition, Sudo et al. [87] have aimed to estimate not only the time but also the places and the situations, when a user uses a system, as an advanced technique for recognizing the user's contexts.

Fifth, I introduce several techniques for calculating reliability of a post. For example, there are the methods for judging whether information has reliability or not, based on factfulness of information [88–90], expertise of the sender [91–93], and machine learning [94, 95].

In this section, I have explained the techniques for calculating the affectors (II to VI in Table 1.3). I am not going to discuss them deeply, because this paper focuses on Influencer Estimation and Bookmarking Forecast. However, the assumed recommender system (Figure 1.1) needs also these techniques to estimate a post that a user would be interested in. I am going to refer to these studies in next paper.

Chapter 2

Preliminary Experiment

2.1 Introduction of This Chapter

The previous chapter has explained that Information Recommendation and SNS are effective in changing people's interests and discussed the system combining them (Figure 1.1) as a vision of the future. The assumed recommender system advertises an item to a user, while showing information of SNS that enables for the user's interest to increase. This chapter discusses how effective the assumed recommender system is in changing people's interests. However, it is difficult to implement the system ideally because several component techniques still have not been established. Therefore, this chapter reports an expected effectiveness of the assumed recommender system by using a prototype system that is similar to it.

2.2 Prototype System

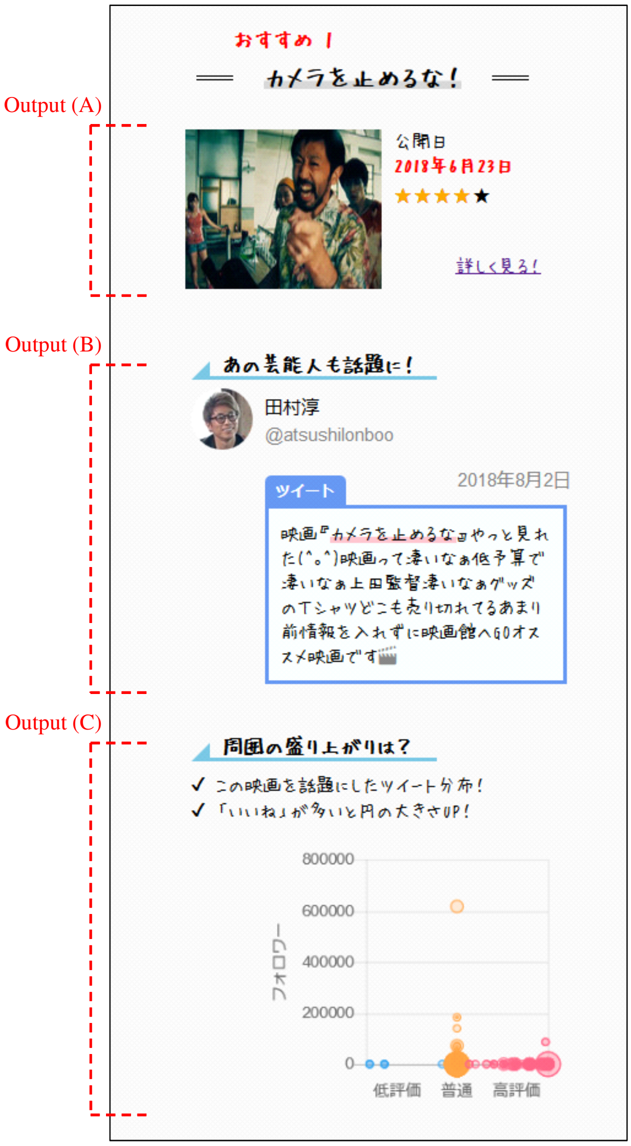
The prototype system has a database about movies to recommend a movie to a user. This database has the titles of several movies, their release dates, and their ratings. These data are extracted from 映画.com[1], which is the web site organizing various kinds of information about different movies.

It is ideal for the assumed recommender system to show a post that fulfills a condition written in Table 1.1 (i.e., a post that a user would be interested in), when the system recommends something to the user. This preliminary experiment focuses on the first, second, and third conditions among those in Table 1.1.

[1] https://eiga.com/

First, I focus on the first and second conditions. Therefore, the prototype system recommends a movie to a user, while showing a post where an influencer for the user introduces the movie positively. An instance of this post is shown in the output (B) of Figure 2.1. The part after this, explains a method for discovering this post. In addition, I analyze the difference of the advertising effectiveness between when simply recommending a movie and when recommending the movie and showing a post where an influencer for the user introduces the movie positively.

Second, I focus on the third condition. Therefore, the prototype system had planned to recommend a movie to a user, while showing a post that introduces the movie and receives many positive reactions from surroundings. To show this post, is the same as to show many users' posts that talk about the movie positively. Actually, it recommends a movie to a user, while showing the distribution of many users' opinions about the movie. An instance of this distribution is shown in the output (C) of Figure 2.1. This preliminary experiment also analyzes the difference of the advertising effectiveness between when simply recommending a movie and when recommending the movie and showing the reviews of surroundings for the movie.



（画像は各々の権利者に帰属する）

Figure 2.1: An instance of the prototype system

2.2.1 Method for Extracting Remarks of Influencers

Experimental Data

This preliminary experiment uses Twitter, and regards a tweet of a celebrity as a post (remark) of an influencer. First, the prototype system extracts Twitter accounts of several celebrities, and divides them into eleven categories (e.g., Actor and Athlete) based on TwiNavi[2]. In addition, it extracts the tweets written by the celebrities, by using TwiLog[3]. However, the maximum number of extracted tweets per account is 100. In addition, replies to someone are not extracted. Table 2.1 shows the number of extracted Twitter accounts and the number of their tweets, every category.

Table 2.1: The number of Twitter accounts and their tweets

Category	Account	Tweets
Comedian	209	10925
TV Personality	210	12935
Idol	77	5062
Actor/Actress	68	3665
Sports Player	81	4403
Politician	58	4183
Musician	384	21841
Artist	40	2455
Painter	85	3764
Investor	91	4903
Others	49	3047

[2] https://twinavi.jp/
[3] https://twilog.org/

Calculation of Suitability as Movie Introduction Tweet

This part explains a method for extracting a tweet where an influencer for a user introduces a movie positively (i.e., Movie Introduction Tweet), from the tweets written by the celebrities. I define four conditions for a tweet to become Movie Introduction Tweet.

- A movie title is included in the tweet.
- The movie is introduced positively in the tweet.
- It is the latest movie (or the date when the tweet is posted is recent).
- The person who wrote the tweet has an influence on surroundings.

In addition, I define a function $\mathrm{score}(t_i)$, which shows how much a tweet t_i is suitable as a Movie Introduction Tweet. The function $\mathrm{score}(t_i)$ has a value from 0.0 to 1.0. The more suitability as Movie Introduction Tweet the tweet t_i is, the closer to 1.0 the value of $\mathrm{score}(t_i)$ gets. On the other hand, if a tweet t_i is not suitable as Movie Introduction Tweet, the value of $\mathrm{score}(t_i)$ is 0.0. In addition, $\mathrm{score}(t_i)$ is calculated by the following formula.

$$\mathrm{score}(t_i) = \mathrm{IM}(t_i) \times \mathrm{RM}(t_i) \times \mathrm{NT}(t_i) \times \mathrm{IP}(t_i)$$

The remainder of this part discusses about the details of these conditions and the formula of $\mathrm{score}(t_i)$.

(i) Whether a movie title is included in a tweet or not

The first condition for a tweet to become a Movie Introduction Tweet, is that a movie title is included in the tweet. Therefore, I define a function $\mathrm{IM}(t_i)$, which shows whether there is a movie title in the text of a tweet t_i or not. A value of the function $\mathrm{IM}(t_i)$ is 0.0 or 1.0. If the text of a tweet t_i includes a movie title, the value of $\mathrm{IM}(t_i)$ is 1.0. On the other hand, if the text of a tweet t_i does not include a movie title, the value of $\mathrm{IM}(t_i)$ is 0.0.

$$\mathrm{IM}(t_i) = \begin{cases} 1.0 & \text{(includes a movie title)} \\ 0.0 & \text{(otherwise)} \end{cases}$$

(ii) The reputation for a movie

The second condition for a tweet to become a Movie Introduction Tweet, is that the introduction/explanation of the movie is positive. Therefore, I define a function RM(t_i), which shows the reputation for a movie written in a tweet t_i (i.e., whether the person who posted the tweet t_i has talked about the movie positively or not). Here, only if there is a movie title in the text of a tweet t_i, RM(t_i) is calculated. A value of the function RM(t_i) is from -1.0 to 1.0. The more positive the comment for a movie in a tweet t_i is, the closer to 1.0 the value of RM(t_i) gets. On the other hand, the more negative the comment for a movie in a tweet t_i is, the closer to -1.0 the value of RM(t_i) gets. In addition, when the value of RM(t_i) gets close to 0.0, it means the reputation for the movie written in the tweet t_i is neutral.

To calculate the reputation for a movie written in a tweet t_i, this preliminary experiment employs a method based on the distance between a movie title and a positive (or negative) word in the text of a tweet t_i. First, the prototype system initializes the value of a variable p_i to 0.0. This variable means the positiveness of a tweet t_i. In addition, it initializes also the value of a variable n_i to 0.0. This variable means the negativeness of the tweet t_i. Second, the prototype system searches for the positive word(s) and the negative word(s) behind the place where there is a movie title in the text of the tweet t_i. The positive/negative words are shown in Table 2.2.

Table 2.2: The positive/negative words showing the reputation for a movie

Positive words	Negative words
最高 / awesome	最低 / terrible
良い / good	悪い / bad
好き / like	嫌い / hate
面白い / interesting	つまらない / boring
楽しい / fun	

Moreover, let each sentence in a tweet t_i, s_j^i $(j = 1, 2, ...)$. If there is a movie title in a sentence s_j^i and there is a positive word in a sentence s_{j+k}^i, the prototype system updates p_i to p_i^{new} by the following formula (here, w is set to 0.8). This formula means that if the distance between the sentence s_j^i and the sentence s_{j+k}^i is far, the prototype system makes only a small update to the positiveness of the tweet t_i.

$$p_i^{new} = p_i + (1.0 - p_i) \times w^k$$

In addition, if there is a movie title in a sentence s_j^i and there is a negative word in a sentence s_{j+k}^i, the prototype system updates n_i to n_i^{new} by the following formula (here, w is set to 0.8).

$$n_i^{new} = n_i + (1.0 - n_i) \times w^k$$

Finally, the function $\mathrm{RM}(t_i)$ is calculated by the following formula. Here, p_i^{fin} and n_i^{fin} in this formula mean the final values of the updated p_i^{new} and the updated n_i^{new}, respectively.

$$\mathrm{RM}(t_i) = \frac{(p_i^{fin} - n_i^{fin}) - (-1.0)}{1.0 - (-1.0)}$$

(iii) The newness of a tweet (or a movie)

The third condition for a tweet to become a Movie Introduction Tweet, is that the tweet or the introduced movie is recent, that is to say, the prototype system had better show a post that introduces a timely movie. Therefore, I define a function $\mathrm{NT}(t_i)$, which shows the newness of a tweet t_i. A value of the function $\mathrm{NT}(t_i)$ is from a default value C to 1.0. The more recent a tweet t_i is, the closer to 1.0 the value of $\mathrm{NT}(t_i)$ gets. On the other hand, the older a tweet t_i is, the closer to the default value C the value of $\mathrm{NT}(t_i)$ gets.

Specifically, the function $\mathrm{NT}(t_i)$ is calculated by the following formula. Here, $r_i \in \{1, 2, ..., N\}$ in this formula shows the rank of a tweet t_i in the case that the tweets t_i $(i = 1, 2, ..., N)$ of the celebrities (Table 2.1) are sorted in order of date when a tweet is posted. For example, $r_i = 1$ means that the date when the tweet t_i is posted, is the latest in the all extracted tweets.

$$\mathrm{NT}(t_i) = C + dif \times (N - r_i)$$

$$dif = \frac{1.0 - C}{N - 1}$$

(iv) The influence of the person who wrote a tweet

The fourth condition for a tweet to become a Movie Introduction Tweet, is that the person who wrote the tweet has an influence on surroundings. Therefore, I define a function IP(t_i), which shows how much the person who wrote a tweet t_i has an influence on surroundings. A value of the function IP(t_i) is from 0.0 to 1.0. The larger the influence of the person who wrote a tweet is, the closer to 1.0 the value of IP(t_i) gets. On the other hand, the smaller the person's influence is, the closer to 0.0 the value of IP(t_i) gets.

This preliminary experiment has a hypothesis that the influence of the person who wrote a tweet depends on the number of followers of the person. Therefore, the function IP(t_i) is calculated by the following formula. Here, f_i (≥ 0) in this formula means the number of followers of the person who wrote a tweet t_i. In addition, f_{max} ($\neq 0$) in this formula shows the maximum value among the numbers of followers of the extracted Twitter accounts.

$$\mathrm{IP}(t_i) = \frac{f_i}{f_{max}}$$

2.2.2 Method for Creating Distribution of Tweets

Next, this subsection discusses about a method for creating a distribution of many users' opinions about a movie, such as the output (C) of Figure 2.1.

First, the prototype system extracts 1500 tweets whose texts include a specific movie title, by using Twitter API. In addition, the prototype system extracts the number of bookmarks that each extracted tweet has received and the number of followers of the person who wrote the tweet. Second, it calculates how positively each extracted tweet t_i evaluates the movie, by using the function RM(t_i). Finally, it creates the distribution of the 1500 tweets. Its vertical axis shows the number of followers of the person who wrote a tweet t_i, and its horizontal axis shows the value of RM(t_i). Here, when plotting a tweet t_i to the distribution, the point size is decided based on the number of bookmarks that the tweet t_i has received. For example, if the number of bookmarks that a tweet t_i has received is large, the tweet t_i is plotted to the distribution as the big point.

2.3 Experiment on Advertising Effectiveness

2.3.1 Overview of Experiment

This preliminary experiment was conducted with 36 subject persons in August, 2018. They consist of 32 males and four females, and most of them are in their 20s. In addition, 23 subject persons often watch various movies on a daily basis, while the others do not.

The prototype system prepares the followings. First, it calculates the function score(t_i), for the tweets t_i ($i = 1, 2, ...$) written by the celebrities (Table 2.1). Second, it divides the tweets t_i ($i = 1, 2, ...$) into eleven categories shown in Table 2.1. If the person who wrote a tweet t_i belongs to Comedian, the tweet t_i also belong to it. This preliminary experiment uses the tweets whose score(t_i) are included in the top 5 of each category as Movie Introduction Tweets. Therefore, there are 55 Movie Introduction Tweets (5 tweets $\times$ 11 categories). Third, it creates the distribution of many users' opinions about the movie that is introduced in each Movie Introduction Tweet. Hence, the prototype system prepares 55 sets of a movie, a tweet, and a distribution (i.e., a set has the output (A), (B), and (C) in Figure 2.1).

In addition, the prototype system executes the following processes for each subject person. First, the system lets each subject person select one among the eleven categories of celebrities. Second, it narrows down the 55 sets to the five sets whose tweets belong to the selected category. Finally, it randomly selects two sets among the five sets, and advertises them to the subject person. The subject user can look at an advertisement (i.e., a set like Figure 2.1), which has the basic information of a movie, the tweet where a celebrity whom the subject person likes introduces the movie, and the distribution of people's opinions about the movie.

In this preliminary experiment, each subject person looks at the set by three types of ways of showing. Advertisement 1[4] shows only the basic information of the movie, like the output (A) of Figure 2.1. Advertisement 2[5] shows both the basic information of the movie and the tweet where a celebrity whom the subject person likes introduces the movie, like the output (A) and the output (B) of Figure 2.1. Advertisement 3[6] shows both the basic information of the movie and the distribution of people's opinions about the movie, like the output (A) and the output (C) of Figure 2.1. Next, each subject person answers each question in Table 2.3 on a scale of 1 to 5, for each advertisement.

Table 2.3: The questions for evaluating the advertisements

	Question
1	Does the advertisement induce the interest for the movie?
2	Does the advertisement give a useful information?
3	Is the advertisement easy to understand?
4	Does the advertisement have a convincing?
5	Do you like the advertisement ?

2.3.2 Results and Discussions

Figure 2.2 shows the mean rating for each question, its details are shown in Table 2.4, Table 2.5, and Table 2.6. Here, High Rating Proportion in each Table shows the proportion of the subject persons who selected 4 or 5 as the answer for a question. In addition, "Mov" shows the answers of the 23 subject persons whose hobbies are to watch movies, and "Oth" shows the answers of the 13 subject persons whose hobbies are not to watch movies.

[4] http://www3.muroran-it.ac.jp/wits/~arasawa/iphs18/1/
[5] http://www3.muroran-it.ac.jp/wits/~arasawa/iphs18/2/
[6] http://www3.muroran-it.ac.jp/wits/~arasawa/iphs18/3/

First, focusing on each mean rating of the question 1 and the question 2 in Figure 2.2, it is revealed that both the advertisement 2 and the advertisement 3 are superior to the advertisement 1. Therefore, it can induce a user's interest for a movie to show not only the basic information about it, but also information where a person who has an influence on the user introduces the movie positively in SNS and information where many SNS users discuss about the movie. It appears this result matches our intuition.

Second, focusing on the mean rating of the question 4 in Figure 2.2, we can confirm that both the advertisement 2 and the advertisement 3 are superior to the advertisement 1. In addition, each of them has a significant difference from the advertisement 1. Furthermore, the mean rating for the question 5 on advertisement 2 is higher than the one on the advertisement 1. A recommender system having a convincing and a recommender system that is easily liked by users, help to a user to accept the thing whom the user was difficult to accept. That is, the recommender system combined with SNS would have an effect that induces a user's interest for a thing that the user had not been interested in.

It would be expected that the assumed system (Figure 1.1) has the effects that are close to these results. Therefore, it appears that there is an effectiveness in changing people's interests in the assumed system of my vision for the future.

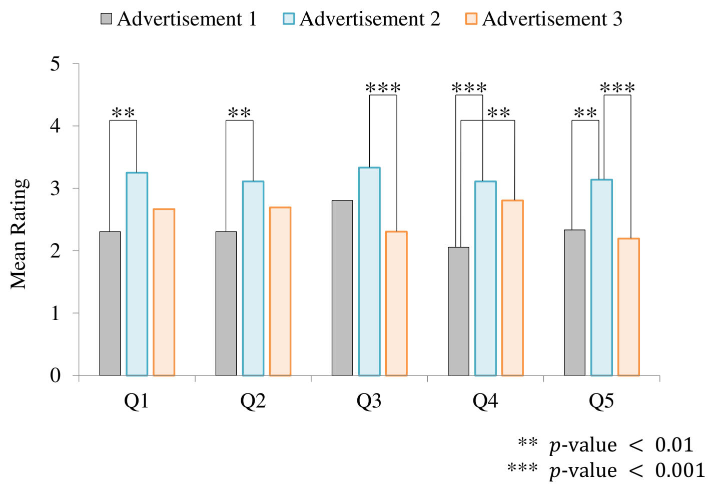


Figure 2.2: The mean rating for each question

Table 2.4: The advertising effectiveness of the output (A)

Question	Q1		Q2		Q3		Q4		Q5	
Hobby of Subject Person	Mov	Oth	Mov	Oth	Mov	Oth	Mov	Oth	Mov	Oth
Sample Mean	2.35	2.23	2.35	2.23	2.87	2.69	2.22	1.77	2.26	2.46
Standard Deviation	1.31	1.19	1.37	1.12	1.42	1.14	1.14	0.70	1.11	1.08
Mode	1	1	1	2	4	2, 3	1	2	1	2
High Rating Proportion	0.26	0.23	0.26	0.23	0.43	0.23	0.17	0.00	0.17	0.23
Significance Probability	0.80		0.80		0.71		0.22		0.61	

Table 2.5: The advertising effectiveness of the output (A)+(B)

Question	Q1		Q2		Q3		Q4		Q5	
Hobby of Subject Person	Mov	Oth	Mov	Oth	Mov	Oth	Mov	Oth	Mov	Oth
Sample Mean	3.35	3.08	3.26	2.85	3.30	3.38	3.17	3.00	3.22	3.00
Standard Deviation	1.17	1.44	1.15	1.35	1.20	1.21	1.13	1.24	1.06	1.18
Mode	4	4	4	4	4	4	4	4	3, 4	3, 4
High Rating Proportion	0.57	0.54	0.52	0.46	0.57	0.54	0.52	0.54	0.43	0.38
Significance Probability	0.55		0.35		0.85		0.68		0.59	

Table 2.6: The advertising effectiveness of the output (A)+(C)

Question	Q1		Q2		Q3		Q4		Q5	
Hobby of Subject Person	Mov	Oth	Mov	Oth	Mov	Oth	Mov	Oth	Mov	Oth
Sample Mean	2.65	2.69	2.70	2.69	2.39	2.15	2.83	2.77	2.26	2.08
Standard Deviation	1.31	1.14	1.33	1.07	1.34	1.03	1.43	1.05	1.15	1.14
Mode	1	3, 4	1, 4	2, 4	1	2	1	3, 4	1	1
High Rating Proportion	0.30	0.31	0.35	0.31	0.30	0.15	0.39	0.31	0.13	0.08
Significance Probability	0.93		0.99		0.59		0.90		0.66	

Chapter 3

Influencer Estimation

3.1 Introduction of This Chapter

This chapter discusses several conditions for a person to become one of the influencers for a user. In addition, I aim to propose several methods for estimating the influencers for the user based on the conditions, and evaluate their estimation performances.

3.1.1 Hypothesis

Cialdini [96] who is a psychologist has claimed six kinds of factors for a user to be affected by other persons (Reciprocation, Consistency, Social Proof, Authority, Liking, and Scarcity). In particular, Liking has been focused by specialists on Information Recommendation. For example, Bonhard et al. [97] and Woerndl et al. [98] have reported that a recommendation from the person whom a user likes has an influence on the user than a recommendation from one who is not.

Therefore, it is assumed that the persons whom a user likes are the part of influencers for the user. Hence, this chapter sets up a hypothesis that either of the following conditions should be fulfilled for a person to become one of the influencers for a user.

1 The user shows favorable reactions to the person's behaviors
This paper embodies this condition as "the user frequently has replied to/bookmarked the posts written by the person."

2 The user shows favorable interests for the person's behaviors
This paper embodies this condition as "a word showing the person, is frequently appeared in the documents including the user's interests."

3.1.2 Overview of Influencer Estimation

This chapter proposes a system that estimates the influencers for each user by analyzing the user's social data, based on the hypothesis.

First, the system calculates Reaction Score $S_{rxn}(t \to i)$ from a target user u_t to each person u_i ($i = 1, 2, ...$), which shows how many the target user u_t has reacted to the posts written by the person u_i. In addition, the system sorts the persons u_i ($i = 1, 2, ...$) based on Reaction Score $S_{rxn}(t \to i)$. Finally, it recognizes several persons whose Reaction Score over a threshold as the influencers for the target user u_t. The detail of these processes is explained in Section 3.2.

Second, the system calculates Interest Score $S_{int}(t \to i)$ from a target user u_t to each person u_i ($i = 1, 2, ...$), which shows how much the target user u_t is interested in the person u_i. In addition, the system sorts the persons u_i ($i = 1, 2, ...$) based on Interest Score $S_{int}(t \to i)$. Finally, it recognizes several persons whose Interest Score over a threshold as the influencers for the target user u_t. The detail of these processes is explained in Section 3.3.

3.2 Influencer Estimation Based on Reaction

3.2.1 Overview of Analysis of Reaction

This subsection proposes a method for estimating the influencers for a target user u_t based on Reaction Score $S_{rxn}(t \rightarrow i)$ from the target user u_t to each person u_i $(i = 1, 2, ...)$. Here, Reaction Score $S_{rxn}(t \rightarrow i)$ shows how many the target user u_t has reacted to the posts written by the person u_i. This method has three steps (Figure 3.1).

First, the system calculates Replying Score $S_{rep}(t \rightarrow i)$ from a target user u_t to each person u_i $(i = 1, 2, ...)$ by analyzing the logs on the replies that the target user u_t has done. In addition, the system also calculates Bookmarking Score $S_{fav}(t \rightarrow i)$ from the target user u_t to each person u_i $(i = 1, 2, ...)$ by analyzing the logs on the bookmarks that the target user u_t has done. Second, the system calculates Reaction Score $S_{rxn}(t \rightarrow i)$ from the target user u_t to each person u_i $(i = 1, 2, ...)$, by combining Replying Score $S_{rep}(t \rightarrow i)$ and Bookmarking Score $S_{fav}(t \rightarrow i)$.

Finally, it sorts the persons u_i $(i = 1, 2, ...)$ based on Reaction Score $S_{rxn}(t \rightarrow i)$, and recognizes several persons whose Reaction Score over a threshold $s_t(\epsilon)$ as the influencers for the target user u_t. Here, the threshold $s_t(\epsilon)$ changes depending on the target user u_t, and it is controlled by a parameter $\epsilon \in [0, 100]$. When the system regards the minimum value among Reaction Scores $S_{rxn}(t \rightarrow i)$ from a target user u_t to every persons u_i $(i = 1, 2, ...)$ as 0% and the maximum value among them as 100%, the parameter ϵ decides what percentage the system employs as the threshold of Reaction Score $S_{rxn}(t \rightarrow i)$.

The remainder of this section proposes several methods for calculating Reaction Score $S_{rxn}(t \rightarrow i)$ form a target user u_t to each person u_i $(i = 1, 2, ...)$ based on Replying Score $S_{rep}(t \rightarrow i)$ and Bookmarking Score $S_{fav}(t \rightarrow i)$.

Step 1 **Calculating Replying Score and Bookmarking Score from a target user u_t to a person u_i**

Other users u_i	Replying score $S_{rep}(t \to i)$
A	0.214
B	0.153
C	0.145

Other users u_i	Bookmarking score $S_{fav}(t \to i)$
D	0.120
E	0.063
A	0.048

Step 2 **Calculating Reaction Score $S_{rxn}(t \to i)$ from the target user u_t to each person u_i**

Candidates u_i of Influencers	Reaction score $S_{rxn}(t \to i)$
A B E	Comparing 6 calculation methods

⇩

Step 3 **Recognizing several persons whose Reaction Score over a threshold as the influencers for the target user u_t**

Influencers	Reaction score $S_{rxn}(t \to i)$
A B	Threshold of more
~~E~~	

Figure 3.1: The overview of Influencer Estimation Based on Reaction

3.2.2 Replying Score and Bookmarking Score

Calculation of Replying Score

For the persons whose posts have been replied by a target user u_t at least once, the system calculates Replying Score $S_{rep}(t \to i)$ from the target user u_t to the persons u_i $(i = 1, 2, ...)$. It is calculated based on the percentage of replies to the posts written by the person u_i in the all of replies that the target user u_t has done. Here, $C_{rep}(t \to i)$ in this formula is the number of replies from the target user u_t to the posts written by the person u_i, and U^t_{rep} in this formula is the set of the persons whose posts have been replied by the target user u_t at least once.

$$S_{rep}(t \to i) = \frac{C_{rep}(t \to i)}{\sum\limits_{u_k \in \boldsymbol{U^t_{rep}}} C_{rep}(t \to k)}$$

$$\boldsymbol{U^t_{rep}} = \{u_k \mid {}^{\forall}k, C_{rep}(t \to k) \geq 1\}$$

Calculation of Bookmarking Score

For the persons whose posts have been bookmarked by a target user u_t at least once, the system calculates Bookmarking Score $S_{fav}(t \to i)$ from the target user u_t to the persons u_i $(i = 1, 2, ...)$. It is calculated based on the percentage of bookmarks for the posts written by the person u_i in the all of bookmarks that the target user u_t has done. Here, $C_{fav}(t \to i)$ in this formula is the number of bookmarks from the target user u_t to the posts written by the person u_i, and U^t_{fav} in this formula is the set of the persons whose posts have been bookmarked by the target user u_t at least once.

$$S_{fav}(t \to i) = \frac{C_{fav}(t \to i)}{\sum\limits_{u_k \in \boldsymbol{U^t_{fav}}} C_{fav}(t \to k)}$$

$$\boldsymbol{U^t_{fav}} = \{u_k \mid {}^{\forall}k, C_{fav}(t \to k) \geq 1\}$$

3.2.3 Calculation of Reaction Score

This subsection compares six kinds of methods for calculating Reaction Score $S_{rxn}(t \to i)$ from a target user u_t to a person u_i. There are the differences of the processes for combining Replying Score $S_{rep}(t \to i)$ with Bookmarking Score $S_{fav}(t \to i)$ in these methods, and their overviews are shown in Table 3.1.

Table 3.1: The overviews of the methods for calculating Reaction Score

Method	Overview
REP	Using only Replying Score
FAV	Using only Bookmarking Score
RaF	Using Product of Replying Score and Bookmarking Score
RoF	Using Larger of Replying Score or Bookmarking Score
RaF(w)	Adjusting Weight to Each Score in **RaF**
RoF(w)	Adjusting Weight to Each Score in **RoF**

Method Using only Replying Score

This method substitutes Replying Score $S_{rep}(t \to i)$ from a target user u_t to a person u_i for Reaction Score $S_{Rxn}(t \to i)$, and it is called **REP**. In this method, for each of the persons whose posts have been replied by the target user u_t at least once, the system calculates Reaction Score $S_{rxn}(t \to i)$ from the target user u_t to the person u_i. Then the persons whose posts have been replied by the target user frequently are estimated easily as the influencers for the target user.

$$S_{rxn}(t \to i) = \begin{cases} S_{rep}(t \to i) & (u_i \in \boldsymbol{U^t_{rep}}) \\ 0 & (\text{otherwise}) \end{cases}$$

Method Using only Bookmarking Score

This method substitutes Bookmarking Score $S_{fav}(t \to i)$ from a target user u_t to a person u_i for Reaction Score $S_{Rxn}(t \to i)$, and it is called **FAV**. In this method, for each of the persons whose posts have been bookmarked by the target user u_t at least once, the system calculates Reaction Score $S_{rxn}(t \to i)$ from the target user u_t to the person u_i. Then the person whose posts have been bookmarked by the target user frequently are estimated easily as the influencers for the target user.

$$S_{rxn}(t \to i) = \begin{cases} S_{fav}(t \to i) & (u_i \in \boldsymbol{U^t_{fav}}) \\ 0 & (\text{otherwise}) \end{cases}$$

Method Using Product of Replying Score and Bookmarking Score

This method substitutes the product of Replying Score $S_{rep}(t \to i)$ and Bookmarking Score $S_{fav}(t \to i)$ for Reaction Score $S_{Rxn}(t \to i)$ from the target user u_t to the person u_i, and it is called **RaF**. In this method, for each of the persons whose posts have been replied **and** bookmarked by the target user u_t at least once, the system calculates Reaction Score $S_{rxn}(t \to i)$ from the target user u_t to the person u_i. Then the person whose posts have been replied and bookmarked by the target user frequently are estimated easily as the influencers for the target user.

$$S_{rxn}(t \to i) = \begin{cases} S_{rep}(t \to i) \times S_{fav}(t \to i) & \\ & (u_i \in \boldsymbol{U^t_{rep}} \text{ and } u_i \in \boldsymbol{U^t_{fav}}) \\ 0 & (\text{otherwise}) \end{cases}$$

Method Using Larger of Replying Score or Bookmarking Score

This method substitutes the larger of Standard Score $Z_{rep}(t \to i)$ on Replying Score $S_{rep}(t \to i)$ or Standard Score $Z_{fav}(t \to i)$ on Bookmarking Score $S_{fav}(t \to i)$ for Reaction Score $S_{Rxn}(t \to i)$ from the target user u_t to the person u_i, and it is called **RoF**. In this method, for each of the persons whose posts have been replied **or** bookmarked by the target user u_t at least once, the system calculates Reaction Score $S_{rxn}(t \to i)$ from the target user u_t to the person u_i. Then the persons whose posts have been replied or bookmarked by the target user frequently, are estimated easily as the influencers for the target user.

$$S_{rxn}(t \to i) = \begin{cases} \max\{Z_{rep}(t \to i), Z_{fav}(t \to i)\} & \\ & (u_i \in \boldsymbol{U^t_{rep}} \text{ or } u_i \in \boldsymbol{U^t_{fav}}) \\ 0 & (\text{otherwise}) \end{cases}$$

Here, Standard Score $Z_{rep}(t \to i)$ and Standard Score $Z_{fav}(t \to i)$ are used to compare Replying Score $S_{rep}(t \to i)$ and Bookmarking Score $S_{fav}(t \to i)$ uniformly. This chapter converts Replying Score $S_{rep}(t \to i)$ from the target user u_t to each person u_i $(i = 1, 2, ...)$ to Z-value, to make the mean μ^t_{rep} of their scores $(S_{rep}(t \to i))$ into 0.0 and the variance v^t_{rep} of them into 1.0, by the following formula. And it also converts Bookmarking Score $S_{fav}(t \to i)$ by the same way. Here, μ^t_{fav} shows the mean of their scores $(S_{fav}(t \to i))$ and v^t_{rep} shows the variance of them.

$$Z_{rep}(t \to i) = \frac{S_{rep}(t \to i) - \mu^t_{rep}}{v^t_{rep}}$$

$$Z_{fav}(t \to i) = \frac{S_{fav}(t \to i) - \mu^t_{fav}}{v^t_{fav}}$$

Method Adjusting Weight to Each Score in RaF

This method adds a weight that shows which is important the persons whom the target user replies to or the persons whom the target user bookmarks, to the method **RaF**, and it is called **RaF(w)**. When calculating the product of Replying Score $S_{rep}(t \to i)$ and Bookmarking Score $S_{fav}(t \to i)$, this method can adjust its ratio by using a parameter $w_a \in [0, 1]$.

$$S_{rxn}(t \to i) = \begin{cases} S_{rep}(t \to i)^{w_a} \times S_{fav}(t \to i)^{1-w_a} \\ \qquad\qquad (u_i \in \boldsymbol{U^t_{rep}} \text{ and } u_i \in \boldsymbol{U^t_{fav}}) \\ 0 \qquad\qquad \text{(otherwise)} \end{cases}$$

Method Adjusting Weight to Each Score in RoF

This method adds a weight that shows which is important the persons whom the target user replies to or the persons whom the target user bookmarks, to the method **RoF**, and it is called **RoF(w)**. When comparing Standard Score $Z_{rep}(t \to i)$ on Replying Score $S_{rep}(t \to i)$ and Standard Score $Z_{fav}(t \to i)$ on Bookmarking Score $S_{fav}(t \to i)$, this method can adjust their values by using a correction value $z_t(w_o)$.

$$S_{rxn}(t \to i) = \begin{cases} \max\{Z_{rep}(t \to i), Z_{fav}(t \to i) + z_t(w_o)\} \\ \qquad\qquad\qquad (u_i \in \boldsymbol{U^t_{rep}} \text{ or } u_i \in \boldsymbol{U^t_{fav}}) \\ 0 \qquad\qquad\qquad \text{(otherwise)} \end{cases}$$

Here, the correction value $z_t(w_o)$ changes depending on the target user u_t, it is controlled by a parameter $w_o \in [0, 1]$. When the value of the parameter w_o is 0.0, this method regards that the persons whose posts have been replied by the target user, are more important than the persons whose posts have been bookmarked by the target user. When the value of the parameter w_o is 1.0, it means the opposite.

3.3 Influencer Estimation Based on Interest

3.3.1 Overview of Analysis of Interest

This subsection proposes a method for estimating the influencers for a target user u_t based on Interest Score $S_{int}(t \to i)$ from the target user u_t to each person u_i ($i = 1, 2, ...$). Here, Interest Score $S_{int}(t \to i)$ shows how much the target user u_t is interested in the person u_i. This method has four steps (Figure 3.2).

First, the system extracts several feature words of each person u_i from the documents including the person's features. Second, the system extracts several frequently appearing words from the documents including the interests of a target user u_t. Third, it calculates Interest Score S_{int} $(t \to i)$ from the target user u_t to the person u_i based on the frequency of appearance of a feature word of the person u_i in the documents including the interests of the target user u_t.

Finally, it sorts the persons u_i ($i = 1, 2, ...$) based on Interest Score $S_{int}(t \to i)$, and recognizes several persons whose Interest Score over a threshold $s_t(\delta)$ as the influencers for the target user u_t. Here, the threshold $s_t(\delta)$ changes depending on the target user u_t, and it is controlled a parameter $\delta \in [0, 100]$. When the system regards the minimum value among Interest Scores $S_{int}(t \to i)$ from a target user u_t to every persons u_i ($i = 1, 2, ...$) as 0% and the maximum value among them as 100%, the parameter δ decides what percentage the system employs as the threshold of Interest Score $S_{int}(t \to i)$.

The remainder of this section investigates various documents that would include the features of a person u_i and various documents that would include the interests of a target user u_t. In addition, it proposes several methods for calculating Interest Score $S_{int}(t \to i)$ from a target user u_t to each person u_i.

Step 1 **Extracting several feature words from the documents that include the features of a person u_i**

YouTuber F		Golfer G	
Word	TFIDF	Word	TFIDF
Makeup	5.325	Driver	3.230
Moisturizing	2.264	Putter	1.174
Lip balm	0.256	US Open	0.159

Step 2 **Extracting several frequently appearing words from the documents that include the interests of a target user u_t**

Word	Frequency
Makeup	15
Skin lotion	7
⋮	

Step 3 **Calculating Interest Score $S_{int}(t \rightarrow i)$ from the target user u_t to each person u_i**

Candidates u_i of Influencers	Interest score $S_{int}(t \rightarrow i)$
F	15×5.325
G	0
⋮	

Step 4 **Recognizing several persons whose Interest Score over a threshold as the influencers for the target user u_t**

Influencers	Interest score $S_{int}(t \rightarrow i)$
F	Threshold of more
~~G~~	
⋮	

Figure 3.2: The overview of Influencer Estimation Based on Interest

3.3.2 Documents Including Interest/Feature of User

Documents Including Interest of Target User

I list three types of documents that may include the interests of a target user.

1. The posts that have been written by the target user
2. The comments in the profile of the SNS account of the target user
3. The posts that have been bookmarked by the target user

It appears that there are various words showing a target user's favorite things, in the posts written by him/her and the comments in the profile of his/her SNS account. In addition, the posts that have been bookmarked by a target user would include various words showing the target user's interests. Therefore, it is expected that the system can extract the target user's interests by analyzing these documents.

On the other hand, there are some problems when analyzing these documents. For instance, the comments in the profile of the SNS account of a target user have not only his/her preferences (i.e., favorite things) but also his/her age and jobs. Furthermore, there may be some soliloquies in the posts written by a target user. Moreover, there is a case that a target user bookmarks a post to show the respect for the person who wrote it even if the target user is not interested in the topics of the post. In this case, it appears that the posts that have been bookmarked by the target user do not always have the target user's interests. Therefore, this chapter also conducts an experiment for revealing the suitable documents for analyzing each target user's interests.

Documents Including Feature of Person

I list three types of documents that may include the features of a person.

1. The handle of the SNS account of the person
2. The comments in the profile of the SNS account of the person
3. The posts that have been written by the person

It appears that the handle of the SNS account of a person directly expresses the person. In addition, the comments in the profile of the SNS account of a person and the posts that have been written by a person, would include what kind of person he/she is (i.e., his/her personality). On the other hand, there is a case that some users use false names as the handles and they remark some jokes in their posts. Therefore, this chapter also conducts an experiment for revealing the documents that the system can exactly extract each person's features.

3.3.3 Calculation of Interest Score

This subsection compares three kinds of methods for calculating Interest Score $S_{int}(t \to i)$ from a target user u_t to a person u_i, and their overviews are shown in Table 3.2. **FW** does not connect with the discussions so far, it tries to simply calculate Interest Score $S_{int}(t \to i)$ from a target user u_t to a person u_i. **DC** is based on the discussions so far, it calculates Interest Score $S_{int}(t \to i)$ from a target user u_t to a person u_i based on the frequency of appearance of a feature word showing the person u_i in the documents including the interests of the target user u_t. Finally, $\mathbf{DC_F}$ combines the method **FW** with the method **DC**. The remainder of this subsection explains the detail of each method.

Table 3.2: The overviews of the methods for calculating Interest Score

Method	Overview
FW	Using only Number of Followers of Person
DC	Analyzing Document Including Interest of User
$\mathbf{DC_F}$	Combining **FW** and **DC**

Method Using only Number of Followers of Person

This method substitutes the number of followers $N_{fw}(i)$ of a person u_i for Interest Score $S_{int}(t \to i)$ from a target user u_t to the person u_i, and it is called **FW**. It appears that the number of followers of a person shows the influence of the person to society, because it is the number of those who want to look at the posts written by the person. This method is based on the way of thinking that if a person has an influence on society, a target user also would be influenced by the person. Therefore, the persons who have the large numbers of followers are estimated easily as the influencers for the target user. Here, the system needs to narrow down the candidates of the influencers for a target user, to several persons who he/she is already interested in. Hence, for each of the persons who are followed by a target user u_t, the system calculates Interest Score $S_{int}(t \to i)$ from the target user u_t to the person u_i.

$$S_{int}(t \to i) = \begin{cases} N_{fw}(i) & (u_i \in \boldsymbol{U^t_{fw}}) \\ 0 & (\text{otherwise}) \end{cases}$$

Method Analyzing Document Including Interest of User

This method is based on the core proposal of this subsection (Figure 3.2), and it is called **DC**. In this method, it is expected that the system can calculate each target user's interest for a person more exactly than the method analyzing only the number of followers of the person, because it analyzes individual social data. Here, when explaining this method, let's assume that the documents that including a person's features and the documents including a target user's interests, are already selected from their candidates (3.3.2).

First, this method calculates Feature Value $tfidf_i(w)$, which shows how much a word w expresses the feature of a person u_i. Feature Value $tfidf_i(w)$ is calculated based on the product of Term Frequency $tf_i(w)$ of the word w and Inverse Document Frequency $idf(w)$ of the word w.

$$tfidf_i(w) = tf_i(w) \times idf(w)$$

Term Frequency $tf_i(w)$ of the word w shows the rate of appearances of the word w in the all words $\boldsymbol{W_i}$ in the documents including the features of the person u_i. Here, $N_{i,w}$ in this formula is the frequency of appearance of a word w in the documents including the features of a person u_i.

$$tf_i(w) = \frac{N_{i,w}}{\sum\limits_{w_k \in \boldsymbol{W_i}} N_{i,w_k}}$$

Inverse Document Frequency $idf(w)$ of the word w shows the inverse of the rate of documents that have the word w in the all documents $\boldsymbol{D}$ including the features of each of all persons. Here, $df(w)$ in this formula is the number of documents that have a word w

$$idf(w) = \log_2 \frac{|\boldsymbol{D}|}{df(w)}$$

Second, this method calculates Interest Score $S_{int}(t \rightarrow i)$ from the target user u_t to the person u_i by the following formula. Here, $c_t(w)$ in this formula is the frequency of appearance of a word w in the documents including the interests of a target user u_t.

$$S_{int}(t \rightarrow i) = \begin{cases} \sum\limits_{w \in \boldsymbol{W_t}} tfidf_i(w) \times c_t(w) & \\ & (u_i \in \boldsymbol{U^t_{fw}}) \\ 0 & (\text{otherwise}) \end{cases}$$

Method Combining FW and DC

This method has two concepts. One is that if a person has an influence on society, a target user also would be interested in the person (i.e., it is the same concept with the method **FW**). The other is that the system analyzes each target user's interests, individually (i.e., it is the same concept with the method **DC**). Therefore, the system substitutes the product of the number of followers $N_{fw}(i)$ of a person u_i and Interest Score $S_{int}(t \rightarrow i)$ of the method **DC**, for Interest Score $S_{int}(t \rightarrow i)$ of this method. Here, $\alpha in[0,1]$ in this formula shows a contribution rate of the number of followers $N_{fw}(i)$ of a person u_i, when calculating Interest Score $sint(t \rightarrow i)$.

$$S_{int}(t \rightarrow i) = \begin{cases} N_{fw}(i)^{\alpha} \times \sum\limits_{w \in \boldsymbol{W_t}} tfidf_i(w) \times c_t(w) & \\ & (u_i \in \boldsymbol{U^t_{fw}}) \\ 0 & (\text{otherwise}) \end{cases}$$

3.4 Overview of Evaluation Experiment

3.4.1 Dataset

This experiment uses Twitter, it has seven subject users who use Twitter on a daily basis. The proposed methods that estimate the influencers for each subject user based on analyzing each subject user's reactions to other persons, need social data about each subject user's replies to the posts written by other persons and his/her bookmarks for the posts written by other persons. On the other hand, the proposed methods that estimate the influencers for each subject user based on analyzing each subject user's interests for other persons, need several documents for analyzing each subject user's interests and several documents for analyzing the feature words of each of the persons other than the subject users.

These data were extracted using Twitter API in July 2019. Table 3.3 shows the mean number of persons whose posts have been replied by seven subject users, the mean number of persons whose posts have been bookmarked by seven subject users, and the mean number of persons who are followed by seven subject users. In addition, Table 3.4 shows the mean frequency of replies that seven subject users have done, and the mean frequency of bookmarks that seven subject users have done. Furthermore, Table 3.5 and Table 3.6 show the mean number of documents for analyzing the interests of seven subject users and the mean number of documents for analyzing the feature words of the persons other than the subject users.

Table 3.3: The persons who are related to the subject users

	Mean	SD
# Replied Persons	31.429	18.446
# Bookmarked Persons	273.714	171.736
# Following Persons	188.286	84.770

Table 3.4: The mean numbers of replies/bookmarks of the subject users

	Mean	SD
# Replying / # Posting	0.409	0.216
# Bookmarking / Day	2.508	3.502

Table 3.5: The extracted data on the subject users

# **Docs**	Mean	SD
Posted Tweets	1763.429	546.530
Bookmarked Tweets	1084.571	606.455
Profile	1.000	0.000
# **Words**	Mean	SD
In Posted Tweets	4069.143	1536.651
In Bookmarked Tweets	3616.143	2108.413
In Profile	5.714	5.573

Table 3.6: The extracted data on the persons other than the subject users

# **Docs**	Mean	SD
Posted Tweets	1744.736	6571.104
Handle (Name)	1.000	0.000
Profile	1.000	0.000
# **Words**	Mean	SD
In Posted tweets	6768.659	4538.493
In Handle (Name)	1.480	1.290
In Profile	7.280	6.111

3.4.2 Correct Data and Criteria

First, this subsection explains the correct data (i.e., the persons who should be estimated by the proposed methods). Each subject user lists 10 persons whom the subject user is actually influenced by, and this experiment evaluates how exactly the proposed methods can estimate the 10 actual influences for each subject user. Actually, the number of actual influencers for each subject user is not always 10. However, this experiment fixes the number of actual influencers of each subject user to 10. The reason is that the standard when each subject user selects the actual influencers changes depending on the subject user, if this experiment allows each subject user to list his/her actual influencers freely.

Second, I explain three kinds of criteria. This experiment employs Recall and Precision as the criteria for evaluating the estimation performances of the proposed methods. Here, Hits in this formulas shows the number of influencers for a subject user, whom the proposed method was able to estimate correctly. Recall is based on the rate of Hits in the 10 influencers selected by a subject user (i.e., correct data). Precision is based on the rate of Hits in the persons estimated by the proposed method.

$$\begin{aligned}\text{Recall} &= \frac{\#\ \text{Hits}}{\#\ \text{Actual influencers}}\\ \text{Precision} &= \frac{\#\ \text{Hits}}{\#\ \text{Estimated persons}}\end{aligned}$$

In addition, this experiment evaluates F-measure as a comprehensive criterion between Recall and Precision, and it is calculated by the following formula.

$$\text{F-measure} = \frac{2 \times \text{Recall} \times \text{Precision}}{\text{Recall} + \text{Precision}}$$

3.4.3 Contents in Evaluation Experiment

First, I evaluate the performances of the proposed methods that estimate the influencers for each subject user based on analyzing his/her reactions to other persons in SNS. Specifically, I compare the estimation performances of the four methods (**REP**, **FAV**, **RaF**, and **RoF**). In addition, I evaluate how much the method **RaF(w)** and the method **RoF(w)** improve the estimation performance of the method **RaF** and the estimation performance of the method **RoF**, respectively.

Second, I evaluate the performances of the proposed methods that estimate the influencers for each subject user based on analyzing his/her interests for other persons in SNS. Specifically, I compare the estimation performances of the three methods (**FW**, **DC**, and $\mathbf{DC_F}$). In addition, I discuss the change of the estimation performance of the proposed method depending on the documents that are analyzed by the system (e.g., the difference of the type of documents for analyzing each subject user's interests, and the difference of the type of documents for analyzing the feature words of each of the persons other than the subject users). Moreover, I evaluate the estimation performance of a method combining the persons estimated based on analyzing each subject user's reactions and the persons estimated based on analyzing each subject user's interests.

3.5 Evaluation of Method Based on Reaction

This section evaluates the performances of the proposed methods that estimate the influencers for each subject user based on analyzing his/her reactions to other persons.

3.5.1 Experiments on REP/FAV/RaF/RoF

This subsection compares the estimation performances of the four methods (**REP**, **FAV**, **RaF**, and **RoF**). Table 3.7 shows the mean F-measure of each method and its standard deviation. It is revealed the mean F-measures of the methods **FAV** and **RoF** are higher than the mean F-measures of the methods **REP** and **RaF**. Furthermore, focusing on the standard deviation of each method, we can confirm that the variabilities on F-measures of the methods **FAV**/**RoF** are less than the variabilities on F-measures of the methods **REP**/**RaF** (i.e., the estimation performances of the methods **FAV** and **RoF** are stable).

Table 3.7: The mean F-measures of **REP**/**FAV**/**RaF**/**RoF**

Method	Threshold	Mean	SD
REP	11.000	0.377	0.257
FAV	12.714	0.517	0.152
RaF	1.857	0.361	0.265
RoF	12.857	0.512	0.137

Threshold shows ϵ% (section 3.2).

Besides, I test its significant difference. Table 3.8 shows the result of one-way analysis of variance for the null hypothesis that there is no difference between each mean F-measure of the four methods (**REP**, **FAV**, **RaF**, and **RoF**). Unfortunately, the null hypothesis is not rejected because its p-value based on F-statistic is higher than 0.05 when testing in the significance level 5%. Therefore, it cannot be concluded that there is statistically significant difference between each mean F-measure of the four methods.

Table 3.8: ANOVA on F-measures of **REP/FAV/RaF/RoF**

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Method	3	0.150	0.050	1.120	0.360
Residuals	24	1.068	0.045		

Next, I compare the four methods by focusing on their the numbers of Hits. Table 3.9 shows the mean number of Hits of each method and its standard deviation. Here, it shows the number of Hits when each method gets the mean F-measure in Table 3.7. As with the discussions on the mean F-measure of each method, we can confirm that the mean numbers of Hits of the methods **FAV** and **RoF** are higher than the mean numbers of Hits of the methods **REP** and **RaF**.

Table 3.9: The mean numbers of Hits of **REP/FAV/RaF/RoF**

Method	Threshold	Mean	SD
REP	11.000	3.857	2.795
FAV	12.714	6.286	2.430
RaF	1.857	3.286	2.563
RoF	12.857	8.000	2.380

Threshold shows ϵ% (section 3.2).

Besides, I test its significant difference. Table 3.10 shows the result of one-way analysis of variance for the null hypothesis that there is no difference between each mean number of Hits of the four methods (**REP**, **FAV**, **RaF**, and **RoF**). Accordingly, we can confirm that the null hypothesis is rejected because its p-value based on F-statistics is lower than 0.01 when testing in the significance level 1%. Therefore, it is assumed that there is a statistically significant difference between each mean number of Hits of the 4 methods.

Table 3.10: ANOVA on the numbers of Hits of **REP**/**FAV**/**RaF**/**RoF**

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Method	3	100.7	33.57	5.174	0.007 **
Residuals	24	155.7	6.49		

** $p < 0.01$

Additionally, I conduct a multiple comparison test on the numbers of Hits of the four methods. Table 3.11 shows the result of the multiple comparison by Tukey's test for the null hypothesis that there is no difference between each mean number of Hits of the four methods (**REP**, **FAV**, **RaF**, and **RoF**). Diff (3rd row) in Table 3.11 is the difference between each mean number of Hits of the four methods. Lwr (4th row) and Upr (5th row) are the lower confidence limit and the upper confidence limit respectively, on 95% confidence interval.

We can confirm that the difference between the method **RoF** and **RaF**, and the difference between the method **RoF** and **REP**, have the statistically significant differences on the mean numbers of Hits, when testing in the significance level 5%.

Table 3.11: Tukey's test on the numbers of Hits of **REP**/**FAV**/**RaF**/**RoF**

Methods		Diff	Lwr	Upr	t value	p value
RaF	**FAV**	−3.000	−6.753	0.753	−2.203	0.151
REP	**FAV**	−2.429	−6.181	1.324	−1.784	0.305
RoF	**FAV**	1.714	−2.038	5.467	1.259	0.597
REP	**RaF**	0.571	−3.181	4.324	0.420	0.975
RoF	**RaF**	4.714	0.962	8.467	3.463	0.010 *
RoF	**REP**	4.143	0.390	7.896	3.043	0.027 *

* $p < 0.05$

The remainder of this subsection summarizes and interprets the discussions so far. Previous section has explained that the method **RaF** regards the persons whose posts has been replied and bookmarked by a target user frequently as the influencers for the target user. At first, this seemed to be appropriate as a condition for the person to become one of the influencers for the target user. However, this experiment revealed that the method **RaF** cannot hit the more influencers for each subject user. This result means this condition is too stick.

In addition, I has explained that the method **REP** regards the persons whose posts has been replied by a target user frequently as the influencers for the target user. Furthermore, I has explained that the method **FAV** regards the persons whose posts has been bookmarked by a target user frequently as the influencers for the target user. At first, they seemed to be too simple to estimate the influencers for each subject user. However, we confirmed the method **FAV** has relatively high mean F-measure. Therefore, it is assumed that there are many influencers for a target user in the persons whose posts has been bookmarked by the target user frequently.

Moreover, the method **RoF** combines the persons whose posts has been replied by a target user frequently, and the persons whose posts has been bookmarked by the target user frequently, without a condition too stick. It regards the persons whose posts has been replied or bookmarked by a target user frequently as the influencers for the target user. It seems to be the best policy for estimating the influencers for each user, because the both its mean F-measure and its mean number of Hits are the highest in the proposed methods.

3.5.2 Experiments on RaF(w)/RoF(w)

This subsection discusses about the estimation performances of the methods **RaF(w)** and **RoF(w)** that are proposed to improve the estimation performance of the methods **RaF** and **RoF**. These methods consider a weight (a parameter w_o) that determines which is important the persons whose posts has been replied by a subject user frequently or the persons whose posts has been bookmarked by the subject user frequently, when estimating the influencers for the subject user.

The first row in Table 3.12 is the result of one-sided t test for the null hypothesis that the mean F-measure of the method **RaF** is higher than the mean F-measure of the method **RaF(w)**. The method **RaF(w)** improves the estimation performance of the method **RaF**, however it does not have a statistical significance. On the other hand, the second row in Table 3.12 is the result of one-sided t test for the null hypothesis that the mean F-measure of the method **RoF** is higher than the mean F-measure of the method **RoF(w)**. We can confirm that the method **RoF(w)** improves the estimation performance of the method **RoF** with a statistical significance.

Table 3.12: The improvement effects of **RaF(w)/RoF(w)**

Methods		Diff	Df	t value	p value
RaF(w)	**RaF**	0.025	6	1.397	0.106
RoF(w)	**RoF**	0.060	6	2.885	0.014 *

* $p < 0.05$

The previous subsection described that the method **RoF** regarding the persons whose posts are frequency replied or bookmarked by a subject user, as the influencers for the subject user, is the best in all methods. On the other hand, this subsection reveals the method **RoF(w)** is superior to the method **RoF**. It is suggested that the estimation performance of **RoF** would be get even better, by considering which is importance the persons whose posts have been replied by the subject user or the persons whose posts have been bookmarked by the subject user, as the influencers for the subject user.

However, one of the problems in the method **RoF(w)** is that it is difficult to decide the value of the parameter w_o. Figure 3.3 shows the change of the performance the method **RoF(w)** depending on the value of the parameter w_o (its horizontal axis is the value of the parameter w_o and its vertical axis is F-measure). When the value of the parameter w_o is 0.0, the method **RoF(w)** gets easier to select the persons whose tweets has been replied by a subject user frequently, as the influencers for the subject user. On the other hand, when the value of the parameter w_o is 1.0, the method **RoF(w)** gets easier to select the persons whose tweets has been bookmarked by a subject user frequently, as the influencers for the subject user.

We can confirm that the parameter w_o is not a common value for all subject users. That is to say, the method **RoF(w)** needs to optimize the value of the parameter w_o for each subject user. Therefore, I am going to consider a method analyzing the logs of a user's behaviors in SNS more deeply, as a feature work. For an instance, the importance of bookmarking may differ between a subject user who bookmarks anything and everything and a subject user who does not. If a subject user bookmarks anything and everything, it is hard to estimate the persons whom the user is conscious. Therefore, in this case, it appears that the method **RoF(w)** should not emphasis the persons whose tweets has been bookmarked by the subject user when estimating the influencers for the subject user. As a future work, I am set to discuss about how to decide the value of the parameter w_o by considering not only the numbers of a user's replies and bookmarks but also the features of the replies and the bookmarks of the subject user.

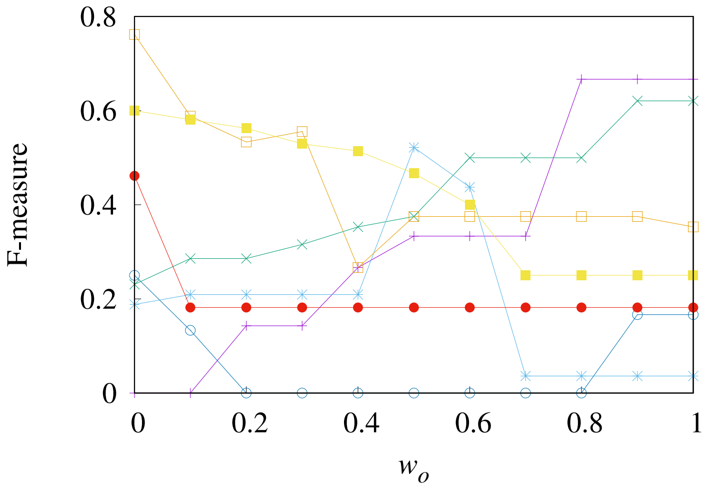


Figure 3.3: Effect of the parameter w_o on F-measure

3.6 Evaluation of Method Based on Interest

This section evaluates the performances of the proposed methods that estimate the influencers of each subject user based on analyzing his/her interests for other persons.

3.6.1 Experiments on FW/DC/DC$_\mathbf{F}$

This subsection compares the estimation performances of the 3 methods (**FW**, **DC**, and **DC**$_\mathbf{F}$). Table 3.13 shows the mean F-measure of each method and its standard deviation. It is suggested that the mean F-measures of the methods **DC** and **DC**$_\mathbf{F}$ are higher than the mean F-measure of the method **FW**. The tendency that the estimation performance of the method **FW** is low matches our intuition, because the method **FW** tries to estimate the influencers for each subject user by using only the number of followers of each of the persons other than the subject users, simply.

Besides, I test its significant difference. Table 3.14 shows the result of one-way analysis of variance for the null hypothesis that there is no difference between each mean F-measure of the 3 methods (**FW**, **DC**, and **DC**$_\mathbf{F}$). Accordingly, we can confirm that the null hypothesis is rejected because its p-value based on statistics F is lower than 0.001 when testing in significance level 0.1%. Therefore, it is assumed that there is a statistically significant difference between each mean F-measure of the 3 methods.

Table 3.13: The mean F-measures of **FW**/**DC**/$\mathbf{DC_F}$

Method	Threshold	Mean	SD
FW	14.286	0.166	0.056
DC	29.000	0.309	0.108
$\mathbf{DC_F}$	18.143	0.372	0.087

Threshold shows δ% (section **??**).

Table 3.14: ANOVA on F-measure of **FW**/**DC**/$\mathbf{DC_F}$

	Df	Sum Sq	Mean Sq	F value	Pr (>F)
Method	2	0.156	0.078	10.46	≈0 ***
Residuals	16	0.134	0.007		

*** $p < 0.001$

Additionally, I conduct multiple comparison test on each F-measure of the 3 methods. Table 3.15 shows the result of multiple comparison by Tukey's test for the null hypothesis that there is no difference between each mean F-measure of the 3 methods (**FW**, **DC**, and $\mathbf{DC_F}$). Diff (3rd row) in Table 3.15 is the difference between each F-measure of the 3 methods. Lwr (4th row) and Upr (5th row) are the lower confidence limit and the upper confidence limit respectively, on 95% confidence interval.

Consequently, the difference between the method **FW** and **DC** has a statistically significant difference on F-measure, when testing in significance level 5%. In addition, the difference between the method **FW** and $\mathbf{DC_F}$ has a statistically significant difference on F-measure, when testing in significance level 0.1%.

Table 3.15: Tukey's test on F-measures of **FW**/**DC**/$\mathbf{DC_F}$

Methods		Diff	Lwr	Upr	t value	p value
$\mathbf{DC_F}$	**DC**	0.063	−0.055	0.181	1.365	0.379
FW	**DC**	−0.143	−0.261	−0.025	−3.098	0.016 *
FW	$\mathbf{DC_F}$	−0.206	−0.324	−0.066	−4.464	≈0 ***

* $p < 0.05$
*** $p < 0.001$

3.6.2 Comparative Experiment on Analyzed Documents

The method **DC** and the method $\mathbf{DC_F}$ estimate the influencers for each subject user by analyzing the frequency of appearance of a feature word of a person in the documents including the subject user's interests. This subsection analyzes the estimation performance of the method $\mathbf{DC_F}$ depending on the difference of the types of analyzed documents. In other words, I aim to reveal the suitable documents for analyzing each subject user's interests, and the suitable documents for analyzing the feature words of each of the persons other than the subject users.

Table 3.16: The performance of $\mathbf{DC_F}$ depending on the analyzed documents

	T	**TC**	**TF**	**TCF**	**C**	**CF**	**F**	Mean
N	0.239	0.239	0.229	0.229	0.124	0.213	0.214	0.212
NC	0.278	0.277	0.302	0.302	0.199	0.322	0.318	0.285
NT	0.296	0.300	0.317	0.317	0.239	0.324	0.324	0.302
NCT	0.286	0.288	0.316	0.316	0.232	0.324	0.323	0.298
C	0.272	0.272	0.307	0.307	0.196	0.323	0.318	0.285
CT	0.284	0.287	0.318	0.318	0.232	0.329	0.326	0.299
T	0.287	0.290	0.315	0.315	0.239	0.329	0.326	0.300
Mean	0.278	0.279	0.301	0.301	0.209	0.309	0.307	

Table 3.17: Effectiveness of $\mathbf{DC_F}$ depending on the analyzed documents

	T	**TC**	**TF**	**TCF**	**C**	**CF**	**F**
N	0.139	0.139	0.139	0.139	0.927	0.136	0.136
NC	0.113	0.113	0.041 *	0.041 *	0.790	0.019 *	0.019 *
NT	0.090	0.093	0.058	0.058	0.820	0.049 *	0.049 *
NCT	0.119	0.120	0.058	0.058	0.808	0.049 *	0.049 *
C	0.040 *	0.040 *	0.013 *	0.013 *	0.802	0.014 *	0.014 *
CT	0.119	0.120	0.058	0.058	0.808	0.049 *	0.049 *
T	0.113	0.114	0.058	0.629	0.820	0.049 *	0.049 *

* $p < 0.05$

I explain Table 3.16 and Table 3.17. The columns of each Table are classified based on the types of documents for analyzing each subject user's interests. When a column includes **T**, the method $\mathbf{DC_F}$ regards the tweets written by each subject user, as the documents including the subject user's interests. When a column includes **C**, the method $\mathbf{DC_F}$ regards the comments in the profile of each subject user's Twitter account, as the documents including the subject user's interests. When a column includes **F**, the method $\mathbf{DC_F}$ regards the tweets that have been bookmarked by each subject user, as the documents including the subject user's interests.

On the other hand, the rows of each Table are classified based on the types of documents for analyzing the feature words of each of the persons other than the subject users. When a row includes **N**, the method $\mathbf{DC_F}$ analyzes the handle of Twitter account of each of the persons other than the subject users, to extract the feature words of the person. When a row includes **C**, the method $\mathbf{DC_F}$ analyzes the comments in the profile of Twitter account of each of the persons other than the subject users, to extract the feature words of the person. When a row includes **T**, the method $\mathbf{DC_F}$ analyzes the tweets written by each of the persons other than the subject users, to extract the feature words of the person.

The value of a cell in Table 3.16 shows the mean F-measure of the method $\mathbf{DC_F}$ when it estimates the influencers for each subject user by analyzing the documents that are shown in the row of this cell and the column of this cell. In addition, the value of a cell in Table 3.17 is calculated as follows. First, the method $\mathbf{DC_F}$ estimates the influencers for each subject user by analyzing the documents that are shown in the row of this cell and the column of this cell. Second, I combine the persons estimated by the method $\mathbf{RoF(w)}$ and the persons estimated by the method $\mathbf{DC_F}$. Then they are defined as the persons estimated by a method $\mathbf{RoF(w){+}DC_F}$. Third, I conduct one-sided t test for the null hypothesis that the mean F-measure of the method $\mathbf{RoF(w)}$ is higher than the mean F-measure of the method $\mathbf{RoF(w){+}DC_F}$. Finally, its p value is shown as the value of this cell.

Therefore, if p-value of a cell in Table 3.17 is low, it means that the method $\mathbf{RoF(w){+}DC_F}$ improves the estimation performance of the method $\mathbf{RoF(w)}$. That is, when the method $\mathbf{DC_F}$ estimates the influencers for each subject user by analyzing the documents shown in the column of this cell and the row of this cell, it can estimate several persons who are not estimated by only the method $\mathbf{RoF(w)}$.

Focusing on the rows in Table 3.17, we can confirm that there are many cells whose p-values are less than 0.05 in the row of **NC** and the row of **C**. On the other hand, focusing on the columns in Table 3.17, we can also confirm that there are many cells whose p-values are less than 0.05 in the column of **CF** and the column of **F**. In addition, it indicates that when the row is **C** and the column is **TF** (or **TCF**), p value of this cell is the least. Therefore, it is revealed that the method $\mathbf{DC_F}$ had better conduct the following analyses to estimate a subject user's influencers who are not estimated by only the method **RoF(w)**. First, the method extracts the feature words of each of the persons other than the subject user, from the comments in the profile of the person's Twitter account. Second, it calculates the frequencies of appearance of these feature words in the tweets written by the subject user or the tweets bookmarked by the subject user. Finally, the method would be able to calculate the subject user's interests for other persons exactly.

3.7 Interaction of Proposed Methods

This subsection discusses about an interaction between the method that estimates the influencers based on a user's reactions to other persons and the method that estimates the influencers based on the user's interests for other persons.

Figure 3.4 shows F-measure of each of three types of methods for each subject user, and Figure 3.5 shows the number of Hits of each type of methods for each subject user. The first type is the method $\mathbf{RoF(w)}$ (i.e., the method estimating the influencers for a subject user by analyzing the user's reactions to other persons), and a set $\boldsymbol{I_r}$ means the influecenrs for the subject user who are estimated by this method. The second type is the method $\mathbf{DC_F}$ (i.e., the method estimating the influencers for a subject user by analyzing the user's interests for other persons), and a set $\boldsymbol{I_i}$ means the influencers for the subject user who are estimated by this method. The third type is a method regarding both the persons estimated by the method $\mathbf{RoF(w)}$ and the persons estimated by the method $\mathbf{DC_F}$ ($\boldsymbol{I_r} \cup \boldsymbol{I_i}$) as the influencers for the subject user. In addition, Table 3.18 shows the mean F-measure of each type of methods and the mean number of Hits of each type of methods.

Accordingly, we can confirm that the mean F-measure of the method regarding $\boldsymbol{I_r} \backslash \boldsymbol{I_i}$ as the influencers for the subject user is higher than both the mean F-measure of the method $\mathbf{RoF(w)}$ and the mean F-measure of the method $\mathbf{DC_F}$. In addition, $\boldsymbol{I_r} \backslash \boldsymbol{I_i} = 5.143$ means that the method analyzing a user's reactions to other persons helps to estimate approximately five influencers who are not estimated by only the method analyzing the user's interests for other persons. Furthermore, $\boldsymbol{I_i} \backslash \boldsymbol{I_r} = 1.000$ means that the method analyzing a user's interests for other persons helps to estimate approximately one influencer who is not estimated by only the method analyzing the user's reactions to other persons.

Table 3.18: The mean performances of methods estimating $\boldsymbol{I_r}/\boldsymbol{I_i}/\boldsymbol{I_r} \cup \boldsymbol{I_i}$

	F-measure	Hits
$\boldsymbol{I_r}$	0.572	6.714
$\boldsymbol{I_i}$	0.356	2.571
$\boldsymbol{I_r} \cup \boldsymbol{I_i}$	0.649	7.714
$\boldsymbol{I_r} \backslash \boldsymbol{I_i}$	–	5.143
$\boldsymbol{I_i} \backslash \boldsymbol{I_r}$	–	1.000

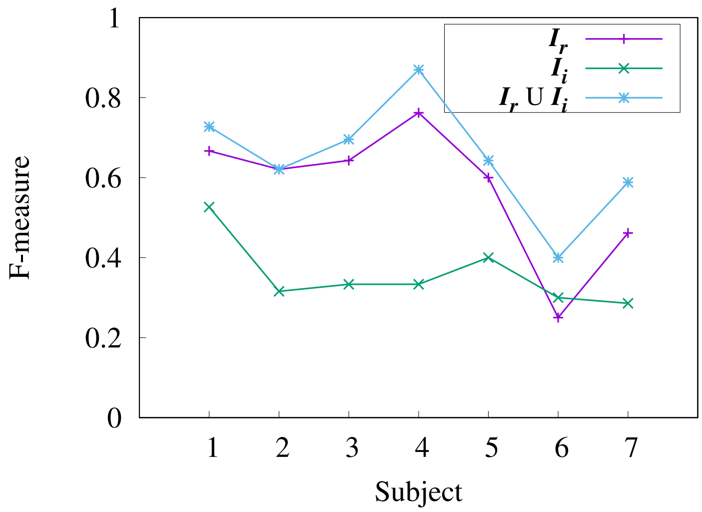


Figure 3.4: F-measure of each of methods estimating $\boldsymbol{I_r}/\boldsymbol{I_i}/\boldsymbol{I_r} \cup \boldsymbol{I_i}$

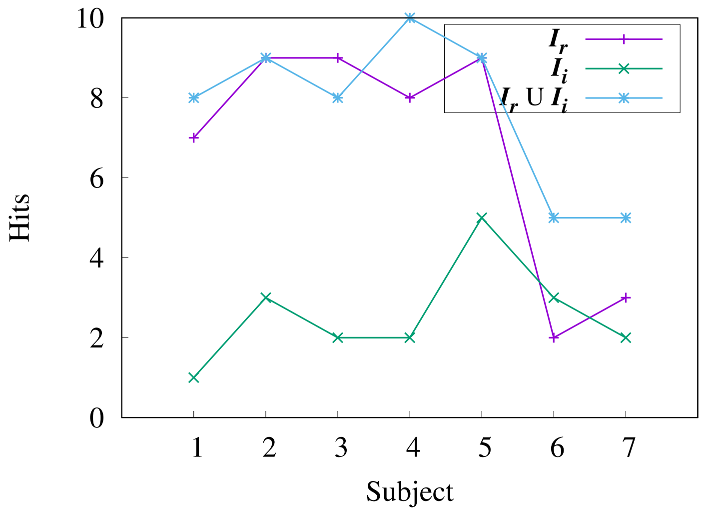


Figure 3.5: The number of Hits of each of methods estimating $\boldsymbol{I_r}/\boldsymbol{I_i}/\boldsymbol{I_r} \cup \boldsymbol{I_i}$

The reminder of this subsection considers the reasons of these results. I described this chapter aims to is to estimate not only a person who has an influence on society (e.g., a celebrity) but also a person who has an influence on only a user (e.g., a friend), as the influencers of the user. It appears that behaviors in SNS to each type of influencers depend on the user.

For example, a user may often write the posts including the topics about a celebrity whom the user is influenced by, and often reply to the posts written by a friend whom the user is influenced by. In this case, if the system aims to estimate the influencers for this user by only analyzing the user's reactions to other persons, it may not be able to estimate a social influencer (i.e., a person who has an influence on society). In addition, if the system aims to estimate the influencers for this user by only analyzing the user's interests for other persons, it may not be able to estimate a personalized influencer (i.e., a person who has an influence on only the user).

On the other hand, another user may often only bookmark both the posts written by his/her favorite celebrities and the posts written by his/her favorite friends. In this case, if the system aims to estimate the influencers for this user by only analyzing the user's interests for other persons, it may not be able to estimate both types of influencers For this reason, it is important for the system to analyze both a user's reactions to other persons and this user's interests for other persons, to estimate more types of influencers for this user exactly.

3.8 Qualitative Evaluation on This Study

3.8.1 Evaluation of Estimated Influencers

This subsection qualitatively discusses about the difference between the influencers estimated by a method that is close to previous studies and the influencers estimated by my proposed methods. Figure 3.6, Figure 3.7, Figure 3.8, and Figure 3.9 show four types of persons..

- **Social** The persons estimated by a method closing to previous studies
- **Corr** The influencers selected by the subject users (i.e., correct data)
- **Rxn** The persons estimated by analyzing the users' reactions
- **Int** The persons estimated by analyzing the users' interests

Here, the horizontal axis of each graph is the logarithm of the number of bookmarks that a person has received from surroundings, and the vertical axis of each graph is the logarithm of the number of replies that a person has received from surroundings.

First, the number of persons who belong to **Social** is 70. I select the 5 persons whose numbers of bookmarks per tweet are ranked in the top 5, from the persons who are followed by each subject user. In addition, I select the 5 persons whose numbers of replies per tweet are ranked in the top 5, from the persons who are followed by each subject user. They come to 70 persons in all (5 persons $\times$ 5 persons $\times$ 7 subject users), and they are plotted as **Social**. The persons belonging to **Social** are selected based on the number of reactions from surroundings. In other words, they are estimated by the approach that is similar to previous studies [27–29]. Therefore, most of them are in the upper right of Figure 3.6, naturally.

Second, the number of persons who belong to **Corr** is 70. They are the persons who selected by the seven subject users, as their actual influencers. The actual influences for each subject user are not only those who have an influence on society but also several familiar persons for him/her. The reason is because the experiment of this paper lets each subject user select the actual influencers by using his/her standard. Therefore, the actual influencers for the seven subject users are uniformly distributed from the lower left to the upper right of each graph. In addition we can confirm that it is difficult for the method that is close to previous studies to cover the actual influencers (Figure 3.6).

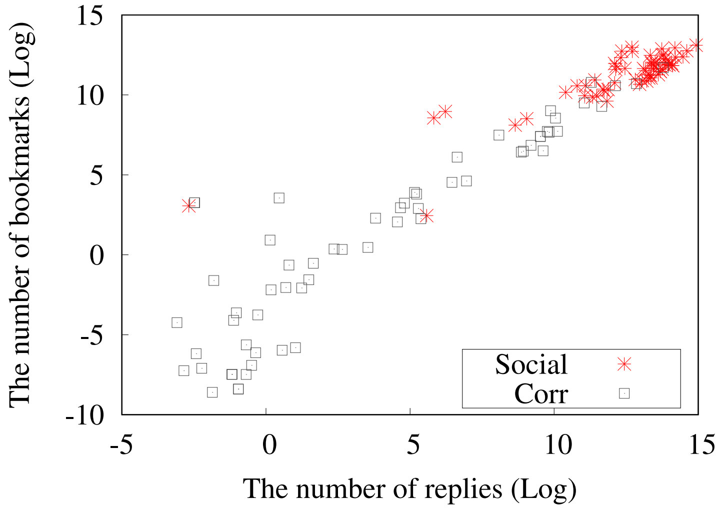


Figure 3.6: The influencers by the method based on previous studies

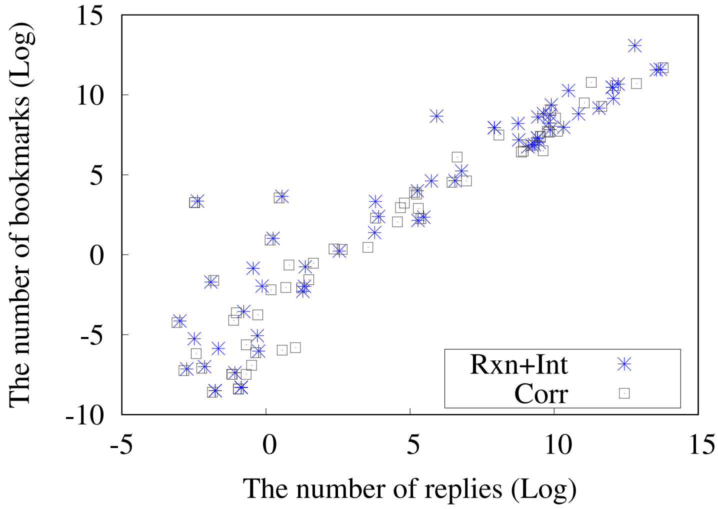


Figure 3.7: The influencers by the method based on both reactions/interests

Third, both the number of persons who belong to **Rxn** and the number of persons who belong to **Int** are 35, respectively. I select the 5 persons whose Reaction Score $S_{rxn}(t \to i)$ calculated by the method **RoF(w)** are ranked in the top 5, from the persons whose posts have been reacted by each subject. They come to the 35 persons in all (5 persons $\times$ 7 subject users), they are plotted as **Rxn**. Focusing on Figure 3.8, we can confirm a tendency that the points of **Rxn** cover its lower left. The result suggests that to analyze a user's reactions to other persons in a social network helps to estimate a familiar influencer for the user (i.e., a friend).

Finally, I select also the 5 persons whose Interest Score $S_{int}(t \to i)$ calculated by the method $\mathbf{DC_F}$ are ranked in the top 5, from the persons who are followed by each subject. They come to the 35 persons in all (5 persons $\times$ 7 subject users), they are plotted as **Int**. The method $\mathbf{DC_F}$ analyzes the documents that have been written by each subject user in SNS to extract the subject user's interests for other persons. I confirmed that several subject users talk about several famous accounts (e.g., a shop and a game of smart phone) in SNS, frequency. In addition, the method $\mathbf{DC_F}$ analyzes also the number of followers of a person. Therefore, focusing on Figure 3.9, we can confirm a tendency that the points of **Int** cover its upper right. The result suggests that to analyze a user's interests for other persons in a social network helps to estimate a social influencer (i.e., a celebrity).

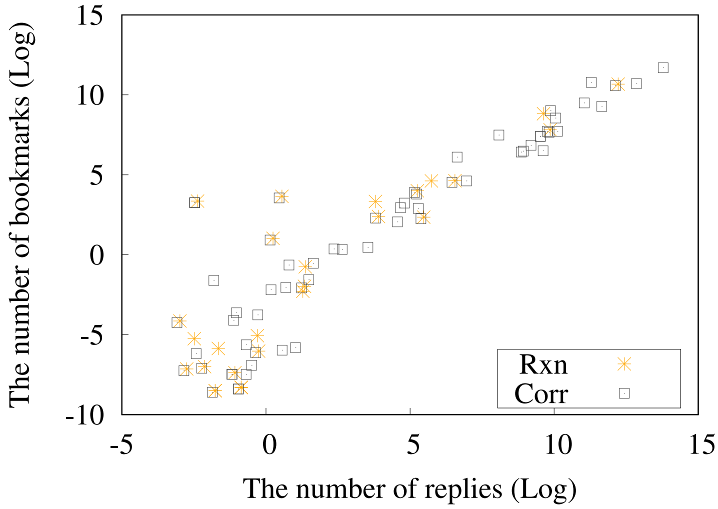


Figure 3.8: The influencers by the method based on only reactions

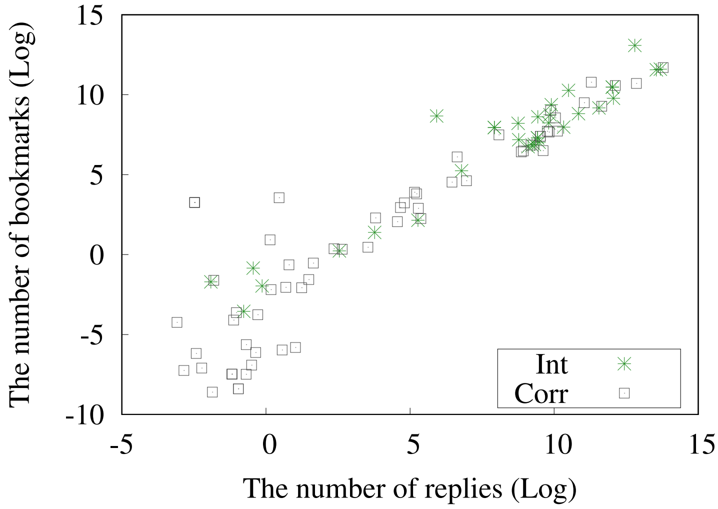


Figure 3.9: The influencers by the method based on only interests

3.8.2 Conditions for Proposed Methods to Work Well

This subsection discusses about the difference of the performance of the proposed method depending on the types of subject users. The seven subject users who participate in this experiment consist of 6 males and 1 females. In addition, the mean age is 24.7, and its standard deviation is 1.25. it appears that the attributes of the subject users are biased, because most of them are males in their 20s. However, it appears that the attribute of each subject user would not affect on the performances of the proposed methods. The proposed methods analyze both the persons whom each subject user has frequently reacted to, and the words in the documents including each subject user's interests. In addition, the methods do not have a parameter that changes depending on the attribute of each subject user. Therefore, it is assumed that the results of this experiment would not change even if the attribute of the subject user changes. On the other hand, subject users whose influencers are difficult to be estimated by the proposed methods may have 2 features.

First, it is difficult for the proposed method to estimate the influencers for a subject user whose number of replies to the posts written by other persons is small or whose number of bookmarks for the posts written by other persons is small. For an instance, there is a case that the total number of replies that a subject user u_t has done is 1 (this is a reply to the post written by a person u_i). In this case, the value of Reply Score $S_{rep}(t \to i)$ from the subject user u_t to the person u_i becomes 1.0 (1/1) based on the proposed formula in this chapter. Then the method may regard the person u_i as an influencer for the subject user u_t by mistake, although the person u_i is not one of the influencers for the subject user u_t (the false-negative rate would get higher), because there is no reliability on Reply Score $S_{rep}(t \to i)$.

Second, it is difficult for the proposed method to estimate the influencers for a subject user who has equality reacted to various persons. For an instance, there is a case that the total number of replies that a subject user u_t has done is 10, and the 10 persons replied by the subject user are all different. In this case, there is no way for specifying a person whom the subject user u_t is conscious.

There are several countermeasures when estimating the influencers for the subject users who have these features. For example, there is a method considering the confidence intervals of each proposed score (i.e., Interest Score and Reaction Score), when calculating them. In addition, there is a method narrowing down only other persons whose numbers of reactions from a subject user over a threshold when analyzing the subject user's reactions to other persons in SNS. I plan to implement these as the future works. Incidentally, it is revealed that the subject users who participate in this experiment do not have the 2 features, focusing on Table 3.19, Figure 3.10, Figure 3.11.

First, Table 3.19 shows the usage period of Twitter of a subject user, the number of tweets of the subject user, the number of replies of the subject user, and the number of bookmarks of the subject user. We can confirm that there is no subject user whose number of replies to the posts written by other persons or whose number of bookmarks for the posts written by other persons is too small. In other words, the seven subject users do not have the first feature.

Second, I explain Figure 3.10 and Figure 3.11. The vertical axis of Figure 3.10 shows the number of replies from a subject user to a person. Figure 3.10 plots the persons whose posts are replied by each subject user, in descending order of numbers of replies whom they received from the subject user. On the other hand, the vertical axis of Figure 3.11 shows the number of bookmarks from a subject user to a person. Figure 3.11 plots the persons whose posts are bookmarked by each subject user, in descending order of numbers of replies whom they received from the subject user. Here, even if the number of replies/bookmarks from a subject user to a person is over 100, the graph regards it 100. Consequently, we can confirm that there is no subject user who has equality reacted to various persons. In other words, the seven subject users do not have the second feature.

Table 3.19: The usage stats of SNS for each subject user

Sub	# Bookmarks	# Replies	# Posts	# Days
1	149	459	1960	2094
2	194	512	1998	1279
3	1497	1508	1990	699
4	1620	950	1987	489
5	1054	99	425	916
6	1406	378	1992	654
7	1672	610	1992	937

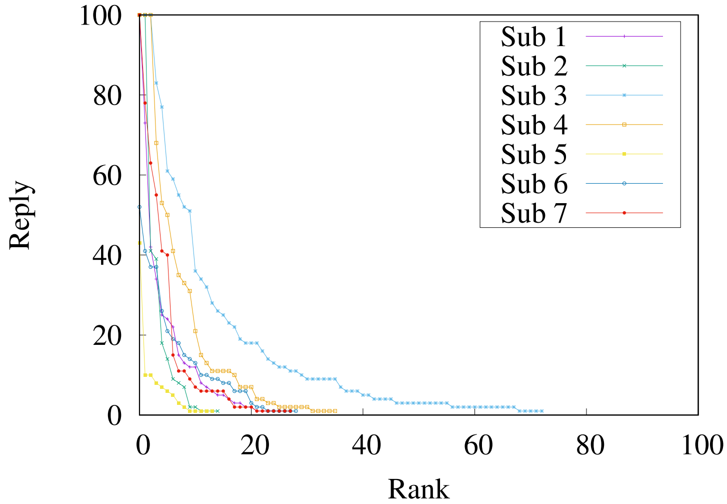


Figure 3.10: The number of each subject user's replies to a person

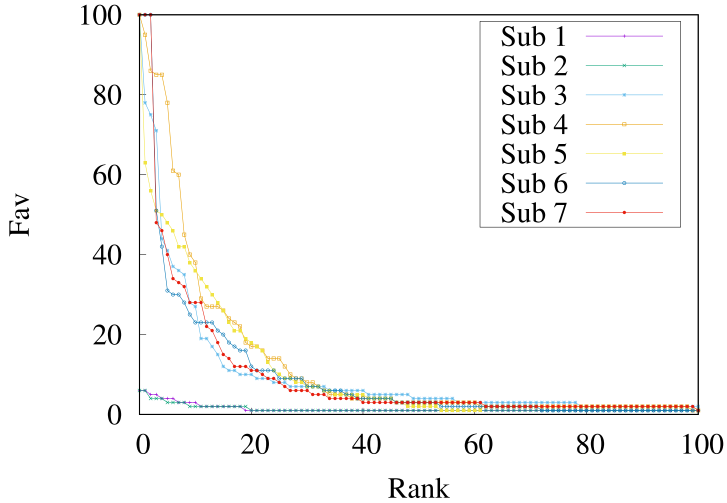


Figure 3.11: The number of each subject user's bookmarks to a person

3.9 Improvement of Proposed Methods

This section discusses about several improvements of the proposed methods.

3.9.1 Improvement on Analysis of Reaction

There is a problem in the method for calculating Reaction Score $S_{rxn}(t \to i)$ from a subject user u_t to a person u_i, which shows how many the subject user u_t has reacted to the posts written by the person u_i. The calculation of Reaction Score $S_{rxn}(t \to i)$ utilizes Replying Score $S_{rep}(t \to i)$ and Bookmarking Score $S_{fav}(t \to i)$, I have explained the methods for calculating them. As one of the future words, I need to consider the problem that the calculation of them depends on the total number of tweets written by a person u_i.

For an instance, there is a case that the total number of tweets written by a person u_1 is 10 and the number of replies from a subject user u_t to the person u_1 is 5. In addition, the total number of tweets written by a person u_2 is 100 and the number of replies from the subject user u_t to the person u_2 is 10. This case has ambiguity. The first way of thinking in this case is that the person u_2 is more important person for the subject user u_t than the person u_1. The reason is because the number of replies from the subject user u_t to the person u_2 (10) is higher than the number of replies from the subject user u_t to the person u_1 (5). The second way of thinking in this case is that the person u_1 is more important person for the subject user u_t than the person u_2. The reason is because the rate of the number of replies from the subject user u_t in the to total number of tweets written by the person u_1 (5/10) is higher than the rate of the number of replies from the subject user u_t in the to total number of tweets written by the person u_2 (10/100). I plan to consider them more deeply and discuss about a method for calculating each score more exactly.

3.9.2 Improvement on Analysis of Interest

There is a problem in the method for calculating Interest Score $S_{int}(t \to i)$ from a subject user u_t to a person u_i, which how much the target user u_t is interested in the person u_i. The calculation of Interest Score $S_{int}(t \to i)$ is based on the frequency of appearance of a feature word of the person u_i in the documents including the interests of the target user u_t. However, it appears that the concept in this method are not enough to extract the feature words of each of the persons other than the subject users, exactly.

For example, Table 3.20 shows the feature words of each of two persons among the persons other than the subject users. One is the account of a college student, the other is the account of a shop. Here, the feature words of each person (account) are extracted from the posts written by the person. In the case of the shop, I confirmed that there are the words showing the features of the shop, in the posts written by the account. On the other hand, in the case of the college student, it appears that there are not the words showing the features of the person, but the words showing the preferences of the person, in the posts written by the account. Therefore, there is a tendency that ordinary persons seems to remark their favorite things, whereas the accounts of organizations remark information of themselves. I plan to consider a system that selects the suitable documents for extracting the feature words of the person depending on the types of accounts.

Table 3.20: The feature words of each of two users

Account 1 (a college student)		**Account 2 (a coffee shop)**	
TF-IDF	Word (Japanese)	TF-IDF	Word (Japanese)
0.104	デレステ	0.710	スネークセンター
0.046	ヒナ	0.247	ヘビ
0.034	幸子	0.110	松寿
0.033	もも	0.100	ホオズキ
0.027	スカチケ	0.099	コーンスネーク
0.027	未央	0.084	アメリ
0.024	フェデラー	0.082	ニジコ
0.021	錦織	0.081	ニジスケ
0.019	ダルビッシュ	0.070	蛇
0.019	キューバ	0.066	パンサー

3.10 Conclusion of This Chapter

This chapter has proposed the methods for estimating the influencers for each user by analyzing the logs of his/her behaviors in SNS that show the user is influenced by other persons . In addition, the experiments have evaluated how exactly the proposed methods are able to estimate the actual influencers selected by each of the seven subject users, using Recall, Precision, and its F-measure.

3.10.1 Novelty

A number of studies on Influencer have proposed several methods for specifying a person who has an influence on society. On the other hand, this chapter has proposed the methods for estimating not only a common influencer for all users but also the difference influencers for each user.

One of the reasons why I have aimed to estimate the personalized influencers, is because I plan to propose the recommender system combined with Social Networking Service. First, the assumed recommender system (Figure 1.1) estimates a post that a user would be interested in. Second, it advertises several contents that are related to the topics of the post, while showing the post to the user. I described that the person who write a post that a user may be interested in, is not always a celebrity. Therefore, even if a person has an influence on only a user and does not have an influence on society, a remark of the person would be important as one of the persuasive grounds for the system to advertise something to user. That is why this chapter tries to discover the personalized influencers.

3.10.2 Main Results

Main Results on Influencer Estimation Based on User's Reactions

The first proposal of this chapter is the method for estimating the influencers for a target user by analyzing the target user's reactions to other persons in SNS. Specifically, the method analyzes the ranking of the persons based on their numbers of replies from the target user and the ranking of the persons based on their numbers of bookmarks from the target user.

The evaluation experiments have suggested that it is difficult for the the methods whose conditions for a person to become an influencer for a user are too stick, to estimate more influencers for the user. One of them is the method regarding the persons whose posts have been replied by a target user frequently, as the influencers of the target user. The other is the method regarding the persons whose posts have been replied and bookmarked by a target user frequently, as the influencers for the target user. I have confirmed each of these methods does not have a high F-measure.

On the other hand, it has been revealed that the best method for estimating the influencers for a target user has two way of thinkings. One is to regard the persons whose posts have been replied or bookmarked by a target user frequently, as the influencers for the target user. The other is to consider the weight that shows which is important, the persons whose posts have replied from the target user or the persons whose posts have bookmarked from the target user, as the influencers for the target user.

Main Results on Influencer Estimation Based on User's Interests

The second proposal of this chapter is the method for estimating the influencers for a target user by analyzing the target user's interests for other persons in SNS. Specifically, the method analyzes the frequency of appearance of the feature word of a person's, in the documents including the target user's interests.

The evaluation experiments have suggested that the proposed methods are superior to the method that is closed to previous studies. The method that is closed to previous studies analyzes the number of followers of each of the persons followed by a target user. On the other hand, the proposed methods analyze not only the interests of surroundings for the person, but also the interest of each target user for the person.

In addition, I have conducted the experiment for discovering the types of documents that are suitable for the analysis in this proposed method. The documents that are suitable for extracting the feature word of a person, are the comments in the profile of the person's SNS account and the posts written by the person. The documents that are suitable for analyzing the interests of a target user, are the posts written by the target user and the posts that have been bookmarked by the target user.

Influencer Estimation Based on Both Reactions and Interests

I also have evaluated the method regarding both the persons estimated by the method based on analyzing a user's reactions to other persons and the persons estimated the method based on analyzing the user's interests for other persons, as the influencers for him/her. I have confirmed there is a tendency that the influencers who are not able to be estimated by one side of the 2 methods, can be estimated by the other method. Therefore, it has been suggested that it is effective to analyze both a target user's reactions to other persons and his/her interests for other persons in SNS, to estimate the influencers for the target user.

3.10.3 Future Works

First, I need to study on a system for determining the optimal value of each of the parameters that are used in the methods proposed in this chapter. This chapter has shown that these parameters change depending on the type of target users I plan to analyze the trends and the difference of the optimal value of the parameter depending on the types of target users, by expanding the scale of the experiment.

Second, I need to analyze the logs of the behaviors in SNS of a target user more deeply. The proposed methods based on a user's reactions to other persons analyze the number of replies/bookmarks from the user to other persons For an instance, there are several meanings in a user's reply to a post of a person. One means one of the silly conversations with the user and the person. Another means that the user is impressed by the person. I plan to propose a method for calculating the degree of the reaction from the user to the person by analyzing also the text in the reply and recognizing the mean of the reply. It would enable to estimate the influencers for the user more exactly. Additionally, the proposed methods based on analyzing a user's interests for other persons analyze only the frequency of appearance of a word in the posts that have been bookmarked/written by the user. It is expected that focusing on the images and hash tags in these posts, would help to recognize what/who the user is interested in more exactly.

Third, the influencers for a user may change depending on the user's context (e.g., time and place). One of the future works analyzes the difference of the estimated influencers for the user depending on the date when analyzing social data on the user.

Chapter 4

Bookmarking Forecast for Post

4.1 Introduction of This Chapter

Chapter 1 has discussed the recommender system combined with SNS (Figure 1.1) as a vision of the future. In addition, I described that one of the component techniques of the assumed recommender system is to estimate the posts that each user would be interested in (Figure 1.1–Step1). Furthermore, I claimed that forecasting a user's various behaviors after he/she looks at a post, helps to estimate whether the user would be interested in the post or not. For example, one of the behaviors that we want to forecast, is a user's bookmark for a post. There are various meanings in a user's bookmark for a post (Table 4.1), and one of them is that the user has an interest for the topic/image of the post. Therefore, forecasting the posts that a user would bookmark seems to become a part of estimating the posts that a user would be interested in.

Table 4.1: Various meanings in a user's bookmark for a post

	Meanings
1	The user's interests for the topics/images of the post
2	The user's favor/respect for the person who wrote the post
3	A simple reply from the user (OK/Thank you)
4	A tag for referring the information/knowledge in the post later

Hence, this chapter aims to propose a method for forecasting each user's bookmarks for various posts. Specifically, I consider 3 kinds of affectors when a user bookmarks a post, 1) the goodness of the content and the format of the post, 2) the influence of the person who wrote the post, and 3) the social assurance for the post. Next, I create a model that forecasts whether the user would bookmark the post or not by using Random Forest that has 22 kinds of independent variables, and evaluate its forecast performance.

4.2 Independent Variables for Bookmaking Forecast

This section examines several independent variables that work for forecasting the posts that each user would bookmark. I focus on three types of independent variables shown in Table 4.2.

Table 4.2: The three types of affectors for a user to bookmark a post

The goodness of the content/format of the post	**Class**
The goodness of the constitution of the post	A-1
The user's interest for the topic of the post	A-2
The influence of the person who wrote the post	**Class**
The user's interest for the person who wrote the post	B-1
The user's reaction to the person who wrote the post	B-2
The attention for the person who wrote the post from people	B-3
The social assurance for the post	**Class**
The people's reactions to the post	C

4.2.1 Goodness of Content/Format of Post

One of the conditions for a subject user u_t to bookmark a post p, is that the content and the format of the post p is good. That is divided into two types (Table 4.2 A-1 and A-2). One (A-1) is the goodness of the constitution of the post p. The other (A-2) is the degree of interest of the subject user u_t for the topic of the post p. They would help for a system to forecast whether the subject user u_t would bookmark the post p or not.

The independent variables in A-2 are calculated by analyzing either the posts that have been bookmarked by the subject user u_t or the posts that have been written by the subject user u_t or the posts that have been written by the persons followed by the subject user u_t. Let one of the three types of documents, be D_t. Then a function $\mathrm{Int}(t \rightarrow p)$ that shows the degree of interest of the subject user u_t for the topic of the post p, is calculated as the maximum value among the frequencies of appearance of each of the all words W_p of the post p in the documents D_s. Here, $tf_t(w)$ in this formula is the frequency of appearance of a word w in the documents D_s.

$$Int(t \rightarrow p) = \max\{tf_t(w)|w \in W_p\}$$

4.2.2 Influence of Person Who Wrote Post

One of the conditions for a subject user u_t to bookmark a post p, is that the person u_i who wrote the post p has an influence on the subject user u_t and society. That is divided into three types (Table 4.2 B-1, B-2, and B-3). One (B-1) is the degree of interest of the subject user u_t for the person u_i who wrote the post p. Another (B-2) is the degree of reaction of the subject user u_t to the person u_i who wrote the post p. The other (B-3) is the degree of attention for the person u_i who wrote the post p from people (society). They would also help to forecast whether the subject user u_t would bookmark the post p or not.

The independent variables in B-1 are calculated by analyzing either the posts that have been bookmarked by the subject user u_t or the posts that have been written by the subject user u_t. Let one of the two types of documents, be D_t. Then a function Int($t \to i$) that shows the degree of interest of the subject user u_t for the person u_i is calculated by the following formula.

$$Int(t \to i) = \sum_{w \in W_i} tfidf_i(w) \times tf_t(w)$$

W_i in this formula shows the all words in the comments in the profile of SNS account of the person u_i. In addition, $tfidf_i(w)$ shows how much a word w expresses the person u_i, and is based on TFIDF Analysis. TF of a word w is calculated based on the frequency of appearance of the word w in the comments in the profile of SNS account of the person u_i. DF of a word w is calculated based on the number of users whose comments in the profile of the SNS account include the word w, among 100 users (the subject users in this chapter).

4.2.3 Social Assurance for Post

One of the conditions for a subject user u_t to bookmark a post p, is that there is a social assurance for the post p. I define the 4 independent variables based on people's reactions to the post p (Table 4.2 C).

Table 4.3: The independent variables for Bookmarking Forecast

No	Class	Name	Range
1	A-1	Txt*	{0,1}
2	A-1	TxtRep*	{0,1}
3	A-1	Img*	{0,1}
4	A-1	ImgRep*	{0,1}
5	A-1	TxtImg*	{0,1}
6	A-1	TxtImgRep*	{0,1}
7	A-2	IntFav	≥ 0
8	A-2	IntTwt	≥ 0
9	A-2	IntFol	≥ 0
10	B-1	IntPerFav	≥ 0
11	B-1	IntPerTwt	≥ 0
12	B-2	FavRate	[0,1]
13	B-2	RepRate	[0,1]
14	B-2	RtwRate	[0,1]
15	B-2	RxnRate	[0,1]
16	B-3	Flw	≥ 0
17	B-3	UsFav	≥ 0
18	B-3	UsRtw	≥ 0
19	C	TarFav	≥ 0
20	C	ChgFav	≥ -1
21	C	TarRtw	≥ 0
22	C	ChgRtw	≥ -1

Table 4.4: The explanations of the independent variables in Class A

Name	Explanation
Txt*	Whether a post p consists of only texts or not
TxtRep*	Whether a post p consists of only texts and is a reply to someone or not
Img*	Whether a post p consists of only images or not
ImgRep*	Whether a post p consists of only images and is a reply to someone or not
TxtImg*	Whether a post p consists of both texts and images or not
TxtImgRep*	Whether a post p consists of both texts and images and is a reply to someone or not
IntFav	The degree of interest of a subject user u_t for the topic of a post p based on analyzing the posts that have been bookmarked by the subject user u_t
IntTwt	The degree of interest of a subject user u_t for the topic of a post p based on analyzing the posts that have been written by the subject user u_t
IntFol	The degree of interest of a subject user u_t for the topic of a post p based on analyzing the posts that have been written by the persons followed by the subject user u_t

Table 4.5: The explanations of the independent variables in Class B

Name	Explanation
IntPerFav	The degree of interest of a subject user u_t for the person u_i who wrote a post p based on analyzing the posts that have been bookmarked by the subject user u_t
IntPerTwt	The degree of interest of a subject user u_t for the person u_i who wrote a post p based on analyzing the posts that have been written by the subject user u_t
FavRate	The rate of the posts written by the person u_i who wrote a post p in the all posts bookmarked by a subject user u_t
RepRate	The rate of the posts written by the person u_i who wrote a post p in the all posts replied by a subject user u_t
RtwRate	The rate of the posts written by the person u_i who wrote a post p in the all posts shared by a subject user u_t
RxnRate	$\max\{\mathbf{FavRate}, \mathbf{RepRate}, \mathbf{RtwRate}\}$
Flw	The number of followers of the person u_i who wrote a post p
UsFav	The mean number of bookmarks from surroundings for the posts written by the person u_i who wrote a post p
UsRtw	The mean number of shares from surroundings for the posts written by the person u_i who wrote a post p

Table 4.6: The explanations of the independent variables in Class C

Name	Explanation
TarFav	The number of bookmarks from surroundings for a post p
ChgFav	The rate of change from the mean number of bookmarks from surroundings for the posts written by the person u_i who wrote a post p to the number of bookmarks from surroundings for the post p $(\mathbf{TarFav} - \mathbf{UsFav})/\mathbf{UsFav}$
TarRtw	The number of shares from surroundings for a post p
ChgRtw	The rate of change from the mean number of shares from surroundings for the posts written by the person u_i who wrote a post p to the number of bookmarks from surroundings for the post p $(\mathbf{TarRtw} - \mathbf{UsRtw})/\mathbf{UsRtw}$

4.3 Experimental Method

4.3.1 Procedures

The procedures in this experiment are as follows.

1. Extracting the posts that a subject user has looked at.
2. Appending a label (correct data) that shows whether the user bookmarked the post or not, to each extracted post, by referring the bookmark-list of the user.
3. Calculating the independent variables for each post, and forecasting whether each post is bookmarked by the user or not, by using Random Forest.
4. Evaluating the forecast model by using criteria shown in 4.3.3.

This experiment implements Random Forest based on caret package in R language. The number of dimension of the feature vector of Decision Tree is optimized based on Grid Search. In addition, I conduct 10-fold cross validation when evaluating the forecast model of the subject user.

4.3.2 Dataset

This experiment has 100 Twitter users as subject users. Their usage situations of Twitter and the numbers of persons who are related to the 100 subject users are shown in Table 4.7 and Table 4.8 respectively. In addition, their Twitter data are extracted by using Twitter API, and Table 4.9 summarizes their details. However, Twitter API does not provide the posts that each subject user has looked at. Therefore, I get several posts among the posts written by the persons followed by the subject user (Table 4.8 $*^1$) and the posts written by the persons whose posts have been bookmarked by the subject user at least once (Table 4.8 $*^2$), and regard them as the posts that the subject user has looked at. Table 4.10 shows the number of test data and the number of training data. FAV in Table 4.10 means the post(s) that a subject user bookmarks, and N-FAV in Table 4.10 means the post(s) that he/she does not.

Table 4.7: The numbers of replies/bookmarks of the subject users

	Mean	SD
# Posting / Day	7.548	10.013
# Bookmarking / Day	5.178	3.766
# Replying / # Posting	0.289	0.245

Table 4.8: The numbers of persons who are related to the subject users

	Mean	SD
Following Persons $*^1$	58.700	44.269
Bookmarked persons $*^2$	145.310	72.816

Table 4.9: The numbers of extracted data

	Mean	SD
# Post (Sub.)	358.040	88.672
# Bookmarked Post (Sub.)	344.390	78.925
# Post (Table 4.8 $*^1$)	190.858	33.525
# Post (Table 4.8 $*^2$)	193.074	28.408

Table 4.10: The numbers of training data and test data

	Training Data		Test Data	
	FAV	N-FAV	FAV	N-FAV
Mean	310.951	311.639	35.439	115.306
SD	70.713	71.148	7.867	26.652

4.3.3 Criteria

This experiment evaluates the performance of the forecast model of each of the 100 subject users by using the confusion matrix shown in Table 4.11, Accuracy, Recall, Precision, and Specificity. These criteria are calculated by the following formulas.

$$\text{Accuracy} = (\text{TP} + \text{TN})/(\text{TP} + \text{TN} + \text{FP} + \text{FN})$$

$$\text{Recall} = \text{TP}/(\text{TP} + \text{FN})$$

$$\text{Precision} = \text{TP}/(\text{TP} + \text{FP})$$

$$\text{Specificity} = \text{TN}/(\text{TN} + \text{FP})$$

In addition, I evaluate F-measure of the forecast model that shows a harmonic mean between Recall and Precision, and Balanced-Accuracy that considers the ratio of the number of FAV to the number of N-FAV in test data. They are calculated by the following formulas.

$$\text{F-measure} = 2 \times \text{Recall} \times \text{Precision}/(\text{Recall} + \text{Precision})$$

$$\text{Balanced-Accuracy} = (\text{Recall} + \text{Specificity})/2$$

Table 4.11: Confusion Matrix

Fcst \ Act	FAV	N-FAV
FAV	**TP**	**FP**
N-FAV	**FN**	**TN**

4.4 Result

4.4.1 Performance of Forecast Model

This subsection discusses the performance of the model that forecasts the posts that each subject user would bookmark. Table 4.12 summarizes the mean performance of the forecast models of the 100 subject users. We can confirm a tendency that Specificity (i.e., True negative rate) is higher than the other criteria. This result seems to match with our intuition. The reason is because the posts that a subject user actually bookmarks account for only 30% of the test data of this experiment (Table 4.10). In other words, it is easier for the proposed model to detect the posts that the user would not bookmark (N-FAV) than to detect the posts that he/she does (FAV). That is why the forecast model has a high detection rate of N-FAV (Specificity). Therefore, it is also a natural tendency that Balanced-Accuracy (i.e., the mean of Specificity and Recall) is higher than F-measure (the mean of Precision and Recall).

Table 4.12: The performance of the forecast models of the subject users

	TP	FN	FP	TN
Mean	24.887	9.552	11.458	103.848
SD	8.588	5.928	5.294	24.885
Max	40.000	31.000	37.000	134.000
Min	2.000	0.000	0.000	18.000
Median	26.000	9.000	11.000	113.000

	Acc	Spe	Pre	Rec	F	B-Acc
Mean	0.856	0.899	0.681	0.715	0.690	0.807
SD	0.053	0.041	0.110	0.169	0.128	0.089
Max	0.994	1.000	1.000	1.000	0.987	0.996
Min	0.681	0.738	0.250	0.143	0.214	0.509
Median	0.857	0.903	0.679	0.744	0.701	0.814

On the other hand, Recall and Precision are more important to evaluate the proposed model because these criteria focus on how exactly the model can forecast the posts that each subject user would bookmark (i.e., how exactly it can detect FAV). Table 4.12 reveals both Recall and Precision are approximately 0.7. In other words, the probability that the proposed model cannot forecast a post that a subject user may bookmark and the probability that the proposed model's forecast that a subject user would bookmark a post is the mistake, are both only 30%. The reason why the detections of FAV by proposed model are incomplete (i.e., a cause of False Negative), may be because the investigations about the conditions for a user to bookmark a post (Section 4.2) are not enough. In addition, the reason why the detections of FAV by proposed model include some faults (i.e., a cause of False Positive), may be because the method for calculating each independent variable is not perfect. I plan to examine more conditions for a user to bookmark a post and several suitable methods for calculating the independent variables more exactly.

However, there are also many cases that a user bookmarks several posts on a whim (i.e., without regularity). Therefore, it is not easy to forecast whether a user would bookmark a post without an omission (False Negative) and a fault (False Positive). Hence, I would like to claim that this result is not bad and the forecast performance of the proposed model seems to be a practical level.

4.4.2 Effect of User's Type on Forecast Performance

This subsection discusses the effects of the difference between the usage stats of Twitter of each subject user on the performance of his/her forecast model. Table 4.13 is a correlation matrix whose columns show several types of the subject users and whose rows are the criteria. First, focusing on **FolU** (3rd column), we can confirm that there is a weak negative correlation between **FolU** and Recall. It is also revealed that the decline of Recall causes the declines of both F-measure and Balanced-Accuracy. A user who has a large number of friends, may overlook several posts and whimsy bookmark a post whose type is differ from the posts that this user usually bookmarks, because he/she looks at many posts written by his/her friends. Therefore, it seems to be hard to forecast the posts that a user who has a large number of friends (i.e., a large number of persons followed by the user) would bookmarks.

Table 4.13: The correlations between the types of user and the criteria

	Fav/D	**Twt/D**	**FolU**	**FlwU**	**FavU**
Acc	0.06	0.02	−0.20	0.08	0.08
Spe	0.11	0.04	0.13	0.08	0.41
Pre	0.09	0.03	−0.08	0.10	0.21
Rec	−0.01	−0.01	−0.39	0.04	−0.24
F	0.04	0.01	−0.29	0.07	−0.07
B-Acc	0.02	0.00	−0.33	0.06	−0.12

Fav/D: The number of posts that a subject user bookmarks per day
Twt/D: The number of posts that a subject user writes per day
FolU: The number of persons whom a subject user follows
FlwU: The number of followers of a subject user
FavU: The number of persons whose posts have been bookmarked by a subject user at least once

Second, it is revealed that there is a weak negative correlation between **FavU** and Recall, and a weak positive correlation between **FavU** and Precision (or Specificity). However there is no correlation between **FavU** and F-measure (or Balanced-Accuracy). Therefore, though the number of persons whose posts have been bookmarked by a subject at least once has an effect on individual criterion, it does not have an effect on its final performance. That means there is not a big difference between the performance of the forecast model of a subject user who bookmarks many posts written by various persons and the performance of the forecast model of a subject user who does not.

4.4.3 Importance of Independent Variable

This subsection analyzes the importances of the independent variables of the forecast model of each of the 100 subject users.

Importances of Independent Variables of Average Forecast Model

Table 4.14 shows the importance of each of the independent variables of the average forecast model of the 100 subject users. It is revealed that ChgFav is the most important in the independent variables when forecasting the posts that each of the 100 subject users would bookmark. In addition, we can confirm that the mean importance of ChgFav of the forecast models of the 100 subject users, are 88.35 ($\pm$22.99). I described ChgFav shows the people's reactions to a post. Therefore, the result that ChgFav is the important independent variable for forecasting whether a user bookmarks a post or not, suggests an average user is easy to bookmark the posts that receives many bookmarks form surroundings.

Besides, the values of IntFav and FavRate are higher than the values of the other independent variables, in the average forecast model of the 100 subject users. IntFav shows the degree of interest of a subject user for the topic of a post. Specifically, it is calculated based on the frequency of appearance of a word in this post, in the posts bookmarked by this subject user. In addition, FavRate shows the degree of reaction of a subject user to the person who wrote a post. Specifically, it is calculated based on the number of bookmarks for the posts written by this person that this subject user has done. Therefore, the result that IntFav and FavRate are the effective independent variables for forecasting whether a user would bookmark a post or not, suggests an average user is easy to be conscious of the posts whose contents are similar to the contents of the posts that the user has bookmarked, and the posts written by the persons whose posts have been bookmarked by the user frequently.

Table 4.14: The importance of each independent variable

	Maen	SD	Max	Min	Median
Txt	17.71	24.30	100.00	0.03	7.26
TxtRep	20.72	29.44	100.00	0.00	6.45
Img	0.55	0.77	8.62	0.00	0.34
ImgRep	0.28	1.16	12.21	0.00	0.00
TxtImg	6.42	8.87	57.14	0.05	2.85
TxtImgRep	0.54	2.29	29.02	0.00	0.00
IntFav	30.79	25.72	100.00	0.09	23.45
IntTwt	8.22	9.92	100.00	0.00	5.22
IntFol	15.09	10.59	71.24	0.06	12.83
IntPerFav	6.11	5.52	54.15	0.35	4.59
IntPerTwt	5.21	4.57	30.70	0.00	3.91
FavRate	35.21	27.20	100.00	0.79	26.32
RepRate	7.32	11.84	100.00	0.00	3.70
RtwRate	3.92	10.08	100.00	0.00	1.20
RxnRate	21.38	18.15	100.00	0.93	16.30
Flw	9.91	7.56	74.27	0.93	7.99
UsFav	8.87	7.37	63.31	0.50	6.49
UsRtw	8.27	7.29	65.82	0.31	5.95
TarFav	38.86	28.66	100.00	0.43	32.54
ChgFav	88.35	22.99	100.00	2.34	100.00
TarRtw	13.72	12.61	65.45	0.15	9.17
ChgRtw	26.48	17.92	93.07	0.49	21.48

Cluster Analysis on Subject Users

I investigate the types of subject users based on the importances of the 22 independent variables of the forecast model of each of the 100 subject users. First, I conduct a hierarchical cluster analysis for the 100 subject users, based on the 22 dimensional feature vector of the forecast model of each of them. Here, the measurement of the distance between each feature vector is based on Euclidean Distance.

The analysis classifies the 100 subject users into three clusters. Here, the number of clusters is optimized based on Silhouette Analysis. Figure 4.1 shows the mean feature vector of each type. Its horizontal axis is an independent variable, and its vertical axis is the importance of the independent variable. The details of these clusters are as follows.

First, we can confirm that the mean feature vector of the forecast models of the subject users in Cluster 1 has a high value of ChgFav. In other words, whether each of the subject users in Cluster 1 would bookmark a post in SNS or not is easy to change depending on the social assurance for the post. Conversely, it is hard to be affected by the influence of the person who wrote the post and the goodness of the content and the format of the post. That means the subject users in Cluster 1 may be easy to have an interest for a topic that many ordinary people are interested in. Therefore, this analysis suggests that the subject users in Cluster 1 are Society Inspired Type Person. In addition, I revealed they account for about 65% of the all subject users.

Second, we can confirm that the mean feature vector of the forecast models of the subject users in Cluster 2 has the high values of ChgFav, Txt, TextRep, and IntFav. In other words, whether each of the subject users in Cluster 2 would bookmark a post in SNS or not is easy to change depending on the goodness of the contents and the format of the post. Conversely, it is hard to be affected by the social assurance for the post and the influence of the person who wrote the post. That means the subject users in Cluster 2 may be easy to have an interest for a topic that is close to their preferences. Therefore, this analysis suggests that the subject users in Cluster 2 are Content Inspired Type Person. In addition, I revealed they account for about 22% of the all subject users.

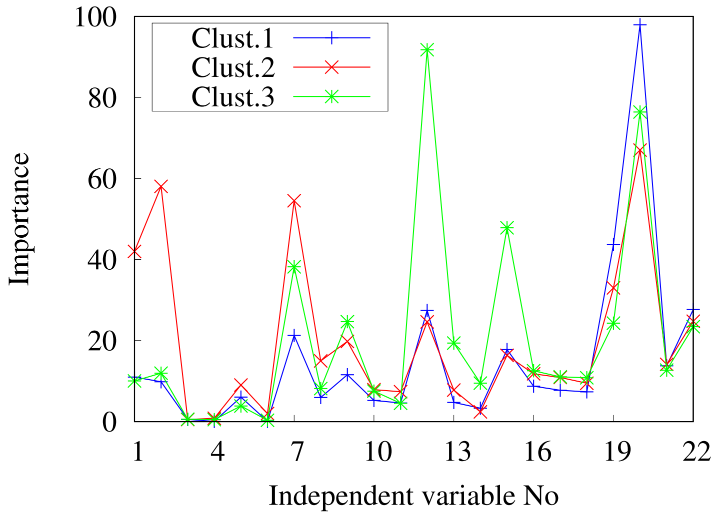


Figure 4.1: The feature vector of the forecast model of each type of users

Third, we can confirm that the mean feature vector of the forecast models of the subject users in Cluster 3 has the high values of ChgFav, FavRate, and RxnRate. In other words, whether each of the subject users in Cluster 3 would bookmark a post in SNS or not is easy to change depending on the influence of the person who wrote the post. Conversely, it is hard to be affected by the social assurance for the post and the goodness of the contents and the format of the post. That means the subject users in Cluster 3 may be easy to have an interest for the topic in a remark of an influencer for the subject users. Therefore, this analysis suggests that the subject users in Cluster 3 are Talker Inspired Type Person. In addition, I revealed they account for about 22% of the all subject users.

4.5 Conclusion of This Chapter

This chapter has discusses a method that forecasts the posts that each user would bookmark by analyzing the logs of his/her behaviors in SNS. Specifically, I have considered the 3 kinds of affectors when a user bookmarks a post, 1) the goodness of the contents and the format of the post, 2) the influence of the person who wrote the post, and 3) the social assurance for the post. In addition, I have proposed a model forecasting whether each user would bookmark a post or not by using Random Forest that has the 22 kinds of independent variables, and evaluated its forecast performance and the importance of each independent variable.

4.5.1 Novelty

There are no previous studies that discuss the same task with this chapter (i.e., forecasting a user's bookmark for a post). Suffice it to say, the research field that is similar to the task of this chapter is the recommendation of a Web content (e.g., a Web page) that a user would like. A web page that a target user would be interested in, is specified by discovering the Web pages that have been viewed by the persons who have the same preferences with the target user and the Web pages whose contents include the preferences of the target user.

However, that is not enough to specify a post in SNS that the target user would be interested in. For example, estimating how much a target user is influenced by a person, would help to judge whether the target user would have an interest for a post written by the person or not. This chapter tries to forecast the posts that a target user would be interested in, by utilizing several unique technologies discussed in the previous chapter that estimates the degree of influence of the person who wrote a post to the target user. Therefore, it was able to discuss the task that has not been studied in the previous papers.

4.5.2 Main Resluts

I have proposed the the method for forecasting whether each user would bookmark a post in SNS or not. In addition, it is assumed the forecast performance of the method is a practical level, because the experiment reveals that F-measure of the mean forecast model of the 100 subject users is 0.690, its Balanced-Accuracy is 0.807.

Moreover, I have demonstrated the following. First, the forecast models of most of the 100 subject users, tend to regard the social assurance for a post as the important affector when forecasting their bookmarks. Next, the forecast models of 22% of the 100 subject users, tend to regard the goodness of the contents and the format of a post as the important affector when forecasting their bookmarks. Finally, the forecast models of 13% of the 100 subject users, tend to regard the influence of the person who wrote a post as the important affector when forecasting their bookmarks.

Chapter 5

Summary

Discovering a method for changing people's interests is an important task. For example, it helps for us to promote various items, to improve students' motivation to learn, and to care dependent patients. I have advocated a new framework for inducing a user's interests for various things, which recommends various contents to the user while showing several posts that have useful information for the user. This paper discusses two phases that are the bases of this framework.

The first phase conducts a preliminary experiment for discussing an expected effect of the assumed framework by using its prototype. Specifically, it shows that a recommender system that advertises items while showing their reasons extracted from SNS, has the effects on the improvement of a user ' s receptivity for a new thing and the induction of his/her interest for it. This result helps for me to insist that the assumed framework has the social significances.

The second phase develops several component technologies of the assumed framework. Especially, it is one of the important technologies of the assumed framework, to estimate several posts that each user would be interested in (e.g., a post written by a person whom the user likes). However, existing methods in Social Computing cannot realize it. This paper aims to propose a new method that analyzes individual social data and estimates them. Its details are as follows.

1. This paper proposes a method that estimates the influencers for each user by analyzing the user's reactions to other persons and the user's interests for other persons in SNS. In addition, it confirms that the proposed method is superior to those of previous researches, because the proposed method can estimate not only a common influencer for all users (Social influencer) but also the different influencers for each user (Personalized influencers).

2. This paper proposes a method that formulates several relationships between each user and a post, and forecasts whether each user would bookmark the post or not. In addition, it examines the differences of the factors for the users to bookmark the post depending on the type of the user by a cluster analysis for them.

Summarizing the above, this paper is for an advanced recommender system combined with SNS. It proposes the method that estimates the influencers for each user and the method that forecasts the posts that each user would bookmark.

Acknowledgements

I always received insightful comments, useful suggestions, and warm encouragements from Professor Yasuo Kudo, Assistant Professor Shun Hattori, and the members of Web Intelligence Time-Space (WITS) Laboratory. I would like to express the deepest appreciation to them. In addition, I want to thank my family, my relatives, and my friends (Wan, Mic., Mas., Mid., and Aka etc.) for their moral support.

Bibliography

[1] B. M. Sarwar, G. Karypis, and J. A. Konstan, “Item-Based Collaborative Filtering Recommendation Algorithms,” Proceedings of the 10th International Conference on World Wide Web, pp.285–295 (2001).

[2] G. Adomavicius and A. Tuzhilin, “Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions,” IEEE Transactions on Knowledge and Data Engineering, Vol.17, Issue 6, pp.734–749 (2005).

[3] D. Bogdanov, G. G. Haro, F. Fuhrmann, E. Gómez, and P. Herrera, “Content-Based Busic Recommendation Based on User Preference Examples,” The 4th ACM Conference on Recommender Systems, Workshop on Music Recommendation and Discovery pp.33–38 (2010).

[4] M. Levy and M. Sandler, “Music Information Retrieval Using Social Tags and Audio,” IEEE Transactions on Multimedia Vol.11, Issue 3, pp.383–395 (2009).

[5] Y. Koren, R. Bell, and C. Volinsky, “Matrix Factorization Techniques for Recommender Systems,” IEEE, Computer Journal, Vol.42, Issue 8, pp.30–37 (2009).

[6] M. Ge, F. Ricci, and D. Massimo, “Health-Aware Food Recommender System,” Proceedings of the 9th ACM Conference on Recommender Systems, pp.333–334 (2015).

[7] N. Shino, R. Yamanishi, and J. Fukumoto, “Recommendation System for Alternative-Ingredients Based on Co-occurrence Relation on Recipe Database and the Ingredient Category,” Proceedings of the 5th IIAI International Congress on Advanced Applied Informatics, pp.986–993 (2016).

[8] D. Bianchini, V. D. Antonellis, N. D. Franceschi, and M. Melchiori, "PREFer: A Prescription-Based Food Recommender System," Computer Standards and Interfaces, Vol.54, Part 2, pp.64–75 (2017).

[9] G. Agapito, M. Simeoni, B. Calabrese, I. Caré, T. Lamprinoudi, P. H. Guzzi, A. Pujia, G. Fuiano, and M. Cannataro, "DIETOS: A Dietary Recommender System for Chronic Diseases Monitoring and Management," Computer Methods and Programs in Biomedicine, Vol.153, pp.93-104 (2018).

[10] Y. Yang and N. C. Marques, "User Group Profile Modeling Based on User Transactional Data for Personalized Systems," Proceedings of the 12th Portuguese Conference on Progress in Artificial Intelligence, pp.337–347 (2005).

[11] I. Seifert, "Collaborative Assistance with Spatio-Temporal Planning Problems," International Conference on Spatial Cognition, Spatial Cognition 2006: Spatial Cognition V Reasoning, Action, Interaction, pp.90–106 (2006).

[12] L. Ardissono, A. Goy, G. Petrone, M. Segnan, and P. Torasso, "Intrigue: Personalized Recommendation of Tourist Attractions for Desktop and Hand Held Devices," Applied Artificial Intelligence, An International Journal, Vol.17, Issue 8–9, pp.687–714 (2003).

[13] X. Yi, J Allan, and W. B. Croft, "Matching Resumes and Jobs Based on Relevance Models," Proceedings of the 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.809–810 (2007)

[14] V. Lavrenko and W. B. Croft, "Relevance-Based Language Models," Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp.120–127 (2011).

[15] A. Singh, C. Rose, K. Visweswariah, V. Chenthamarakshan, and N. Kambhatla, "PROSPECT: A System for Screening Candidates for Recruitment," Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp.659–668 (2010).

[16] X. Zang, T. Yamasaki, K. Aizawa, T. Nakamoto, E. Kuwabara, S. Egami, and Y. Fuchida, "Analysis of People's Attractiveness in Dating Activities by Machine Learning," The Institute of Image Informa-

tion and Television Engineers, Winter Annual Convention 2016, 13C-5 (2016).

[17] R. Rothe, R. Timofte, and L. V. Gool, "Some Like It Hot-Visual Guidance for Preference Prediction," Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition 2016, pp.5553–5561 (2016).

[18] J. Hannon, M. Bennett, and B. Smyth, "Recommending Twitter Users to Follow using Content and Collaborative Filtering Approaches," Proceedings of the 4th ACM Conference on Recommender Systems, pp.199–206 (2010).

[19] L. Backstrom and J. Leskovec, "Supervised Random Walks: Predicting and Recommending Links in Social Networks," Proceedings of the 4th ACM International Conference on Web Search and Data Mining, pp.635–644 (2011).

[20] P. Domingos and M. Richardson, "Mining the Network Value of Customers," Proceedings of the 7th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.57–66 (2001).

[21] M. Richardson and P. Domingos, "Mining Knowledge-Sharing Sites for Viral Marketing," Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.61–70 (2002).

[22] C. L. Hsu and J. C. C. Lin, "Acceptance of Blog Usage: The Roles of Technology Acceptance, Social Influence and Knowledge Sharing Motivation," Information and Management, Vol.45, Issue 1, pp.65–74 (2008).

[23] S. Pookulangara and K. Koesler, "Cultural Influence on Consumers' Usage of Social Networks and Its' Impact on Online Purchase Intentions," Journal of Retailing and Consumer Services, Vol.18, Issue 4, pp.348–354 (2011).

[24] C. Lorenzo, E. Constantinides, and M. C. Alarcón-del-Amo, "Consumer Adoption of Social Networking Sites: Implications for Theory and Practice," Journal of Research in Interactive Marketing, Vol.5, Issue 2–3, pp.170–188 (2011).

[25] R. K. Merton, "Social Theory and Social Structure Revised and Enlarged Edition," The Free Press of Glencoe (1964).

[26] D. J. Watts and P. S. Dodds, "Influentials, Networks, and Public Opinion Formation," Journal of Consumer Research, Vol.34, pp.441–458 (2007).

[27] M. Cha, H. Haddadi, F. Benevenuto, and K. P. Gummadi, "Measuring User Influence in Twitter: The Million Follower Fallacy," Proceedings of International AAAI Conference on Weblogs and Social Media, pp.10–18 (2010).

[28] H. Kwak, C. Lee, H. Park, and S. Moon, "What is Twitter, a Social Network or a News Media?," Proceedings of the 19th International Conference on World Wide Web, pp.591–600 (2010).

[29] B. Wang, C. Wang, J. Bu, C. Chen, W. V. Zhang, D. Cai, and X. He, "Whom to Mention: Expand the Diffusion of Tweets by @ Recommendation on Micro-blogging Systems," Proceedings of the 22nd International Conference on World Wide Web, pp.1331–1340 (2013).

[30] J. Weng, E. P. Lim, J. Jiang, and Q. He, "Twitterrank: Finding Topic-Sensitive Influential Twitterers," Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pp.261–270 (2010).

[31] Z. Y. Ding, Y. Jia, B. Zhou, Y. Han, L. He, and J. F. Zhang, "Measuring the Spreadability of Users in Microblogs," Journal of Zhejiang University SCIENCE C, Vol.14, Issue 9, pp.701–710 (2013).

[32] M. Katagiri, M. Etoh, and H. Takemura, "Identifying Influencers through Mobile Applications Usage," Transactions of Information Processing Society of Japan, Vol.5, No.3, pp.75–85 (2012).

[33] E. Bakshy, J. M. Hofman, W. Mason, and D. J. Watts, "Everyone's an Influencer: Quantifying Influence on Twitter," Proceedings of the 4th ACM International Conference on Web Search and Data Mining, pp.65–74 (2011).

[34] D. Kempe, J. Kleinberg, and Éva Tardos, "Maximizing the Spread of Influence through a Social Network," Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.137–146 (2003).

[35] M. Kimura, K. Saito, and H. Motoda, "Blocking Links to Minimize Contamination Spread in a Social Network," ACM Transactions on Knowledge Discovery from Data, Vol.3, Issue 2, Article No.9 (2009).

[36] Y. Yoshikawa, K. Saito, H. Motoda, K. Ohara, and M. Kimura, "Estimating Method of Expected Influence Curve from Single Diffusion Sequence on Social Networks," The IEICE Transactions on Information and Systems, Vol.94, No.11, pp.1899–1908 (2011).

[37] P. H. Reingen and J. B. Kernan, “Analysis of Referral Networks in Marketing: Methods and Illustration,” Journal of Marketing Research, Vol.23, No.4, pp.370–378 (1986).

[38] S. W. Bither and P. Wright, “Preferences Between Product Consultants: Choices vs. Preference Functions,” Journal of Consumer Research, Vol.4, Issue 1, pp.39–47 (1977).

[39] F. Toriumi, T. Sakaki, M. Yoshida, K. Shinoda, S. Kurihara, K. Kazama, and I. Noda, “Finding Important Users for Information Diffusion,” The 28th Annual Conference of the Japanese Society for Artificial Intelligence, 1H2NFC02a1 (2014).

[40] L. Terveen, W. Hill, B. Amento, D. W. McDonald, and J. Creter, “PHOAKS: A System for Sharing Recommendations,” Communications of the ACM, Vol.40, No.3, pp.59–62 (1997).

[41] U. Shardanand and P. Maes, “Social Information Filtering: Algorithms for Automating ‘word of mouth’,” Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.210–217 (1995).

[42] P. Resnick, N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, “GroupLens: An Open Architecture for Collaborative Filtering of Netnews,” Proceedings of the 1994 ACM conference on Computer Supported Cooperative Work, pp.175–186 (1994).

[43] J. S. Breese, D. Heckerman, and C. Kadie, “Empirical Analysis of Predictive Algorithms for Collaborative Filtering,” Proceedings of the 14th Conference on Uncertainty in Artificial Intelligence, pp.43–52 (1998).

[44] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme, “Information Retrieval in Folksonomies: Search and Ranking,” European Semantic Web Conference, ESWC 2006: The Semantic Web: Research and Applications, pp.411–426 (2006).

[45] A. Hotho, R. Jaschke, C. Schmitz, and G. Stumme, “Trend Detection in Folksonomies,” International Conference on Semantic and Digital Media Technologies, SAMT 2006: Semantic Multimedia, pp.56–70 (2006).

[46] S. Niwa, T. Doi, and S. Honiden, “Web Page Recommender System based on Folksonomy Mining for ITNG’ 06 Submissions,” Proceedings of the 3rd International Conference on Information Technology: New Generations, pp.388–393 (2006).

[47] Z. K. Zhang, T. Zhou, and Y. C. Zhang, "Personalized Recommendation Via Integrated Diffusion on User-Item-Tag Tripartite Graphs," Physica A: Statistical Mechanics and its Applications, Vol.389, No.1, pp.179–186 (2010).

[48] J. Rucker, and M. J. Polanco, "Siteseer: Personalized Navigation for The Web," Communications of the ACM, Vol.40, No.3, pp.73–75 (1997).

[49] R. Wetzker, W. Umbrath, and A. Said, "A Hybrid Approach to Item Recommendation in Folksonomies," Proceedings of the WSDM'09 Workshop on Exploiting Semantic Annotations in Information Retrieval, pp.25–29 (2009).

[50] Y. Xu, L. Zhang, and W. Liu, "Cubic Analysis of Social Bookmarking for Personalized Recommendation," Asia-Pacific Web Conference, AP-Web 2006: Frontiers of WWW Research and Development, pp.733–738 (2006).

[51] D. Davidov, O. Tsur, and A. Rappoport, "Enhanced Sentiment Learning Using Twitter Hashtags and Smileys," 23rd International Conference on Computational Linguistics, pp.241–249 (2010).

[52] L. Barbosa and J. Feng, "Robust Sentiment Detection on Twitter from Biased and Noisy Data," 23rd International Conference on Computational Linguistics, pp.36–44 (2010).

[53] I. S. Silva, J. Gomide, A. Veloso, W. Meira, and R. Ferreira, "Effective Sentiment Stream Analysis with Self-Augmenting Training and Demand-Driven Projection," Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, pp.475–484 (2011).

[54] L. Jiang, M. Yu, M. Zhon, X. Liu, and T. Zhao, "Target-dependent Twitter Sentiment Classification," Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, pp.151–160 (2011).

[55] S. Brody and N. Diakopoulos, "Cooooooooooooooollllllllllllll!!!!!!!!!!!!!! Using Word Lengthening to Detect Sentiment in Microblogs," Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp.562–570 (2011).

[56] Z Yang, J. Guo, K. Cai, J. Tang, J. Li, L. Zhang, and Z. Su, "Understanding Retweeting Behaviors in Social Networks," Proceedings of

the 19th ACM International Conference on Information and Knowledge Management, pp.1633–1636 (2010).

[57] J. L. Iribarren and E. Moro, "Impact of Human Activity Patterns on the Dynamics of Information Diffusion," Physical Review Letters, Vol.103, No.3, pp.8–11 (2009).

[58] T. T. Kuo, S. C. Hung, W. S. Lin, S. D. Lin, T. C. Peng, and C. C. Shih, "Assessing the Quality of Diffusion Models Using Real-World Social Network Data," 2011 International Conference on Technologies and Applications of Artificial Intelligence, pp.200–205 (2011).

[59] Y. Matsubara, Y. Sakurai, B. A. Prakash, L. Li, and C. Faloutsos, "Rise and Fall Patterns of Information Diffusion: Model and Implications," Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.6–14 (2012).

[60] J. Kleinberg, "Bursty and Hierarchical Structure in Streams," Proceedings of the 8th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp.91–101 (2002).

[61] J. Bollen, H. Mao, and X. Zeng, "Twitter Mood Predicts the Stock Market," Journal of Computational Science, Vol.2, Issue 1, pp.1–8 (2011).

[62] S. Asur and B. A. Huberman, "Predicting the Future with Social Media," Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, Vol.1, pp.492–499 (2010).

[63] D. Quercia, M. Kosinski, D. Stillwell, and J. Crowcroft, "Our Twitter Profiles, Our Selves: Predicting Personality with Twitter," 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing, pp.180–185 (2011).

[64] J. Golbeck, C. Robles, M. Edmondson, and K. Turner, "Predicting Personality from Twitter," 2011 IEEE 3rd International Conference on Privacy, Security, Risk and Trust and 2011 IEEE 3rd International Conference on Social Computing, pp.149–156 (2011).

[65] J. Golbeck, C. Robles, and K. Turner, "Predicting Personality with Social Media," CHI'11 Extended Abstracts on Human Factors in Computing Systems, pp.253–262 (2011).

[66] D. Quercia, R. Lambiotte, D. Stillwell, M. Kosinski, and J.Crowcroft, "The Personality of Popular Facebook Users," Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work, pp.955–964 (2012).

[67] M. Kosinski, D. Stillwell, and T. Graepel, "Private Traits and Attributes are Predictable from Digital Records of Human Behavior," Proceedings of the National Academy of Sciences of the United States of America, Vol.110, No.15, pp.5802–5805 (2013).

[68] L. Backstrom, E. Sun, and C. Marlow, "Find Me If You Can: Improving Geographical Prediction with Social and Spatial Proximity," Proceedings of the 19th International Conference on World Wide Web, pp.61–70 (2010).

[69] Z. Cheng, J. Caverlee, and K. Lee, "You Are Where You Tweet: A Content-Based Approach toGeo-locating Twitter Users,"Proceedings of the 19th ACM International Conference on Information and Knowledge Management, pp.759–768 (2010).

[70] B. Hecht, L. Hong, B. Suh, and E. H. Chi, "Tweets from Justin Bieber's Heart: The Dynamics of the "Location" Field in User Profiles," Proceedings of the International Conference on Human Factors in Computing Systems, pp.237–246 (2011).

[71] J. D. Burger, J. Henderson, G. Kim, and G. Zarrella, "Discriminating Gender on Twitter," Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing, pp.1301–1309 (2011).

[72] D. Rao, D. Yarowsky, A. Shreevats, and M. Gupta, "Classifying Latent User Attributes in Twitter," Proceedings of the 2nd International Workshop on Search and Mining User-Generated Contents, pp.37–44 (2010).

[73] J. Ito, T. Hoshide, H. Toda, T. Uchiyama, and K. Nishida, "What is he/she like?: Estimating Twitter user attributes from contents and social neighbors," 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining, pp.1448–1450 (2013).

[74] J. Marui, N. Nori, T. Sasaki, and J. Mori, "Empirical Study of Conversational Community Using Linguistic Expression and Profile Information," International Conference on Active Media Technology 2014, Active Media Technology, pp286–298 (2014).

[75] J. Chen, R. Nairn, L. Nelson, M. Bernstein, and E. H. Chi, "Short and Tweet: Experiments on Recommending Content from Information Streams," Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp.1185-1194 (2010).

[76] F. Abel, Q. Gao, G. J. Houben, and K. Tao, "Analyzing User Modeling on Twitter for Personalized News Recommendations," International Conference on User Modeling, Adaptation, and Personalization 2011, User Modeling, Adaption and Personalization, pp.1–12 (2011).

[77] P. Kapanipathi, P. Jain, C. Venkataramani, and A. Sheth, "User Interests Identification on Twitter Using a Hierarchical Knowledge Base," European Semantic Web Conference 2014, The Semantic Web: Trends and Challenges, pp.99–113 (2014).

[78] M. Michelson and S. A. Macskassy, "Discovering users' topics of interest on twitter: A first look," Proceedings of the 4th Workshop on Analytics for Noisy Unstructured Text Data, pp.73–80 (2010).

[79] F. Zarrinkalam, H. Fani, E. Bagheri, M Kahani, and W. Du, "Semantics-Enabled User Interest Detection from Twitter," 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, pp.469–476 (2015).

[80] P. Kapanipathi, F. Orlandi, A. Sheth, and A. Passant, "Personalized Filtering of the Twitter Stream," Proceedings of the 2nd International Conference on Semantic Personalized Information Management: Retrieval and Recommendation, Vol.781, pp.6–13 (2011).

[81] P. Bhattacharya, M. B. Zafar, N. Ganguly, S. Ghosh, and K. P. Gummadi, "Inferring User Interests in the Twitter Social Network," Proceedings of the 8th ACM Conference on Recommender Systems, pp.357–360 (2014).

[82] J. Han and H. Lee, "Characterizing the Interests of Social Media Users: Refinement of a Topic Model for Incorporating Heterogeneous Media," Information Sciences, Vol.358–359, pp.112–128 (2016).

[83] G. Chen and D. Kotz, "A Survey of Context-Aware Mobile Computing Research," Department of Computer Science, Darthmouth College, TR2000-381 (2000).

[84] M. V. Setten, S. Pokraev, and J. Koolwaaij, "Context-Aware Recommendations in the Mobile Tourist Application COMPASS," International

Conference on Adaptive Hypermedia and Adaptive Web-Based Systems, Adaptive Hypermedia and Adaptive Web-Based Systems, Vol.3137, pp.235–244 (2004).

[85] A. Asthana, M. Crauatts, and P. Krzyzanowski, "An Indoor Wireless System for Personalized Shopping Assistance," 1st Workshop on Mobile Computing Systems and Applications, pp.69–74 (1994).

[86] G. D. Abowd, C. G. Atkeson, J. Hong, S. Long, R.Kooper, and M. Pinkerton, "Cyberguide: A mobile context-aware tour guide," The Journal of Mobile Communication, Computation and Information, Wireless Networks, Vol.3, No.5, pp.421–433 (1997).

[87] k. Sudo, S. Tatsuta, R. Sugihara, and T. Kato, "TPO-Based Personal Information Assistance by Wearable Sensors and Application to Recommendation Service," JSKE Journal, Vol.8, No.3 pp.843–849 (2009).

[88] X. Yin, J. Han, and P. S. Yu, "Truth Discovery with Multiple Conflicting Information Providers on the Web," IEEE Transactions on Knowledge and Data Engineering, Vol.20, Issue 6, pp.796–808 (2008).

[89] J. Pasternack and D. Roth, "Latent Credibility Analysis," Proceedings of the 22nd International Conference on World Wide Web, pp.1009–1019 (2013).

[90] X. Dong, E. Gabrilovich, K. Murphy, V. Dang, W. Horn, C. Lugaresi, S. Sun, and W. Zhang, "Knowledge-Based Trust: Estimating the Trustworthiness of Web Sources," Proceedings of the VLDB Endowment, Vol.8, Issue 9, pp.938–949 (2015).

[91] M. A. Suryanto, E. P. Lim, A. Sun, and R. H. L. Chiang, "Quality-aware collaborative question answering: methods and evaluation," Proceedings of the 2nd ACM International Conference on Web Search and Data Mining, pp.142–151 (2009).

[92] B. T. Adler and L. D. Alfaro, "A content-driven reputation system for the wikipedia," Proceedings of the 16th International Conference on World Wide Web, pp.261–270 (2007).

[93] A. Dong, Y. Chang, Z. Zheng, G. Mishne, J. Bai, R. Zhang, K. Buchner, C. Liao, and F. Diaz, "Towards recency ranking in web search," Proceedings of the 3rd ACM International Conference on Web Search and Data Mining, pp.11–20 (2010).

[94] C. Castillo, M. Mendoza, and B. Poblete, "Information credibility on twitter," Proceedings of the 20th International Conference on World Wide Web, pp.675–684 (2011).

[95] W. W. Yang, "Liar, Liar Pants on Fire: A New Benchmark Dataset for Fake News Detection," Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, Vol.2, pp422–426 (2017).

[96] R. B. Cialdini, "Influence: The Psychology of Persuasion," Harper Business (2006).

[97] P. Bonard and M. A. Sasse, "'Knowing me, Knowing you' - Using Profiles and Social Networking to Improve Recommender Systems," BT Technology Journal, Vol.24, Issue 3, pp.84–98 (2006).

[98] W. Woerndl and G. Groh, "Utilizing Physical and Social Context to Improve Recommender Systems," Proceedings of the 2007 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology - Workshops, pp.123–128 (2007).