Dissertation for the Degree of Doctor

A Personalized Recommender System based on Friendship Strength using Social Big Data

by
Young-Duk Seo

Department of Computer Science and Engineering
Graduate School
Korea University

December 2017

A Personalized Recommender System based on Friendship Strength using Social Big Data

Advisor: Professor Doo-Kwon Baik

by
Young-Duk Seo

A thesis submitted to the Graduate School of the Korea University in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Engineering

Department of Computer Science and Engineering
Korea University

December 2017

The Dissertation of **Young-Duk Seo** is Approved

by

Doo-Kwon Baik, Committee Chair

Hae-Chang Rim, Committee Member

Hoh Peter In, Committee Member

Young-Gab Kim, Committee Member

Jangwon Gim, Committee Member

December 2017
Korea University

A Personalized Recommender System based on Friendship Strength using Social Big Data

ABSTRACT

The rapid growth of social network services has produced a considerable amount of data, called social big data. Social big data are helpful for improving personalized recommender systems because these enormous data have various characteristics. Therefore, many personalized recommender systems based on social big data have been proposed, in particular models that use people relationship information. However, most existing studies have provided recommendations on special purpose and single-domain social big data sets that have a set of users with similar tastes, such as MovieLens and Last.fm; nonetheless, they have considered closeness relation. In this dissertation, we introduce an appropriate measure to calculate the closeness between users in a social circle, namely, the friendship strength. Furthermore, we propose a friendship strength-based personalized recommender system that recommends topics or interests users might have to analyze social big data, using Twitter in particular. The proposed measure provides precise recommendations in multi-domain environments that have various topics. We evaluated the proposed system based on various evaluation metrics using one month’s Twitter data. Our experimental results show that our personalized recommender system outperforms the baseline systems, and friendship strength is of great importance in personalized recommendation.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In this chapter, we introduce the motivation and the purpose of this dissertation. Furthermore, we also present the taxonomy and organization of the dissertation.

1.1 Research Motivation

Recently, considerable real-time data have been generated because of the increase in the use of social network services (SNSs) such as Facebook[1], Twitter[2], Delicious[3], and Last.fm[4]. With the development of SNSs, more than one billion users are now

1 http://www.facebook.com/
2 http://twitter.com/
3 http://delicious.com/
4 http://www.last.fm/

using these services. Through SNSs, users can express their opinions in an unconstrained manner, and share their interests with others. This spontaneous participation of users in SNSs results in the generation of enormous amounts of data with various characteristics, called social big data [Cambria13]. Social big data have been used in various studies in many research fields because of their massiveness and variety [Manovich11]. In these fields, active research on personalized recommender systems has been conducted to provide appropriate information to users according to their demands and preferences [Guy13].

Traditional personalized recommender systems employ mainly a collaborative filtering (CF) algorithm. A CF provides recommendations to users by analyzing their individual characteristics to utilize the information of other users who are highly similar to them [Herlocker99]. Similarity between users is measured by their preference value of a specific item. A user's preference value is obtained either explicitly or implicitly [Bellogin13]. Explicit information refers to the user's rating of the items that can be represented by a numerical value, for example, from 0 to 5 points. On the other hand, implicit information is acquired by monitoring the user's behavior. For example, the number of times that the user has listened to a particular song, the number of applications downloaded, and the web pages visited can be implicit indicators of the user's preferences.

Table 1. The main properties of social big data used in the personalized recommender systems

	Contents Generated by Users	Relationship Information	Interaction Information
Definition	All contents that are created by the users themselves, and their meta-information	A social circle representing directly connected relationship among users on SNSs	Messages or contents exchanged between users
Example	• Tweets on Twitter • Posts on blog • Posting time of contents	• The follower-followee relationship on Twitter • Friend list on Facebook	• Mention and retweet on Twitter • Review sharing on Yelp • Message on Facebook

Social big data enable us to consider new diverse features to calculate the similarity between users, which cannot be used in traditional personalized recommender systems [Bellogin13]. These features have three main properties:

“Contents generated by users”, “Relationship information” and “Interaction information”. “Contents generated by users” refers to all contents that are created by the users themselves, and their meta-information, such as tweets on Twitter, posts on blog, posting time of contents and tag information [Bobadilla13]. “Relationship information” constitutes a social circle representing directly linked or connected relationships among users on SNSs, such as the follower-followee relationship on Twitter and the friends list on Facebook. “Interaction information” refers to messages or contents exchanged between users, such as mention and retweet on Twitter, review sharing on Yelp, and message on Facebook [Nepal13]. Using these features, we can improve the recommendation quality as compared to that of traditional systems. Table 1 summarizes the above three properties mentioned above describing the social big data used in personalized recommender systems.

1.2 Research Purpose

In order to calculate the similarity between users, conventional CF methods use similarity measures such as the Pearson correlation coefficient (PCC) [Herlocker 99] and Jaccard mean squared difference (JMSD) [Bobadilla10]. However, these measurements are not suitable for social big data-based personalized recommendations. This is because existing similarity measures are suitable for utilizing explicit rather than implicit information, whereas most social big data comprise implicit information [Chen13][Ma13]. Further, users trust closely related users' information in a social circle, and many research studies have verified that their information is useful for upgrading personalized services in practice [Servia14]. However, existing similarity measures are not appropriate for calculating the closeness among users in a social circle, because they cannot easily consider the various characteristics of social big data, except for contents generated by users [Bobadilla10][Liu13][Liu14a].

Moreover, most personalized recommender systems are based mainly on the relation information among users on social circle. If users are closely connected or linked to each other in their social circle, there is a high probability that they have similar interests and interact with each other actively [Nepal13]. In many studies, the degree of closeness between users was measured through the social big data. However, most of them provide recommendations using only the data of single-domain social big data sets, such as MovieLens and Last.fm [Konstas09][Servajean14]. In other words, they utilize sets of users who have similar tastes for specific domains. Users generate contents about numerous topics and form a relationship with other users who have various interests in a variety of topics. However, existing personalized recommender systems do not consider the number of topics users share with each other, because they are used only in a single topic domain. In addition, little works have been conducted on the multi-domain social circles that are formed by users with various topics and interests.

In this dissertation, to overcome the limitations of existing research, we propose a novel approach for measuring closeness between users that considers various features of social big data, particularly “Contents generated by users”, “Relationship information” and “Interaction information”. We refer to the closeness measure as friendship strength. Our proposed personalized recommender system can utilize the information of closely connected users; furthermore, this system can recommend appropriate interests. In other words, the recommended items in this dissertation are the interests or topics, especially smartphone, music, movie, and drama in which users might be interested.

The main contributions of this dissertation are as follows:

(1) The implicit information of social big data, which has not been utilized in the existing similarity measures, such as PCC and JMSD, can be used to calculate friendship strength between users. More specifically, friendship strength is determined by dividing it into interaction, group, and preference similarity, and is calculated by the combination of the three similarities.

(2) The proposed personalized recommender system provides appropriate recommendation results for users by using the information of other users who have a high level of friendship strength with them.

(3) We use data from Twitter, which is a multi-domain rather than single-

domain social big data sets for measuring performance. Through an extensive experiment, we verify that the performance of our system is high level for multi-domain.

(4) In this dissertation, to evaluate the superiority of our proposed friendship strength-based recommender system, we used various evaluation metrics for personalized recommender systems: precision, recall, F1-measure, mean absolute error (MAE), and normalized discounted cumulative gain (NDCG). We verify that the performance of our approach is better than that of baselines in all metrics. In addition, friendship strength plays an important role in personalized recommendation.

1.3 Research Taxonomy

This section describes the taxonomy of our dissertation in the recommender system area as shown in Figure 1.

A recommender system is largely classified into group recommendation [Jameson07] [Seo18] and personalized recommendation [Bobadilla13] [Seo17]. Group recommendation aims to provide the appropriate information for group members to analyze the characteristics and propensity of a group [Jameson07] [Seo18]. Although group recommendation resolves the cost problem of

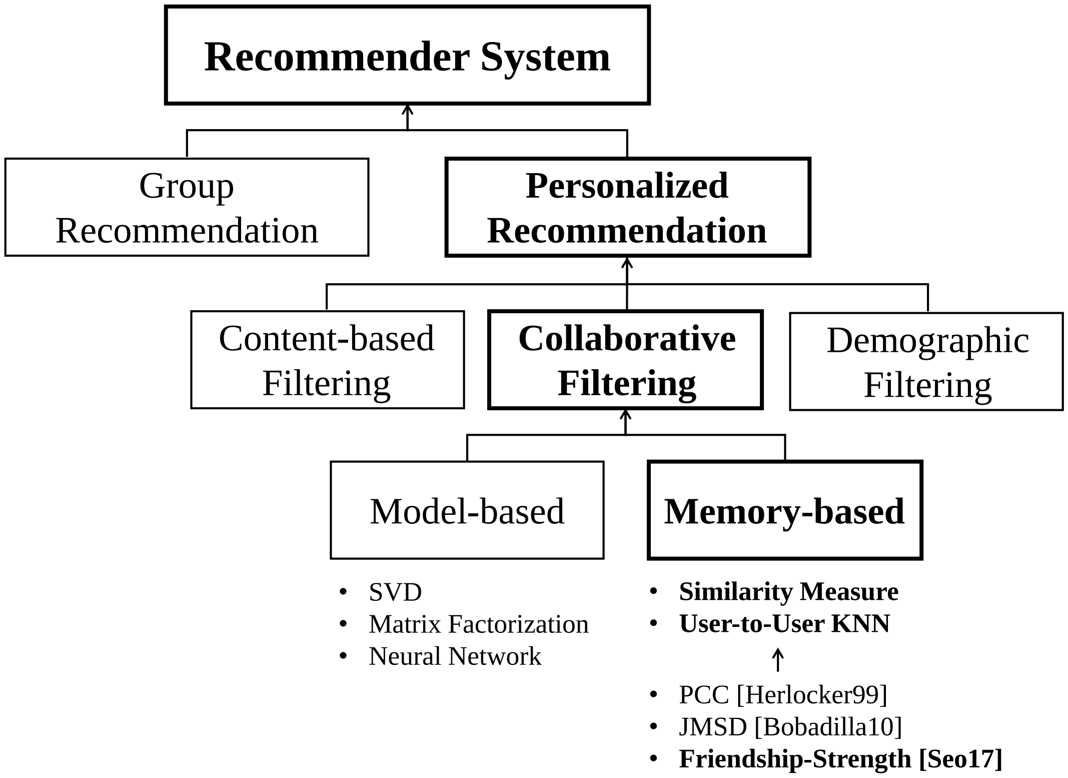


Figure 1. Taxonomy of the personalized recommender system based on social big data

personalized recommendation, most recommender systems that are used in e-commerce, social commerce, and SNSs focus on providing the appropriate items or interests for users according to their personal preferences [Bobadilla13] [Seo17]. Therefore, in this dissertation, we also focus on personalized recommendation.

In the personalized recommendation research area, filtering algorithms have been mainly used in most studies [Pazzani99] [Adomavicius05] [Bobadilla13]. The most general classifications of filtering algorithms are as follow: demographic filtering (DF) [Krulwich97], content-based filtering (CB) [Li08] [Sen09], CF

[Golbeck06] [Konstas09], and hybrid filtering (HF) [Guy10]. First, DF is the most primitive filtering technique, and it classifies groups according to the basic attributes of people, such as sex, age, and country [Krulwich97]. DF based-CF provides recommendations to these classified groups. CB recommends items that are similar to those for which the user has shown interest in the past [Li08] [Sen09]. This filtering approach has relatively high performance compared to the others if all users have enough preference values. If not, it causes a cold-start problem. In the CF approach, items are recommended to users based on other individuals who are found to have similar preferences or taste [Golbeck06] [Konstas09]. CF requires a long time to produce the recommendation results, but it can consider more data compared to CB. Finally, hybrid filtering is the combination of CF and DF [Vozalis07], or CF and CB [Guy10]. These combinations complement each disadvantage and highlight each advantage. Among all filtering techniques, CF is the most used in academic fields[Herlocker99] [Golbeck06] [Bobadilla10] [Liu13] as well as commercial areas [Saleem08] [Ekstrand11] [Victor11], and it is of great importance in personalized recommender systems [Bobadilla13].

Finally, CF can be divided into model-based approach and memory-based approaches. The model-based approach uses a specific model to find the hidden features of user preference, and provides recommendation to users. The widely used models in this approach are the Bayesian classifiers [Park07], neural networks [Wu17] [He17], fuzzy systems [Yager03], genetic algorithm [Babadilla11], singular value decomposition (SVD) [Toroslu10], and matrix factorization (MF) [Koren09]. Conversely, the memory-based approach mainly uses similarity metric between two users or two items based on matrix of user ratings for items. Cosine [Adomavicius05], PCC [Herlocker 99], and JMSD [Bobadilla10] are the famous similarity measurements in this approach. Among the two approaches, in this dissertation, we focus on memory-based approach, especially in calculating the similarity measurement among users [Seo17].

1.4 Organization of the Dissertation

The remainder of this dissertation is structured as follows. In Section 2, we describe the existing personalized recommender systems that leverage social big data, particularly considering the people relationships. Section 3 explains our friendship strength-based personalized recommender system, which calculates the closeness among users based on social big data. Section 4 describes an evaluation framework for evaluating the performance of the proposed system and displays a comparative evaluation and the usage of social big data between the proposed system and the existing systems. In Section 5, we discuss whether our proposed friendship strength is applicable to other recommender system based on social big data and our limitation. Section 6 presents conclusion and describes the direction of future work.

CHAPTER 2

Related Work

In this chapter, we describe the related works on personalized recommender systems based on social big data, especially, focusing on the people relationship information. We divided the existing studies into cases of only using user-generated contents which is not using relationship information and cases of using relationship information.

2.1 Personalized Recommender System based on User-Generated Data

A personalized recommender system collects information about the preference of its users for items. Using this preference information, it recommends items that its users may wish to acquire. In previous systems, this information could be obtained by using the explicit or implicit information. As a result of using social big data such as follower-followee, friends' lists, tweets, blog posts, and tags, we can acquire more information to enhance personalized recommendations, compared with the recommendations provided by systems that not use such data. Therefore, there has been extensive research on the use of social big data in personalized recommender systems.

2.1.1 Personalized Recommender System based on Explicit Data

Many methods measure the similarity between users by utilizing preference values, particularly explicit ratings like PCC [Herlocker99]. [Liu13] utilized PCC and extended it to a version that can handle contextual information. [Bobadilla10] proposed JMSD, which combines the Jaccard measure and MSD. They considered the ratio of common ratings as well as the absolute difference of ratings between two users. [Liu14a] proposed the new heuristic similarity model (NHSM) and calculated similarity by using not only users' own ratings, but also the global preference reflected in user behaviors. [Zhu14] gave a weight to popular items in every users-to-items rating matrix, and used cosine similarity to calculate similarity between users. Then, they predicted the interest of users with respect to them.

However, the similarity measures of the above approaches are mainly based on explicit information, although they use public social big data, such as MovieLens, Netflix, FilmAffinity, and Epinions. Therefore, they do not yield an appropriate similarity measures for social big data-based personalized recommender systems, because it is difficult to utilize implicit information properly.

2.1.2 Personalized Recommender System based on Implicit Data

Social big data include various implicit data, particularly user-generated data, such as tags and profiles, because of their intrinsic nature. Therefore, many researchers utilized these implicit data to improve their recommender systems.

[Firan07] studied personalized track recommendations using data from Last.fm. They analyzed the tag usage statistically and showed that the user profiles based on these tags could produce better recommendations than the conventional ones based on track usage. [Li08] discovered the common interests shared by groups of users by using the user-generated tags on the social bookmarking site Delicious. [Liu14b] proposed a personalized tag recommendation system on Flickr, which matched new updated photos with geo specific tags. They used both the tagging history of users and the geographic information to generate recommended tags based on a learning method. [Yin14] [Yin15] proposed user behavior model, namely temporal context aware mixture model (TCAM) and extended TCAM to dynamic temporal context aware mixture model (DTCAM). They observed rating behaviors of users based on two factors: user implicit preferences and temporal attentions of the whole social circle on the SNS. [Servajean14] tried to find the relevant users set (i.e., cluster) for a specific user in order to provide recommendations. They proposed a new clustering algorithm for recommendations based on their proposed similarity measure, namely usefulness and evaluated it by using MovieLens, Flickr and Last.fm.

However, the above approaches are based on the information of unspecified individuals who are very similar to the target users, but are not linked with them. Therefore, they have difficulty reflecting personal tendencies sufficiently and yield a low recommendation accuracy as compared with approaches that consider users' relationship information on social circle for recommendations.

2.2 Personalized Recommender System based on Relationship Information

Relationship information, which is a unique feature of social big data, is highly appropriate information for improving the performance of personalized recommender systems. For this reason, many research studies on exploiting this information have been conducted.

Personalized recommendation using relationship information on social big data is classified as recommended by influentials [Lin14] or friends [Guy10]. Traditional personalized recommender systems recommend items by using information about unspecified individuals who are not connected to the users. However, we can utilize the people relationship information on a social circle to recommend items to users by using information about the acquaintances connected to them. The degree of similarity of the preference between users who are connected to each other on a social circle is higher than that of users who are not connected, and the influence of the connected users on the social circle is an important factor affecting personalized recommendations [Servia14].

2.2.1 Personalized Recommender System based on Simply Connected Users

Some studies have proven that this has significantly improved the performance of personalized recommendation to use the information of the connected users.

[Geyer08] encouraged user participation through the About You platform that recommends useful contents to users for writing their profiles, particularly using people relationship and user-generated contents. They used a binary score to determine whether the users are connected or not, and gave a weight to the contents of the connected users. [Xu13] discovered the preference of users on microblog based on the information of their connected users. They focused on filtering out unnecessary connected users to predict the preference of specific user, as opposed to general approaches finding relevant users.

The aforementioned research uses a relationship with a number of users on social circle, but a connected relationship itself does not guarantee that two users have a friendly relationship. However, the above approaches do not consider the closeness between users on a social circle and treat them all equally. Therefore, most of the studies have provided recommendations only for single-domain social big data sets that have a set of users with similar interests and preferences.

2.2.2 Personalized Recommender System based on Influentials

An influential is one who plays an important role in the social circle [Kwak10], and influential-based recommendation is processed by the information of experts in their field or area. Therefore, influential-based recommendations are used mainly in recommender systems that require expertise, such as those in academic fields or related to news.

[Zhen09] defined a collaborative team as people with expert knowledge and provided recommendations using their information. [Tang08] [Tang12] collected researcher profiles from the web and developed the ArnetMiner system, which recommends experts and papers relevant to users. [Lin14] defined a person who exerts influence on the news community as an expert. They proposed PRemiSE, which uses their defined expert information and improves the quality of personalized news recommendations.

However, people tend to trust the opinions of acquaintances they know directly. According to a report by Harris Interactive [Heckathorne10], the opinion of friends has more credibility than that of an influential, when a person decides to purchase products. Therefore, influential-based recommendation is suitable for knowledge-intensive domains, such as academic fields or news [Tang08] [Tang12]

[Zhen09] [Lin14], but not for general domains including taste, interest or hobby, such as movie, music and drama; recently, a few studies have been conducted on general domains, such as Twitter [Bhattacharya14].

In addition, influence measurement does not consider the closeness among users, but is generally measured by focusing on the impact over the entire network. In other words, influential-based recommendation does not utilize relationship and communication information among users.

2.2.3 Personalized Recommender System based on Friends

In the case of the social big data environment, as mentioned previously, it is more efficient to provide recommendations based on the information of friends than that of influentials, because people tend to pay more attention to the opinions of acquaintances and friends than to those of influentials [Heckathorne10]. Thus, personalized recommendations have been provided using the information of friends in most of the domains, except in those related to professional or specialized knowledge.

In this dissertation, we call the strength of the connection friendship strength, which is a property for quantifying the closeness between users. Different terms for friendship strength were used in other research studies, such as tie strength [Granovetter73] [Servia14], intimacy [Rau08] [Seol15] and trust [Golbeck06] [Deng14]. Recently, several personalized recommender systems that consider the friendship strength have been proposed to improve the quality of recommendations.

[Konstas09] provided music recommendation based on Last.fm through Random Walk with Restart (RWR) by using users' play count, tag and friendship information. [Guy10] proposed a hybrid approach that combines CF and CB to recommend social media considering both people relationships and user-generated tags. [Golbeck06] presented FilmTrust, which is a web site using the trust among the users in a web-based SNS to provide predictive movie recommendations. He called this the reliability of users "trust," and assigned a weight to the information of high trust users for recommendations. [Servia14] took into account the interaction and the social circle information of users to calculate the tie strength between them. Further, they proposed a personalized model based on the tie strength to enhance social services. [Lai13] considered three types of influence factors for recommendation: social, interest and popularity. They calculated the social influence by measuring the rate of photo sharing between directly connected users. [Yu13] considered the interaction between users on social circle, called the users' popularity, to predict the interest of users. [Qian13] calculated the rating based on the interpersonal influence (i.e., trust relation) and the interest similarity between users, and proposed personalized recommender system based on these two factors along with personal preferences information. In particular, they calculated the similarity measures of multi-level of items to classify their category. [Ma14] proposed user recommendations on SNS considering both the relationship in the social circle and the topic similarity between users.

Through the previous research, it is validated that personalized recommender systems based on the friendship strength can provide higher quality recommendations than systems that treat all relations as the same. However, thus far, a multi-dimensional analysis has not been performed on the factors influencing the friendship strength used in existing studies. Not considering the various elements affecting the friendship strength and not combining them appropriately lead to an inappropriate measurement of the closeness between users. Consequently, a new method is required that analyzes the factors affecting the friendship strength appropriately and combines them to calculate friendship strength.

CHAPTER 3

Personalized Recommender System based on Friendship Strength

In this chapter, we first describe the principal methodology of our proposed personalized recommender systems. There are 10 steps to complete in order to provide the appropriate recommendation results to users.

Furthermore, more importantly, we explain the definition and the measurement of friendship strength to analyze social big data. In this dissertation, we consider various implicit social big data to calculate a similarity which is different from existing similarity measures in CF. We call our similarity friendship strength with three types of properties namely interaction, group, and preference similarities.

Finally, we propose a personalized recommender algorithm based on friendship strength. The proposed personalized recommender algorithm is based primarily on CF. CF-based recommender systems provide recommendations to users by using the information of their directly connected users or friends which is very useful for improving the recommendation quality. In particular, strongly connected users have a greater positive influence on each other than weakly connected users. Therefore, our proposed personalized recommender algorithm uses friendship strength as a similarity between users.

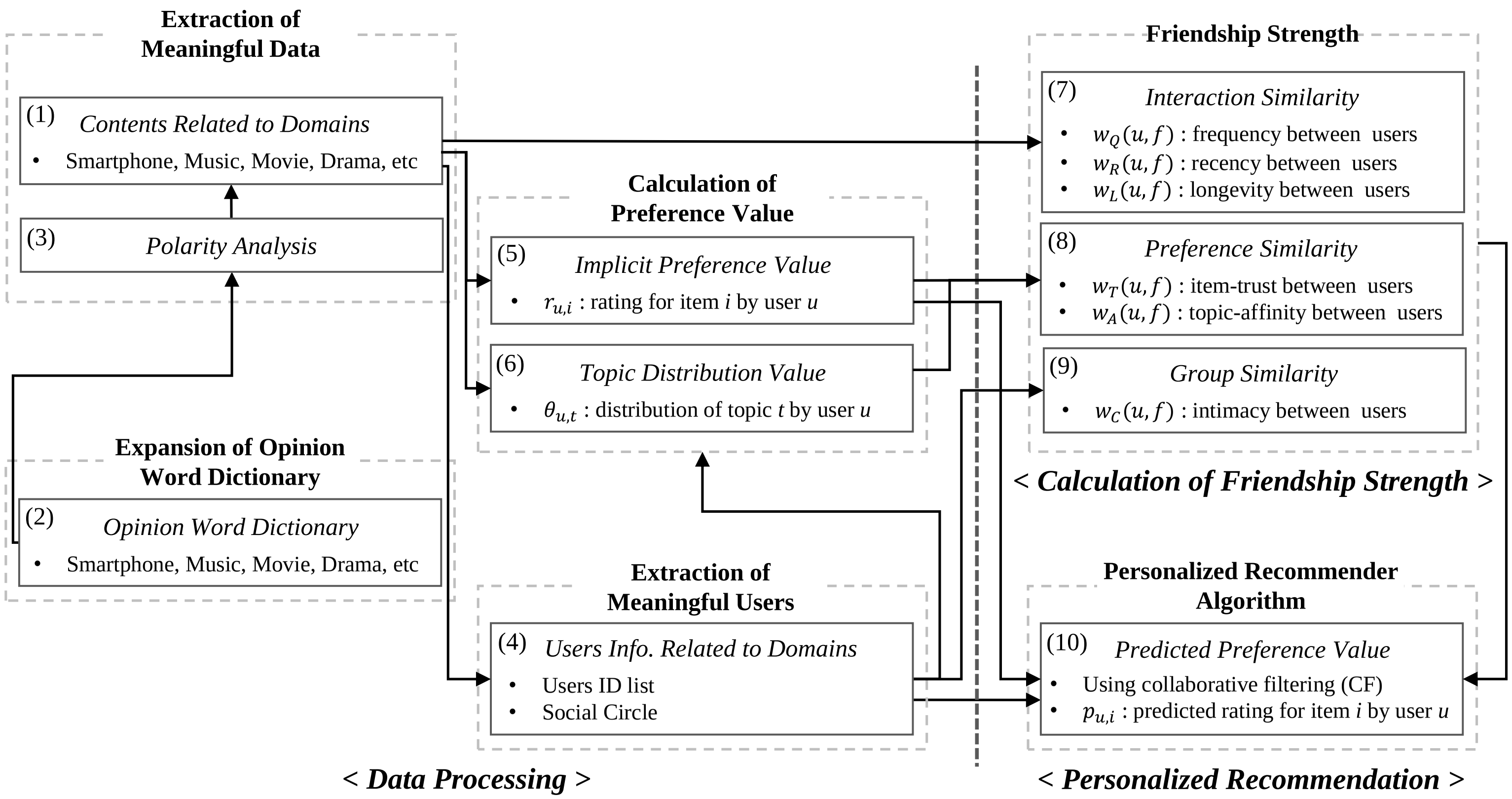


Figure 2. Methodology for the proposed personalized recommender system

3.1 Methodology

We propose the methodology of personalized recommender system to find relevant interests of users as shown in Figure 2. The methodology largely consists of three phases: data processing, calculation of friendship strength, and personalized recommendation phases. First, we process raw data for the first phase, and then calculate friendship strength between users. Finally, we find the interests of users to use friendship strength-based personalized recommender algorithm.

In the data processing phase, we process the data necessary for constructing our system. First, we extract data associated with the smartphone, music, movie, and drama domains (Step 1). Second, we extend the opinion word dictionary to add positive and negative words related to the music, movie, and drama domains based on the existing opinion word dictionary related to the dictionary for smartphone and prior research using English opinion word dictionary (Step 2). Third, using the opinion dictionary, we analyze the polarity of user-generated contents, such as tweets on Twitter, posts on Facebook, and reviews on Yelp (Step 3). Then, we extract users associated with the four domains and their relationship information. Using the extracted user information, we store the user's id list and create a users-to-users matrix for forming a social circle (Step 4). Finally, using scores of the polarity analysis of contents and the number of times the users refer to item, we measure the implicit preference value (Step 5) and the topic distribution value (Step

6) used in the proposed friendship strength and personalized recommender system.

In the calculation of friendship strength phase, we calculate the friendship strength between users. First, we calculate interaction similarity using communication information between users such as, retweet and mention in Twitter, reviews on Facebook and Yelp, and contact information in Delicious (Step 7). Second, we measure preference similarity using preference values, which are implicit preference value and topic distribution value (Step 8). And then, we calculate group similarity using users' social circle information on SNS (Step 9).

Finally, using friendship strength between users and implicit preference values, we calculate predicted preference value based on CF algorithm and provide appropriate recommendation results for users (Step 10). The detailed explanations for core steps are discussed in sections below.

Table 2. Example of our data set

Document_ID	Content	Keyword	Posting_Time	Authour_ID	Is_RT	Shared_ID	Reply_Author_ID	Polarity
219181556185567233	@_Yeo**** iphone is awesome	Iphone	20120701063209	295093845	N	Null	298506140	Positive
219103656413368320	Galaxy's battery is bad	Galaxy	20120701012237	115300182	N	Null	Null	Negative
220846507283329025	RT @Uly**** new firmware for optimus has been uploaded	Optimus	20120705204805	115300182	Y	220846361908744192	Null	Neutrality

3.2 Data Extraction (Step 1 in the Methodology)

Our data set is divided into contents-related data and the user information related to the contents. The former consists of the content id (Document_ID), user-generated contents (Content), keyword of contents (Keyword), the posting time of the contents (Posting_Time), and the retweeted content document id (Shared_ID); the latter consists of the user id (Author_ID) and the mentioned user id (Reply_Author_ID). We determined the polarity of contents (Polarity) to use opinion word dictionary and contents syntax analysis information. Table 2 represents the example of our stored data set.

3.3 Polarity of Contents (Steps 2 and 3 in the Methodology)

For personalized recommendation in our data set, we need to convert preference information of users into rating value. In this subsection, we explain the polarity analysis method for our proposed recommender system.

Users express their emotions or feelings in the contents they post on SNSs, such as Twitter, and Facebook. Thus, we should consider the polarity of contents when using the user-generated contents on twitter for making personalized recommendations. To classify polarity of Korean contents, we utilize a smartphone-related opinion word dictionary provided by DaumSoft[5], which is a data analysis company in Korea. Further, based on a previous study [Esuli06] using an English opinion word dictionary called SentiWordNet, we find representative words in Korean of negative and positive opinions, and create our Korean opinion dictionary expanded to fit the certain domains (e.g., music, movies, and drama). Our opinion word dictionary includes both negative and positive words, as shown in Table 3.

Table 3. Opinion word dictionary

Topic	**Polarity**	
	Positive #	**Negative #**
Smartphone	384	509
Music	220	182
Movie	225	211
Drama	227	211

On the basis of this dictionary, we classify a user-generated content D as positive, neutral or negative by providing a rating of 1, 0.5, or 0, and call it the

5 http://www.daumsoft.com/

polarity score of content D, defined as equation (1):

$$o_D = \begin{cases} 0 & if \quad D\ is\ negative \\ 0.5 & if \quad D\ is\ neutrality \\ 1 & if \quad D\ is\ positive \end{cases} \tag{1}$$

3.4 Social Circle (Step 4 in the Methodology)

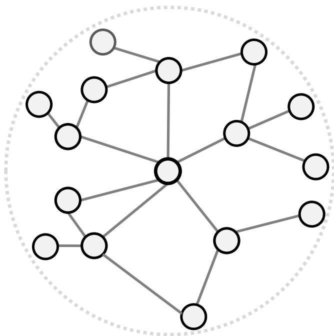
We extract users associated with the four domains and their relationship information. Using the extracted user information, we store the user's id list and create a users-to-users matrix for forming a social circle. Figure 3 represents the process of forming a social circle in our dissertation.

14696334
15187867
15296359
16731282
17228247
17460958
17615156

Users ID List

0	0	0	0	1	0	0
0	0	0	1	0	0	0
0	0	0	0	0	0	0
0	1	0	0	0	0	0
1	0	0	0	0	0	0
0	0	0	0	0	0	0
0	0	0	0	0	0	0

Users-to-Users Matrix



Social Circle

Figure 3. The process of forming a social circle to use the relationship information

3.5 Preference Value of Users

(Steps 5 and 6 in the Methodology)

Most of the social big data-based personalized recommender systems measure the users' preferences by using implicit information such as user-generated contents.

In this dissertation, to measure preference value of users, we extract keywords (i.e., items) from the user-generated contents and call the domain containing the keywords a topic. In other words, topic is a higher level of items. The classification of item and topic layers in this dissertation is shown in Figure 4. Then, we calculate item and topic preferences of users respectively.

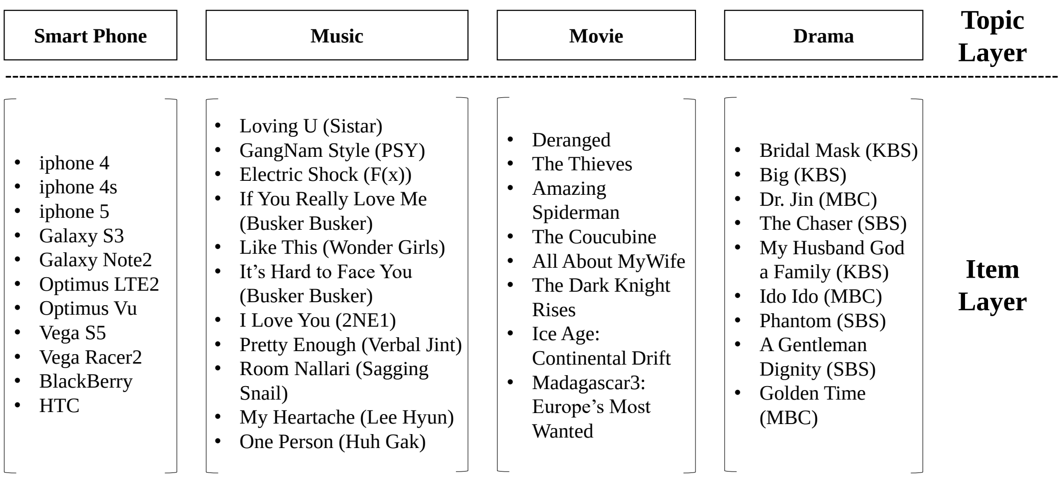


Figure 4. The classification of item layer and topic layer

First, we calculate the implicit indicator of item preference based on the polarity of contents and the total number of the user-generated contents. If the polarity of the content D is o_D and the set of the contents about item i by user u is $D_{u,i}$, the item preference value $r_{u,i}$ is defined as equation (2):

$$r_{u,i} = \frac{\sum_{D \in D_{u,i}} o_D}{|D_{u,i}|} \tag{2}$$

We also measure topic distribution value as a topic preference of users. Topic is a high level of items, and topic distribution value is calculated by using distribution of the total number of the user-generated contents associated with topic t. If D_u is the set of the contents by user u and $D_{u,t}$ is the set of the contents about topic t by u, the topic distribution value $\theta_{u,t}$ is defined as equation (3):

$$\theta_{u,t} = \frac{|D_{u,t}|}{|D_u|} \tag{3}$$

3.6 Friendship Strength

(Steps 7, 8 and 9 in the Methodology)

We consider various elements that affect the friendship strength on SNSs to calculate the friendship strength between users suitable for a personalized recommender system. The proposed friendship strength is classified into three types of similarity: interaction, group, and preference similarities. Table 4 provides the definition of the properties of friendship strength.

Table 4. Representation and definition of friendship strength

	Property	Weight	Definition
Interaction Similarity	Frequency	$w_Q(u, f)$	How much user u communicates with his/her friend f
	Recency	$w_R(u, f)$	How recently user u lasts encountered his/her friend f
	Longevity	$w_L(u, f)$	How long is the contact between user u and his/her friend f
Group Similarity	Intimacy	$w_I(u, f)$	How similar are the social circles of user u and his/her friend f
Preference Similarity	Item-Trust	$w_T(u, f)$	How similar are the interests of user u and f
	Topic-Affinity	$w_A(u, f)$	How many topics have user u and f discussed

3.6.1 Interaction Similarity

Interaction similarity is measured through "*the breadth and the depth of interaction*" between users [Rau08]. If two users have similar tastes or preferences, they share information actively [Nepal13]. Therefore, the interaction similarity between users is an important factor of friendship strength. In the case of Twitter, users share information and communicate with each other using the "mention" or "retweet" function. Similarly, the users of Facebook and Yelp can interact with each other using the reviews. We calculate the interaction similarity to use these functions. In this dissertation, we use the frequency, recency, and longevity properties to measure the interaction similarity.

First, we count the number of times the users communicate with each other on SNSs to measure frequency. In previous research [Nepal13], it is likely that strongly connected users, who communicate actively, exchange more useful information with each other. Thus, frequency should be assigned for personalized recommendations as a weight. This count is based on the number of contacts between users u and f, and is defined as $q(u, f)$. Frequency is measured by normalization in the range between 0 and 1 with a log function, which was applied to a personalized Twitter search by [Vosecky14], and is defined as equation (4):

$$w_Q(u, f) = \begin{cases} \log_{10}\{1 + q(u, f)\} & if \quad q(u, f) < 10 \\ 1 & if \quad q(u, f) \geq 10 \end{cases} \tag{4}$$

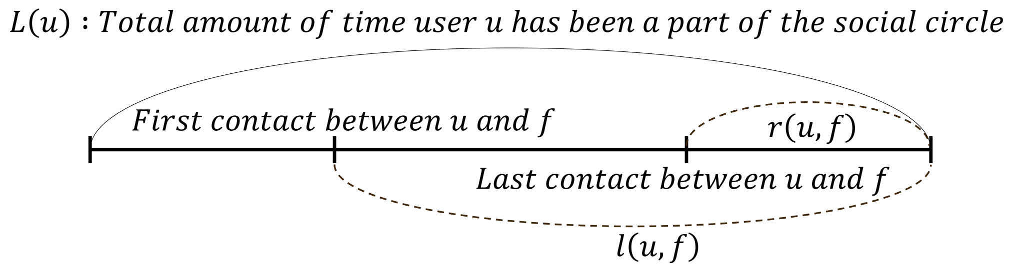


Figure 5. Concept of functions $L(u)$, $r(u, f)$ and $l(u, f)$

While frequency is a weight related to the number of contacts, recency and longevity are weights related to the contact-time between users. To calculate the latter two weights, we define three functions: $L(u)$, $r(u, f)$ and $l(u, f)$, as shown in Figure 5.

Recency measures how recently the users have contacted or communicated with each other. A considerable amount of real-time data is generated on SNSs, and therefore, the most recent information is very important [Dai10]. Further, recent contact means that the users share a current interest. Therefore, whether the users have contacted each other recently or not is a key factor in measuring the closeness relationship. $r(u, f)$ denotes a value that is the measure of how recently users u and f contacted each other, and it is defined as the elapsed time from the last contact to the current time. $L(u)$ represents the total amount of time during which the data of user u were collected. Recency is calculated as the ratio of $r(u, f)$ to $L(u)$, and is expressed as equation (5):

$$w_R(u, f) = \frac{L(u) - r(u, f)}{L(u)} \tag{5}$$

Longevity measures the duration of the users' contact with each other. The information of a person who communicates with another for a relatively long period of time is more important than that of one who does not [Daly09]. Where $l(u, f)$ denotes a period of contact between users u and f, longevity is defined as equation (6):

$$w_L(u, f) = \frac{l(u, f)}{L(u)} \tag{6}$$

Interaction similarity is calculated by using a weighted sum of frequency, recency and longevity in order to consider all the communication-based friendship strength properties as equation (7):

$$\begin{gathered} Sim_I(u, f) = \alpha_I w_Q(u, f) + \beta_I w_R(u, f) + \chi_I w_L(u, f) \\ (\alpha_I + \beta_I + \chi_I = 1) \end{gathered} \tag{7}$$

3.6.2 Group Similarity

Group similarity is an element of friendship strength related to the similarity of the social circles to which the users belong on the SNS, and we define it as intimacy. In general, the users' social circles provide important information regarding their tastes or preferences, because they tend to add users who are in close contact with them offline or have similar interests to their list of friends. In other words, two

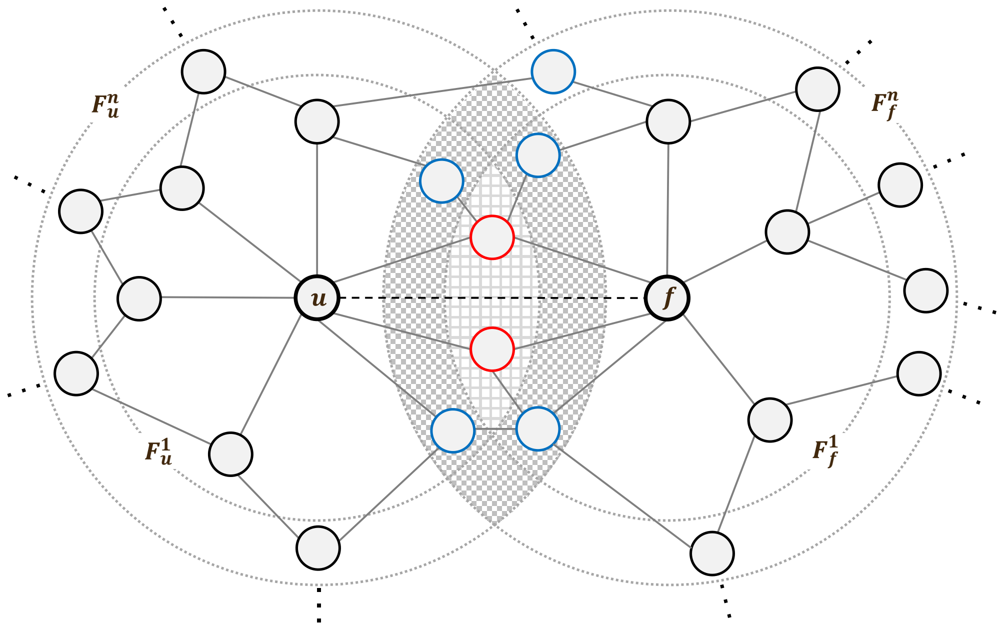


Figure 6. Intimacy (Group Similarity) between user u and his/her friend f

users with similar friend lists have a high probability of closeness [Lee09] [Zanda12] [Seol15]. Therefore, the proposed intimacy is calculated using the ratio of shared friends between groups of users u and f.

The existing intimacy is used for calculating the group similarity between users by using only directly related friends' sets [Seol15]. However, by taking indirectly related friends into account, we can obtain information and resources beyond those available in the users' own social circle. Further, indirectly connected users play a significant role in the flow of information in the SNS [Granovetter73]. Therefore, it is important to consider the information of indirectly related friends. Hence, the proposed intimacy applies the concept of user friend level to make sufficient use of the information of indirectly connected users, as shown in Figure 6. We define level-one F_u^1 as consisting of those users who are directly connected to user u, and level-n F_u^n as a set of indirectly connected users, such as friends of friends of user u. For example, F_u^2 denotes a set of users who are connected within two edges to user u. In this dissertation, we set n as six because almost all users on SNSs are connected within six edges [Kwak10].

An intimacy that consider both directly and indirectly connected groups of users u and f, who are connected to each other, is used for calculating the group similarity to utilizing the Jaccard measure [Koutrika09]. It is defined as equation (8):

$$Sim_G(u, f) = w_I(u, f) = \alpha_G \frac{\left|F_u^1 \cap F_f^1\right|}{\left|F_u^1 \cup F_f^1\right|} + \beta_G \frac{\left|F_u^n \cap F_f^n\right|}{\left|F_u^n \cup F_f^n\right|} \quad (\alpha_G + \beta_G = 1) \tag{8}$$

3.6.3 Preference similarity

Preference similarity denotes the degree of similarity between user-generated contents on social big data, and it is calculated by the preference value of items and the distribution of their topics. In this dissertation, we calculate the item-trust, which is the similarity of preference for items, and the topic-affinity, which is the similarity of the distribution of topics based on a classification of user-generated contents.

Measuring the degree of preference similarity between users is important in CF-based personalized services [Bobadilla10] [Liu13] [Liu14], because SNS users in a relatively intimate relationship with each other have a similar interest. In a personalized recommender system, several similarity measures can be used for calculating the preference similarity between users. Among them, PCC is the most popular measure and is thus the measure mainly used. However, PCC is not appropriate for a social big data-based recommender system that use implicit information, which does not constitute an explicit numerical preference value of items, because it is more suitable for recommender systems that include explicit information such as the ratings for items.

Because users mention a variety of interests or preferences on SNSs, in particular on Twitter, the data of which were used for the experiments in this dissertation, it is vital to consider the ratio of common interests rather than to measure only the numerical similarity of preferences. Therefore, we utilize JMSD [Bobadilla10], which considers the ratio of common interests as well as the numerical interest similarity, to measure the similarity of the preference value between users. JMSD is calculated as the product of the MSD and the Jaccard measure, which measure the preference similarity and the ratio of common interest, respectively. MSD is the average of the difference between the preference values for an item for users u and f. $d_{u,f}^{i}$ denotes the square of the difference between the preference values for item i for users u and f; this value is considered only when both users have a preference value for item i. Further, $d_{u,f}$ represents a set of $d_{u,f}^{i}$. Then, MSD is defined as equation (9):

$$
\begin{aligned}
MSD(u,f) &= \frac{\sum_{i \in I} d_{u,f}^{i}}{\left| d_{u,f} \right|} \\
d_{u,f}^{i} &= \begin{cases} (r_{u,i} - r_{f,i})^2 & if \quad r_{u,i} \neq null \wedge r_{f,i} \neq null \\ null & if \quad r_{u,i} = null \vee r_{f,i} = null \end{cases}
\end{aligned} \tag{9}
$$

The Jaccard measure is calculated as the ratio of common mentioned item on SNSs. If r_u and r_f denote a set of preference value for users u and f, respectively, then the Jaccard measure is defined as equation (10):

$$
Jaccard(u,f) = \frac{\left| r_u \cap r_f \right|}{\left| r_u \cup r_f \right|} = \frac{\left| d_{u,f} \right|}{\left| r_u \right| + \left| r_f \right| - \left| d_{u,f} \right|} \tag{10}
$$

According to MSD and Jaccard, we calculate item-trust, which is a similarity of preference value between users u and f as equation (11):

$$
w_T(u,f) = \{1 - MSD(u,f)\} \times Jaccard(u,f) \tag{11}
$$

SNS users can generate a large number of contents related to various topics, such as music, movies, and books etc. The more intimate the relationship between the users, the wider is the range of topics exchanged among the users. In addition, the similarity of topics between users on SNSs is one of the main elements of a personalized Twitter search with the frequency providing the appropriate retrieval result in accordance with the individual's propensity [Vosecky14].

Therefore, we should consider the number of topics that the users share, which is different from previous research that measured only the item-trust, to provide high quality of recommendations in multi-domain. In this dissertation, the proposed preference similarity considers not only item-trust but also the similarity of topics, and we call it topic-affinity.

To measure topic-affinity, we compute the distribution of specific topic t in all of the users' contents as $\theta_{u,t}$. After the measurement of the distribution of topics for each user is completed, the topic-affinity between users u and f is calculated by using the Kullback-Leibler divergence [Kullback51] [Vosecky14], which can measure the difference between two distributions. We define the set of topics as T. Then, the topic-affinity is calculated as equation (12):

$$\begin{aligned} w_A(u, f) &= \frac{1}{KL(\theta_{u,t} \parallel \theta_{f,t}) + 1} \\ KL(\theta_{u,t} \parallel \theta_{f,t}) &= \sum_{t \in T} \theta_{u,t} \log \frac{\theta_{u,t}}{\theta_{f,t}} \end{aligned} \tag{12}$$

Preference similarity is calculated as a weighted sum of both the item-trust and the topic-affinity, which denote the similarity of the item and the topic mentioned by users; and it is defined as equation (13):

$$\begin{gathered} Sim_P(u, f) = \alpha_P w_T(u, f) + \beta_P w_A(u, f) \\ (\alpha_P + \beta_P = 1) \end{gathered} \tag{13}$$

The range of the values of the properties defined above is 0 to 1 for all six, but the mean, standard deviation, and distribution of numerical values of each type of property are different. Therefore, all properties of similarity, $w_n(u, f)$, are normalized as follows: $\{w_n(u, f) - \overline{X}\} / s$, where $\overline{X}$ and s represent the mean and the standard deviation of the similarities between a user u and f.

$w(u, f)$ denotes friendship strength and is calculated by a combination of the elements, such as $Sim_I(u, f)$, $Sim_G(u, f)$, and $Sim_P(u, f)$. When σ^T is a weighted vector that provides the difference in the weight value according to the importance of each element, $w(u, f)$ is calculated as equation (14):

$$w(u, f) = \sigma^T \begin{pmatrix} Sim_I(u, f) \\ Sim_G(u, f) \\ Sim_P(u, f) \end{pmatrix} \tag{14}$$

3.7 Personalized Recommender Algorithm (Step 10 in the Methodology)

The proposed personalized recommender algorithm based on friendship strength recommends items (i.e., interests) to users by considering their tendency. We calculate the friendship strength by applying various characteristics of social big data and use it as the similarity measure between users. Furthermore, the information of connected users who are linked in their social circle is more important than that of the not connected users. Therefore, we use only the information of directly or indirectly connected users.

The predicted preference value of user u for item i is defined as equation (15), where $\overline{r_u}$ and $\overline{r_f}$ denote the mean preferences of users u and f for all items, respectively.

$$p_{u,i} = \overline{r_u} + \frac{\sum_{f \in F_u^1 \cup F_u^n} (r_{f,i} - \overline{r_f}) w(u, f)}{\sum_{f \in F_u^1 \cup F_u^n} w(u, f)} \tag{15}$$

CHAPTER 4

Experiment and Evaluation

In this chapter, we describe our extensive experiments using the Twitter data set and the analysis of this data. Before beginning the experiment, we first verify the accuracy of the polarity analysis based on our opinion word dictionary. Then, through the results of the experiments, we validate the recommendation quality of our friendship strength-based recommender system to confirm whether it outperforms the baselines by using various metrics, such as MAE, precision, recall, F1-measure, and NDCG.

4.1 Data Set

In this dissertation, we used the data of 120 million crawled Korean contents and

160 million users on Twitter provided by DaumSoft for the period from July 1, 2012, to July 30, 2012. We briefly describe the provided raw data. The raw data is categorized as content related information, user relationship information, and opinion word dictionary for smartphones.

User-generated contents in our social big data set is divided into information about contents and the user information related to the contents. The former consists of the content document id, user-generated contents, parsed contents, the posting time of the contents, and the retweeted content document id, and the latter consists of the user id and mentioned user id.

The information of user relationship consists of a list of user ids identifying each user and sets of the user's followees. The provided opinion word dictionary consists of positive and negative words lists for smartphones to determine the polarity of contents related to smartphone.

In order to verify whether the recommendation performance is improved in multi-domain by using our friendship strength-based personalized recommender system, we extracted 933,499 contents related to the smartphone, music, movie, and drama domains from the provided data. Further, we used only the data of 6,318 active users who mentioned all four domains to ensure a precise experiment.

In the case of the smartphone domain, we chose 11 smartphone devices that were on the market during the study period. In the case of the music domain, 11 songs, which were ranked in the top 10 on South Korea's music streaming web sites, Bugs Music[6], were selected. In the case of the movie domain, we chose 4 Korean movies and 4 foreign movies screened in Korea during the study period. In the case of the drama domain, we chose 9 dramas that were aired by Korean terrestrial television companies, such as KBS, MBC, and SBS. All the data sets used in our experiments are shown in Table 5.

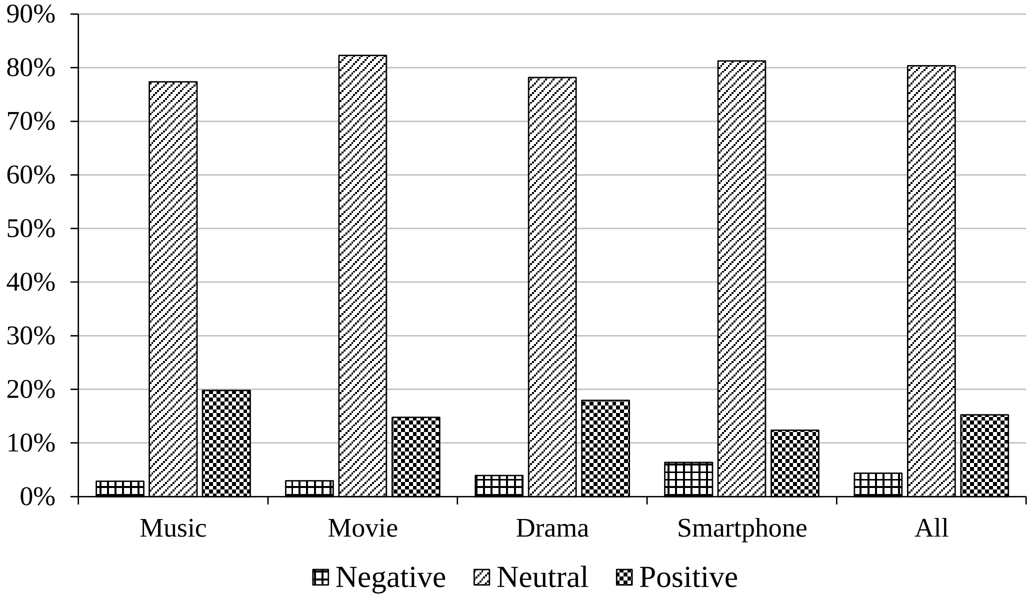
Table 5. Data set for experiment

	Topic			
	Music	**Movie**	**Drama**	**Smartphone**
#Item	11	8	9	11
#Contents	78,148	284,084	246,276	324,991
#Users	47,221	89,372	82,526	89,036
#Active users	6,318			
#All contents	933,499			
Time span	2012 - 07 - 01 ~ 2012 - 07 - 31			

6 http://www.bugs.co.kr

4.2 Data Analysis

In this chapter, before explaining the comparative experiment, we cover the analysis of the data set used in this dissertation shown in Table 5 (Section 4.1). We measure and analyze the four aspects derived from our data set: the distribution of contents' polarity, the distribution of ratings' polarity, the density of the users-to-users matrix, and the density of the users-to-items matrix. The analyzed results are used as the basis of the experimental results to be discussed in the following chapters.

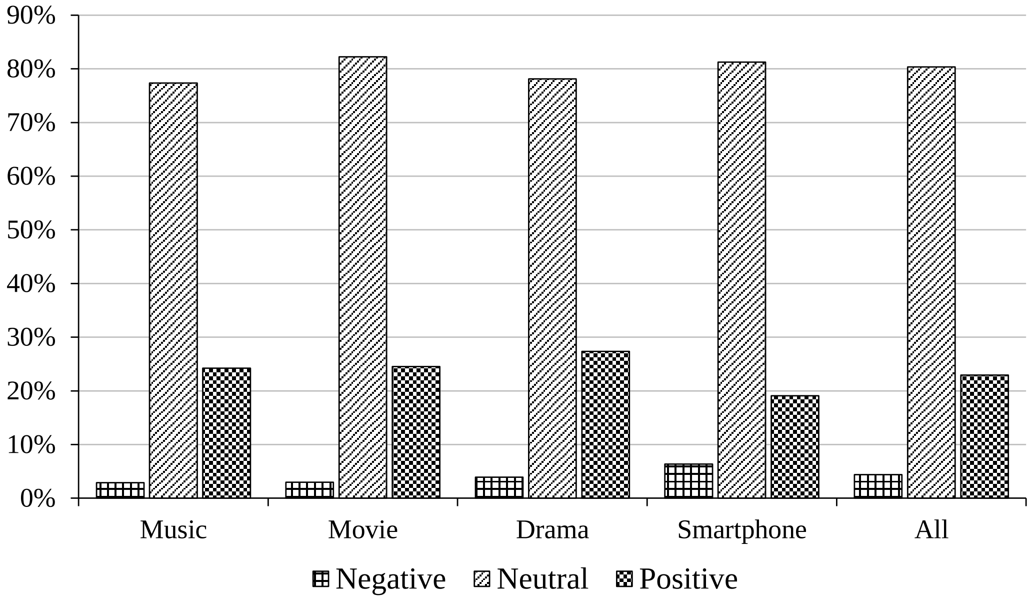


%	Music	Movie	Drama	Smartphone	All
Negative	2.87%	2.96%	3.92%	6.37%	4.37%
Neutral	77.36%	82.26%	78.15%	81.26%	80.37%
Positive	19.78%	14.78%	17.93%	12.37%	15.26%

Figure 7. The distribution of contents' polarity

First, we analyze the distribution of the contents' polarity for each domain as well as all domains. As shown in Figure 7, most content is neutral, which is about 80%. In contrast, the ratio of content with polarity is low, and especially, the ratio of negative is very low. There are two main reasons for this result. The first reason is that polarity analysis of short documents such as tweets on Twitter is difficult, and the second is that users on Twitter are aiming to transfer information rather than to communicate an emotional state. For each domain, the ratio of positive contents is high in music and drama domains, whereas the ratio of negative contents is high in the smartphone domain. In the case of the music domain, the proportion of positive contents increased immediately after some recorded music or albums were released. In the case of the drama domain which is similar to the music domain, a lot of positive contents is generated when the drama is airing on TV. This real-time generation of contents for items is a main reason for the high proportion of positive contents in music and drama domains rather than movie and smartphone domains. In the case of the smartphone domain, the ratio of negative contents is high rather than other domains. This is because smartphone technology at the time of the experimental data (2012) was not mature yet and users' likes and dislikes were clearly distinguished for each smartphone brand.

The distribution of ratings' polarity ratio is shown in Figure 8. Although the ratio of neutral ratings is higher than the polarity ratings, the ratio of polarity ratings is higher than that of contents. The ratios of both positive and negative ratings were twice as high as that of contents. The reason for this is that the ratings are measured by the average of the score of contents for the items. If there is only one content with polarity in all contents for items, it becomes a positive or negative ratings. However, it is still difficult to say that the rating with polarity is high, and renders it difficult to grasp the tendency of users compared to other users in public data.



	Music	Movie	Drama	Smartphone	All
Negative (0 ≤ r < 0.5)	5.01%	5.16%	5.74%	13.13%	8.49%
Neutral (r=0.5)	70.76%	70.33%	66.89%	67.79%	68.57%
Positive (0.5 < r ≤ 1)	24.24%	24.51%	27.36%	19.08%	22.94%

Figure 8. The distribution of ratings' polarity

Table 6. The density of the users-to-items matrix

	Music	**Movie**	**Drama**	**Smartphone**	**All**
Density	$1.2 \times e^{-1}$	$2.6 \times e^{-1}$	$2.4 \times e^{-1}$	$3.5 \times e^{-1}$	$2.4 \times e^{-1}$

Next, the density for ratings in users-to-items matrix is as shown in Table 6. According to the users-to-items matrix, we can compare the ratings between users and calculate the preference similarity. The density of all ratings in users-to-items matrix is $2.4 \times e^{-1}$. The density of each domain is related to the amount of contents shown in Table 5 (Section 4.1). The density of the smartphone domain is the highest for all domains, because the amount of content is the largest. Conversely, the amount of content in the music domain is the smallest, hence its density is also the lowest. The density of the movie and drama domains are not much different, but the movie domain density is slightly higher than that the drama because it has slightly more content.

We also looked at the ratio of user ratings for each item as shown in Figure 9. In the case of drama and smartphone domains, the deviation among the ratio of ratings for each item is low. In the case of the music and movie domains, however, the deviation of the ratio of ratings is high. The density and distribution of the ratings for items are affected by issues at the time of the data collection. Therefore, according to the issues, the density and distribution can be changed.

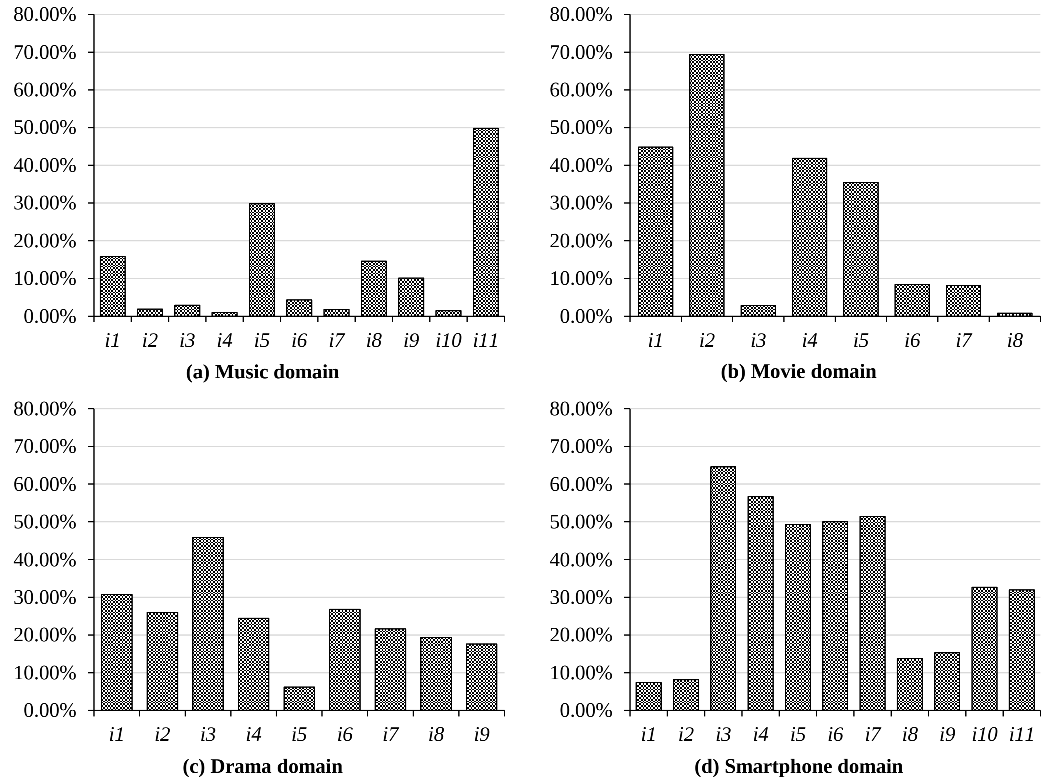


Figure 9. The ratio of each item for all domains

Finally, we measured the density of the users-to-users matrix for all similarity in this dissertation. As shown in Table 7, we found that the density of the interaction similarity was smaller than the density of the preference and the group similarities The low density of interaction similarity might have a negative impact on the performance of proposed recommender system. On the contrary, if there is an interaction similarity between users, they can be regarded as in a more intimate relationship. Therefore, from this point of view, interaction similarity can provide better performance. In the case of group similarity, we consider the group for each user as all users within 6 edges, because the density of the group similarity when only considering one edge is too low. [Kwak10] proved that almost all relationships between users are within 6 edges, therefore, we did not consider the groups with over 6 edges. Preference similarity can be measured in almost all user relationships, which is similar to other public data used in recommender systems.

Table 7. The density of the users-to-users matrix

Interaction Similarity			**Group Similarity**		**Preference Similarity**	
Frequency	**Recency**	**Longevity**	**n = 1**	**n ≤ 6**	**Item Trust**	**Topic Affinity**
$2.9 \times e^{-4}$	$2.9 \times e^{-4}$	$2.9 \times e^{-4}$	$2.5 \times e^{-3}$	$6.5 \times e^{-1}$	$9.6 \times e^{-1}$	1

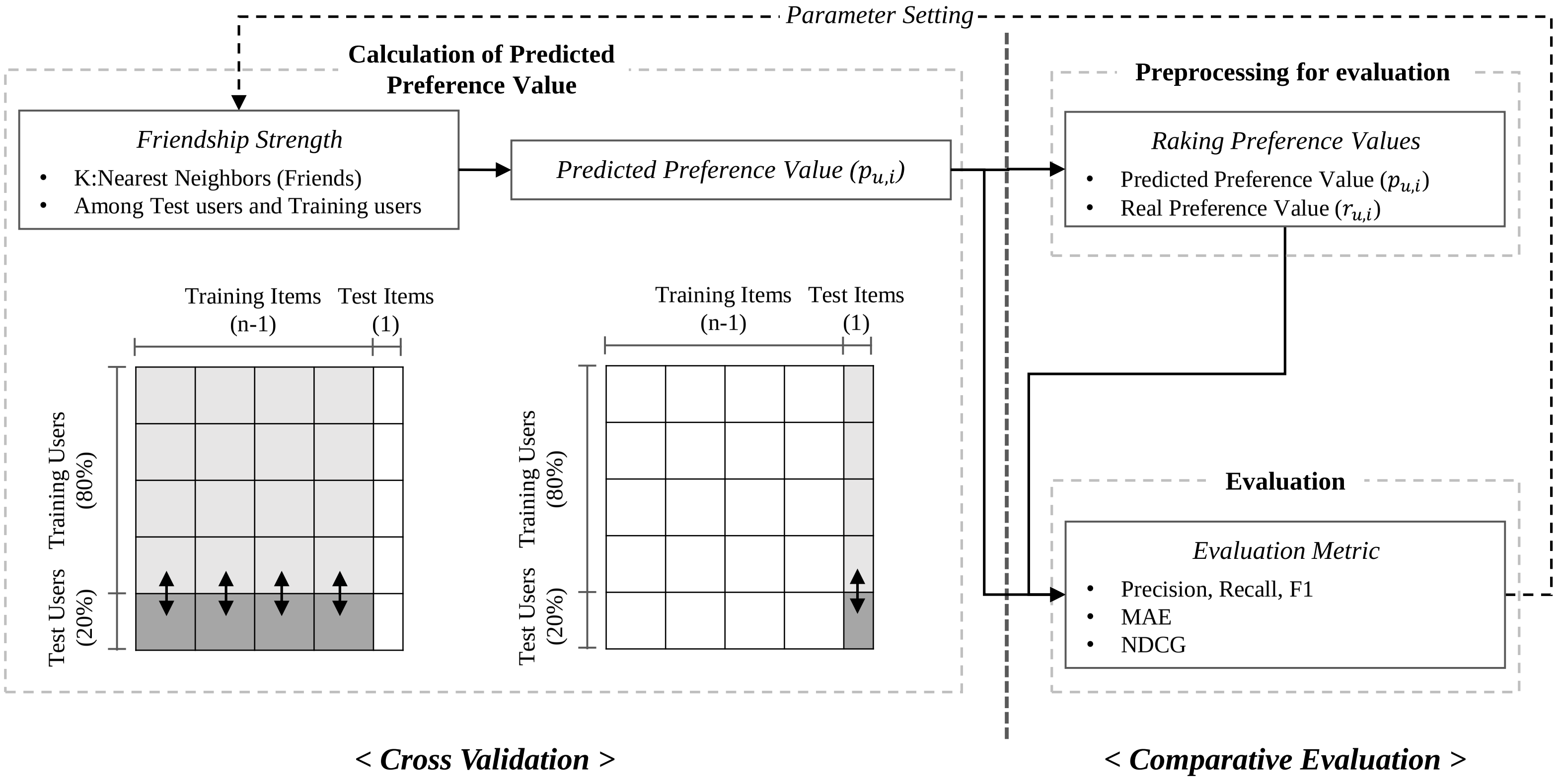


Figure 10. Evaluation framework.

4.3 Evaluation Process

In this dissertation, we conducted a cross validation based on the processing data provided by DaumSoft to evaluate the proposed personalized recommender system. The framework of our evaluation is shown in Figure 10.

In the validation step, we conducted a cross validation for the experiment based on data processing. First, we divided the data into a test set and a training set. In the case of the set of users, we considered 20% of the active users as the test users and 80% as the training users. After building a users-to-users friendship strength matrix for the experiment, we found that the density of the interaction similarity was smaller than the density of the preference and the group similarities as shown in Table 7 (Section 4.2). Therefore, we considered 20% of the active users who interact actively with other users to be test users in order to reflect the influence of interaction similarity.

In the case of the set of items, it was not suitable to use n-fold cross validation as for the set of users. In our data set, there were 39 items used in the experiments; however, the mean of the number of the items mentioned by active users was roughly 10. Therefore, we used the leave-one-out cross validation for the set of items. After the separation of the test set and the training set, we calculated the friendship strength for the test users. The interaction and the group similarity were measured by using the users-to-users friendship strength matrix, and the preference similarity was calculated using the preference value and the topic distribution value of training items between test users and training users.

In the evaluation step, we validated the superiority of the proposed friendship strength-based personalized recommender system using a comparative evaluation. First, we determined the friendship strength of the k-nearest neighbors (i.e., k-friends close to the test users) and then, calculated the predicted preference value ($p_{u,i}$) of test users according to the proposed system based on the information of k-nearest neighbors. Second, we determined the weighted value of three friendship strength elements (α_I , β_I , χ_I , α_G , β_G , α_P , and β_P) and the weighted vector (σ^T) by adjusting it to use the predicted preference value. Finally, we conducted a comparative evaluation to verify the performance of the proposed system using various metrics such as MAE, precision, recall, F1-measure, and NDCG.

4.4 Accuracy of Polarity Analysis

Before measuring the effectiveness of the proposed recommender system, we verify the accuracy of polarity analysis based on our opinion word dictionary. The accuracy is measured by the oracle test through the participation of our lab members. We randomly extract 200 ~ 400 contents used by our experiment data set for each domain, and then check whether the polarity is properly classified as true or false mark. The accuracy is measured by the ratio of true mark, and the result is shown in Table 8.

Table 8. The accuracy of polarity analysis based on oracle test

(%)	User a	User b	User c	User d	Avg.
Smartphone	68.06	70.37	78.67	77.03	73.53
Movie	81.72	84.69	78.61	77.58	80.65
Music	71.52	72.20	82.11	82.35	77.05
Drama	78.70	81.40	80.51	79.82	80.11
Avg.	**75.00**	**77.17**	**79.98**	**79.20**	**77.83**

Polarity analysis usually utilize the opinion word dictionary or the classification algorithms. The famous English opinion word dictionaries are MPQA [Stoyanov05] and SentiWordNet [Esuli07]. For Korean, KOSAC [Shin16] of Seoul national university is famous. In case of the methods using the classification algorithm, they traditionally utilize the SVM (support vector machine), and Naïve Bayes [Pang04]. Recently, polarity analysis for English has a high accuracy which reach the approximately 95% accuracy based on convolutional neural network [Kim14]. In the case of Korean polarity analysis, it is much harder than English because morphological analysis in Korean is difficult [Jang13]. Therefore, Korean polarity analysis is generally less accurate than English.

In this dissertation, we actually do not focus on polarity analysis, but verify that the accuracy is approximately 78% according to oracle test. Furthermore, previous study shows that if the accuracy of the polarity analysis method is higher than 70%, that method is similar to polarity analysis by person [Ogneva10]. Therefore, the accuracy of our Korean polarity analysis is reasonable in this dissertation.

4.5 Decision of the Combination of Friendship Strength Elements

We measured the similarity between users through the combination of three friendship strength elements: interaction similarity (I), group similarity (G), and preference similarity (P). There are seven combinations, namely I, G, P, IG, IP, GP, and IGP, and we need to pre-determine which combination exhibits the best performance. Therefore, before the comparative evaluation with baselines, we conducted an experiment to find the combinations with the highest performance by using precision, recall, and F1-measure shown in Figure 11.

When we used I, G and P as a single element of friendship strength, the performance of I was inferior to the performance of G or P. This is because the users-to-users friendship strength matrix density of I is less than that of the other elements. However, we can confirm that I has a greater effect on the performance improvement when used together with G and P than when used alone. From Figure 11, we can infer that G and P show a similar performance, which is better than the performance of I when used as a single element of the friendship strength; however, when they are used in combination with the other elements, their effect is less than that of I. Among all the combinations, IP shows the best performance; its performance is better than the performance of the combination of all elements (IGP). Therefore, in this dissertation, we set IP as the friendship strength.

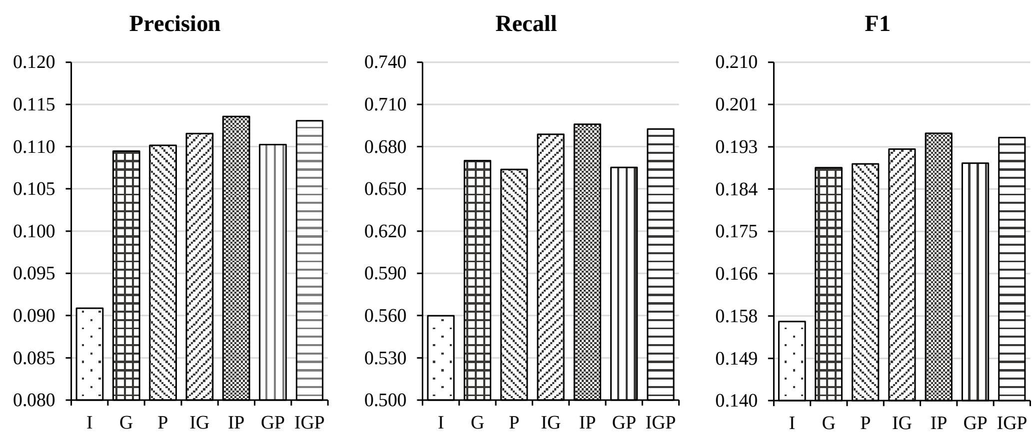


Figure 11. Results of the combination of friendship strength elements.

4.6 Quantitative Evaluation

To prove the quality of the recommendation generated by the proposed friendship strength-based personalized recommender system, we evaluated the performance of the proposed system and compared it with that of baselines. The baselines used for the comparative evaluation were CF based on PCC and JMSD.

CF has been used widely for personalized recommender system, both academically and commercially. In practice, many web services such as Amazon, Reddit, and YouTube, are based on CF [Saleem08] [Ekstrand11]. Furthermore, SNS such as Last.fm, Facebook, and LinkedIn use CF to recommend items or friends [Ekstrand11] [Victor11].

We chose PCC for a baseline because it is the most popular with CF-based personalized recommender systems [Herlocker99] [Liu13] [Yin15]. PCC was originally proposed in 1999 [Herlocker99], but it has been widely used in recent studies. [Konstas10] measured the similarity of play count, tag, and friendship relationship based on PCC to use the baseline methods. [Liu13] measured the similarity based on the contextual information of users by extending the PCC. [Yin14;15] utilized the PCC as an element of measuring similarity. Furthermore, many actual recommender systems are based on PCC. For instance, Ringo (music), BellCore (movie), and FilmTrust (movie) are all based on PCC [Golbeck06] [Ekstrand11].

JMSD is chosen because it can consider the ratio of common interest. Therefore, JMSD is more suitable for Twitter-based recommender systems than other numerical value-based similarity measures. JMSD-based recommender systems do not exist in practice, but their effectiveness has been proven using real recommender system databases, such as Movielens, FilmAffinity, and Netflix [Bobadilla10;12].

The proposed system utilizes the information of users who are connected directly or indirectly on the SNS, whereas the baselines provide recommendations for users using the data of unspecified individuals. Thus, we needed to verify whether the information of connected users is helpful in the case of using PCC and JMSD. To prove this, we measured the performance of the baselines based on PCC and JMSD by using the information of connected users (S_PCC and S_JMSD), and compared it with that of the baselines using the information of unspecified individuals and the proposed system.

We validated the quality of recommendation made by the proposed system using RMSE, precision, recall, F1-measure, and NDCG which are the main metrics used for measuring the performance of the personalized recommender systems.

4.6.1 Accuracy of Predicted Preference Value

First, we calculated the difference between the predicted preference value and the actual user's preference value to evaluate the accuracy of the predicted value generated by the personalized recommender systems [Bobadilla13]. MAE is mainly used for the error measurement of predicted values and is defined as equation (16):

$$MAE = \frac{1}{|I|}\sum_{i \in I}\left\{\frac{1}{|O_{u,i}|}\sum_{u \in O_{u,i}}\left|p_{u,i} - r_{u,i}\right|\right\}$$
$$O_{u,i} = \{u \in U \mid p_{u,i} \neq null \wedge r_{u,i} \neq null\} \quad (16)$$

$O_{u,i}$ denotes the set of users, the $p_{u,i}$ and $r_{u,i}$ of whom are both not null values in the test users' set U. The error of the $p_{u,i}$ is calculated as the absolute value of the difference between $p_{u,i}$ and $r_{u,i}$. The results of MAE for the comparative evaluation are shown in Figure 12; in this case, the range of k-neighbors (K) is 100 to 2,000.

As shown in Figure 12, the proposed friendship strength-based system has a lower error rate than PCC and JMSD. The Figure 12 shows that the performance of the predicted preference value's accuracy is significantly improved by 15 to 20% and 13 to 19% as compared to that of the PCC and S_PCC, respectively, in the overall range.

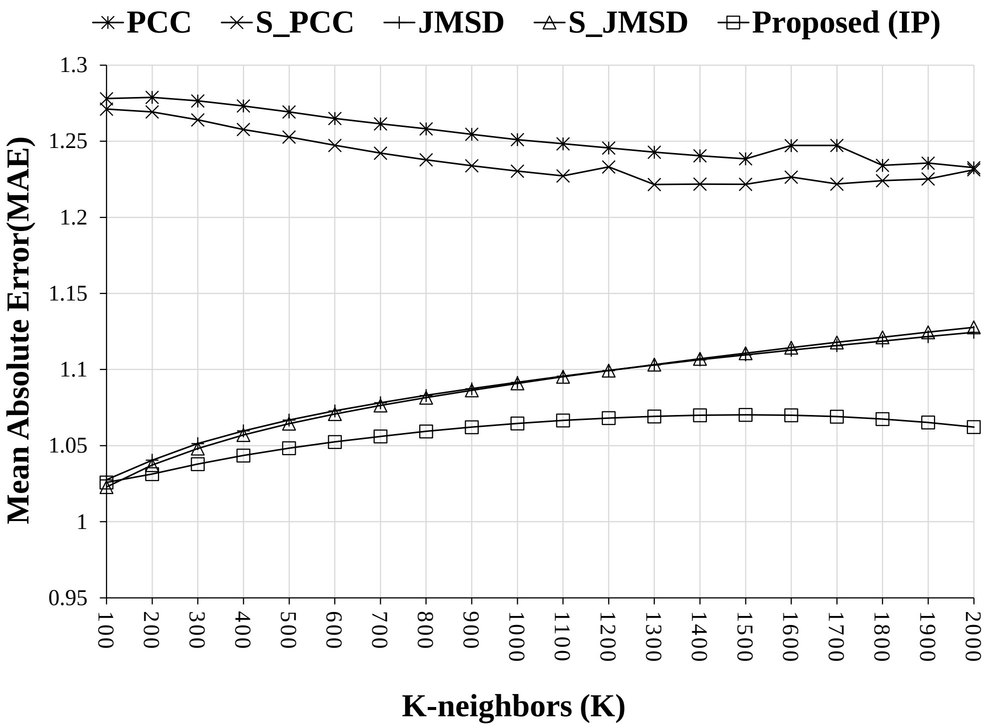


Figure 12. Comparative evaluation to measure predicted value's quality

As compared with the JMSD and S_JMSD, its performance is nearly the same or slightly better until the number of K reaches 1,000. However, when the number of neighbors is greater than 1,000, the performance is improved by approximately 5%. Further, the larger the number of K, the higher is the performance improvement.

The comparison results for PCC and JMSD show that the accuracy of the JMSD is higher, but the range of performance improvement decreases with the increase in the number of K.

Finally, we find that the information of connected users affects the accuracy of the predicted preference value more than that of unspecified users in the existing JMSD- and PCC-based personalized recommender systems. S_PCC and S_JMSD which use the information of the connected users exhibit better quality than PCC and JMSD, even when S_JMSD is almost similar to JMSD, but slightly better as a whole.

4.6.2 Quality of Recommendation List Set

Users' reliance on the recommendation results received through personalized recommendations is not determined by the accuracy of the predicted values. MAE may be a good metric to measure the performance of the recommender algorithms. However, it cannot measure the satisfaction with the recommendation results that the users feel. Users trust results (i.e., recommended item lists) obtained through personalized recommender systems if they are in fact satisfied with these results. Therefore, it is necessary to measure the users' satisfaction with the recommended results; This is calculated by determining whether the result is relevant to the user or not.

Precision, recall, and F1-measure are representatively used for measuring the effectiveness of the personalized recommendation's results [Bobadilla13]. They are defined as equations (17), (18) and (19), where Z_u denotes a set of N recommended items provided by users and ϑ represents the threshold value.

$$precision = \frac{1}{|U|}\sum_{u \in U}\frac{\left|\{i \in Z_u \mid r_{u,i} \geq \vartheta\}\right|}{|N|} \tag{17}$$

$$recall = \frac{1}{|U|}\sum_{u \in U}\frac{\left|\{i \in Z_u \mid r_{u,i} \geq \vartheta\}\right|}{\left|\{i \in I \mid r_{u,i} \geq \vartheta\}\right|} \tag{18}$$

$$F1 = \frac{2 \times precision \times recall}{precision + recall} \tag{19}$$

First, to measure the precision, recall and F1-measure, we not only ranked the lists of the recommended items according to the high predicted preference value but also determined the threshold for judging the relevance of the recommended items. In this dissertation, the $r_{u,i}$ is defined as from 0 to 1. If $r_{u,i}$ is higher than 0.5, user u give a positive opinion about item i more than once in all his/her contents. Therefore, we set the threshold value as 0.5, because it can be judged that recommended item i is relevant if $r_{u,i}$ is higher than 0.5.

As shown in Figure 13, in terms of the metrics for the set of recommendations, such as precision, recall and F1-measure, the proposed system achieves a higher value than the baselines. In terms of precision, the proposed system exhibits an improvement of about 7% and 5% as compared to PCC and

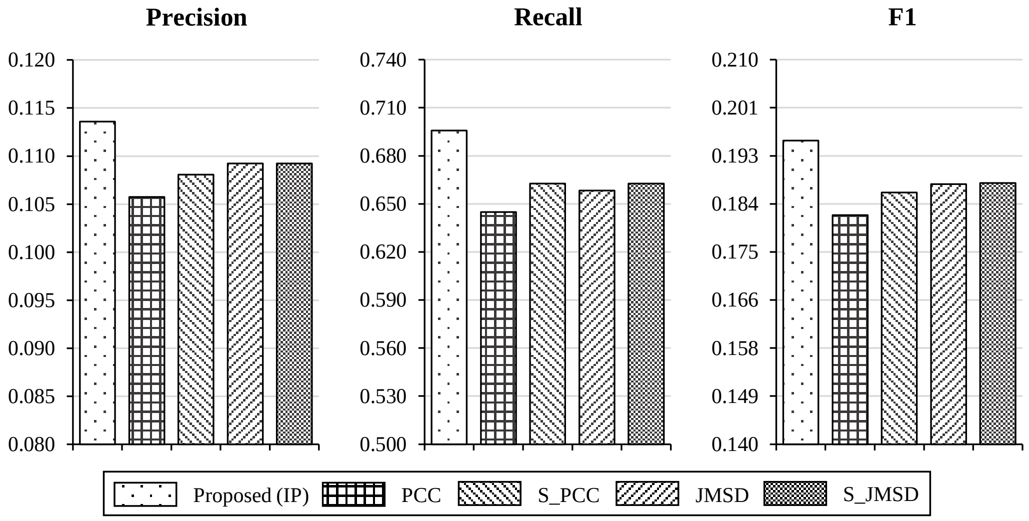


Figure 13. Comparative evaluation to measure the quality of the set of recommendations

S_PCC, respectively. The system is also shown to enhance performance by approximately 4% as compared to JMSD and S_JMSD.

In terms of recall, the proposed system exhibits a performance improvement of about 8% as compared to PCC and of about 6% as compared to JMSD. As compared with S_PCC and S_JMSD, it exhibits a performance improvement of about 5%.

For the F1-measure, the results are similar to those obtained for precision. The proposed system exhibits a performance that is approximately 7% better than that of PCC, 5% than that of S_PCC, and 4% than that of JMSD and S_JMSD.

4.6.3 Quality of Ranked Recommendation List

In personalized recommender systems, we consider mainly the relevance of high ranked items because users tend to look only at the top ranked results among all the recommended items to find the relevant items [Baltrunas10]. Therefore, if N items are recommended to users, the first recommended item has the highest importance. Furthermore, when the high ranked items are incorrect or not relevant to users, a more serious error is generated than when the low ranked items are incorrect or not relevant.

On the basis of these assumptions, DCG and Ideal DCG (IDCG) are calculated according to equations (20) and (21). NDCG [Baltrunas10] is the value of DCG divided by IDCG, as defined in equation (22).

$$DCG = \frac{1}{|U|} \sum_{u \in U} \left\{ r_{u,p_1} + \sum_{i=2}^{k} \frac{r_{u,p_i}}{\log_2 i} \right\} \tag{20}$$

$$IDCG = \max(DCG) \tag{21}$$

$$NDCG = \frac{DCG}{IDCG} \tag{22}$$

$p_1, \ldots, p_n$ denotes the list of ranked items according to the predicted preference value provided by the users. r_{u,p_i} represents the actual preference value of the p_i th item given by user u. In this dissertation, we conducted a comparative evaluation using NDCG when k, which is the number of all recommended items, was set at 5, 10, or 20. Table 9 shows the results of NDCG for the proposed system and the baselines.

Table 9. Comparative evaluation to measure the quality of ranked list

	NDCG@k		
	k=5	**k=10**	**k=20**
PCC	0.7978	0.8588	0.8779
S_PCC	0.7963	0.8578	0.8771
JMSD	0.7963	0.8576	0.8736
S_JMSD	0.7986	0.8584	0.8752
Proposed (IP)	**0.8039**	**0.8607**	**0.8784**

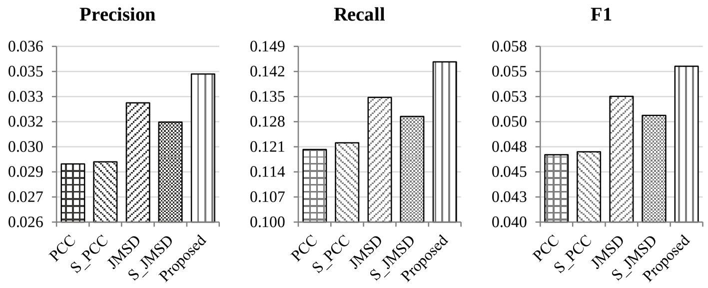
As for the measurement of the quality of the ranked results as shown in Table 9, the differences in NDCG between the proposed system and the baselines is not very large. Nevertheless, in all cases, when k is 5, 10, or 20, the proposed system's NDCGs are higher than that of the baselines.

4.6.4 Quality for Each Domain

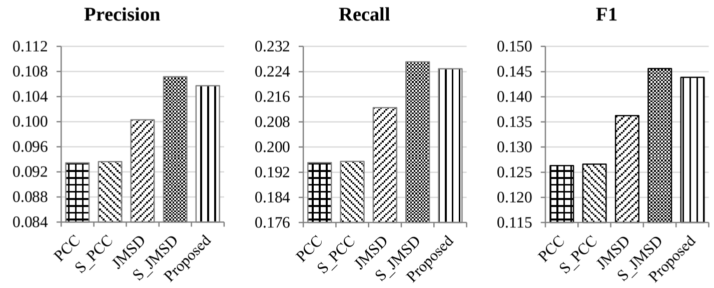
We also conducted the experiments for each domain to validate that our proposed method has a higher performance in a multi-domain environment. As shown in Figure 14, the friendship strength-based recommender system outperforms the baselines in all domains, except in the movie domain.

In the case of the music domain, the amount of contents is one-third that of the other domains as shown in Table 5 (Section 4.1). The smartphone domain, on the other hand, has the most content, when compared to the other domains. Further, the density of the users-to-items matrix of the music domain is the lowest, and the density of the smartphone domain is the highest as shown in Table 6 (Section 4.2). Because of the density of the matrix and the amount of contents, the music domain has the lowest effectiveness and smartphone domain has the highest in all domains.

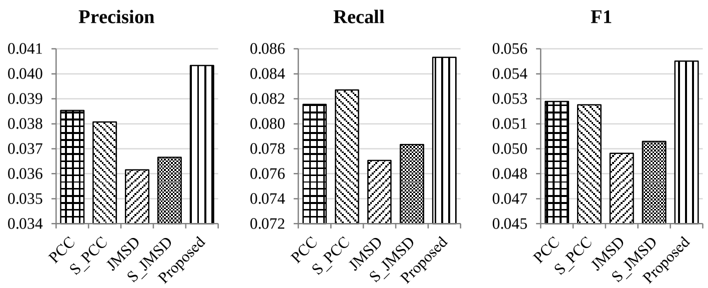
In the case of the movie domain, items having many ratings or few ratings are clearly divided, as shown in Figure 9 (Section 4.2). The effectiveness of JMSD is measured to be relatively higher than other baselines, including our method, in the movie domain. This is because the higher the number of ratings for items, the higher the likelihood of liking the items at the same time. Our friendship strength-based method can also affect the movie domain significantly, because JMSD is an element of friendship strength. In the movie domain, we verify that preference similarity is of a higher impact than interaction similarity.



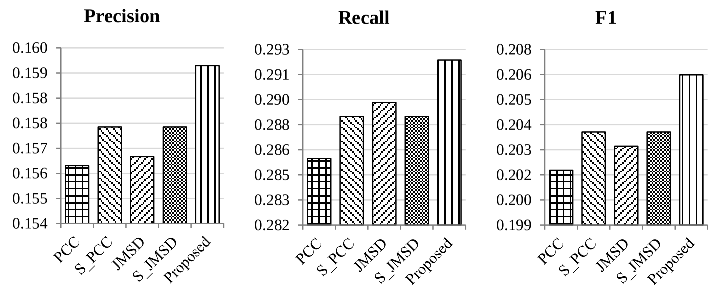
(a) Music domain



(b) Movie domain



(c) Drama domain



(d) Smartphone domain

Figure 14. Comparative evaluation to measure the effectiveness for each domain

In the case of the drama domain, the ratings' ratio for each item is distributed uniformly, but there are a few items having many ratings. As a result, JMSD is measured to be relatively low in the drama domain. However, the performance of the proposed IP is the highest because users share a lot of information about drama.

We evaluated our friendship strength-based personalized recommender system by using various metrics: MAE, precision, recall, F1-measure, and NDCG. The results show that the proposed system exhibits a better performance than the baselines in terms of all the metrics. Furthermore, we verified that our proposed recommender system has higher effectiveness for each domain rather than baselines.

The proposed system based on the friendship strength provides recommendations for users by using the information of other users who are closely connected with them on the SNS; therefore, its recommendations are more valuable for users than those of the baselines based on PCC and JMSD, which use the information of unspecified users. Further, the performance of S_PCC and S_JMSD is better than that of PCC and JMSD, respectively. That is, the use of the information of connected users' results in better recommendations than does that of the information of not connected users. Consequently, we verified that it is very efficient to use the information of connected users in personalized recommender systems based on people relationship information.

Finally, the performance of proposed system is superior to S_PCC and S_JMSD. This result reveals that the measurement of the proposed friendship strength appropriately represents the closeness between users with respect to their relationship and the proposed personalized recommender system based on friendship strength is useful for improving the accuracy of recommendations.

4.7 Qualitative Evaluation

Finally, we compared the proposed system with previous personalized recommender systems, based on the method of using social big data shown in Tables 10 to 13. We mainly categorized recommender systems based on whether they used the people relationship information or not, as described in Section 2.

Calculating the item similarity between users based on social big data appropriately is a prerequisite condition for all personalized recommender systems, especially based on unspecified individuals' information. As shown in Table 10, [Firan07] [Li08] [Bobadilla10] [Liu14a] [Liu14b] [Servajean14] [Zhu14] [Yin14] [Yin15] only considered item similarity, and their systems are based on the information of unspecified individuals highly similar to the target users. Furthermore, there are no distinction between topic and item layers in their studies. [Yin14] [Yin15] defined topic as a higher level of item, but they did not exist topic similarity. Finally, they did not consider the connections between users in a social circle, but formed the clusters based on a set of unspecified individuals with similar interests. A few studies [Zhu14] [Yin14] [Yin15] considered the information of an entire social circle, but this is not the information between connected users.

Most social big data-based personalized recommender systems consider the connectivity among users on SNSs (i.e., group information). However, as shown in Table 11, influential-based recommendations [Zhen09] [Tang08] [Tang12] [Lin14] only find the expert users in an entire network. They did not consider the connection or relationship between two users.

The information of friends is important for a social big data-based personalized recommender system; furthermore, the measurement of friendship strength between users is a key factor. As shown in Table 12, a few studies [Geyer08] [Xu13] just use the connection relationship between users without considering the distance between them; in other words, they did not consider friendship strength. As shown in Table 13, group information was mainly used to calculate the friendship strength in many studies [Golbeck06] [Konstas09] [Guy10] [Qian13] [Ma14]. They assigned a weight to intimate friends based on their measurement of the group similarity. For example, they calculated group similarity to use RWR [Konstas09], or called it familiarity relationship score [Guy10], trust relationship [Golbeck06] [Qian13] [Ma14], or interpersonal influence [Qian13].

Table 10. The usage of social big data on personalized recommender systems based on information of unspecified individuals

	Interaction information			Group information	Personal information		Data set (Experiment domain)
	Frequency	Recency	Longevity		Item similarity measure	Topic similarity measure	
[Firan07]	Not support	Not support	Not support	Not support	Cosine	Not support	Last.fm (Music)
[Li08]	Not support	Not support	Not support	User clusters for topics	Similarity of intra- and inter-topics (based on cosine)	Not support (do not exist higher level of topic)	Delicious (URLs)
[Babadilla10]	Not support	Not support	Not support	Not Support	JMSD	Not support	MovieLens (Movie), NetFlix (Movie), FilmAffinity (Movie)
[Liu14a]	Not support	Not support	Not support	Not Support	New Heuristic Similarity Model (NHSM)	Not support	MovieLens (Movie), Epinions (Multi-domain)
[Liu14b]	Not support	Not support	Not support	Not support	Gaussian kernel, Cosine	Not support	Flickr (Geo-Specific Tag)
[Servajean14]	Not support	Not support	Not support	Not support	Usefulness score (based on Jaccard)	Not support	MovieLens (Movie) Flickr (Photo) Last.fm (Music)
[Zhu14]	Popularity of items (in a whole network)	Not support	Not support	Not support	Cosine	Not support	MovieLens (Movie)
[Yin14] [Yin15]	Not support	Temporal context (recent interests in a whole network)	Not support	Not support	Vector Space Similarity (VSS), Jaccard, PCC	Not support (topic is higher level of item, but do not exist topic similarity)	Digg (News) MovieLens (Movie) Douban Movie (Movie) Delicious (URLs)

Table 11. The usage of social big data on personalized recommender systems based on information of the influentials

	Interaction information			Group information	Personal information		Data set (Experiment domain)
	Frequency	Recency	Longevity		Item similarity measure	Topic similarity measure	
[Zhen09]	Not support	Not support	Not support	Collaborative team (a set of influentials)	Relationship Similarity Coefficient (RSC), Influence Coefficient (InfC)	Not support	The environment of a manufacturing enterprise (Enterprise knowledge)
[Tang08] [Tang12]	Not support	Not support	Not support	Random Walk with Restarts (RWR)	Cross-domain Topic Learning (CTL)	Not support	Aminer (Academic items)
[Bhattacharya14]	Not support	Not support	Not support	Topic experts (based on list meta-data)	Interest vector	Not support	Who Likes What (Multi-domain)
[Lin14]	Not support	Not support	Not support	Expert model (a way to find implicit influentials)	Cosine	Not support	Google News (News)

Table 12. The usage of social big data on personalized recommender systems based on information of the simply connected users

	Interaction information			Group information	Personal information		Data set (Experiment domain)
	Frequency	Recency	Longevity		Item similarity measure	Topic similarity measure	
[Geyer08]	Not support	Not support	Not support	Relationship strength (based on binary score)	Candidate relevance score	Not support (do not exist higher level of topic)	About You (Topics for user profile)
[Xu13]	Not support	Decay factor (weight of posting time)	Not support	Followee Influence (based on PageRank)	Not support (item is lower level of topic, but do not exist item similarity)	Cosine	Sina Weibo (Multi-domain)

Table 13. The usage of social big data on personalized recommender systems based on information of the friends

	Interaction information			Group information	Personal information		Data set (Experiment domain)
	Frequency	**Recency**	**Longevity**		**Item similarity measure**	**Topic similarity measure**	
[Golbeck06]	Not support	Not support	Not support	Trust value	PCC	Not support	FilmTrust (Movie)
[Konstas09]	Not support	Not support	Not support	RWR	RWR	Not support	Last.fm (Music)
[Guy10]	Not support	Not support	Not support	Familiarity relationship score	Similarity score (based on Jaccard)	Not support	SaND (Social media items)
[Lai13]	Social influence (the ratio of the sharing favorite photo between friends)	Time factor (weight of posting time)	Not support	Popularity influence (the total count of photo in a whole network)	Interest influence (interest similarity)	Not support	Flickr (Photo)
[Qian13]	Not support	Not support	Not support	Interpersonal influence, Trust value	Interpersonal interest similarity (Second level of category, based on cosine)	Interpersonal interest similarity (First level of category, based on cosine)	Yelp (Restaurant) MovieLens (Movie) Douban Movie (Movie)
[Yu13]	Users popularity (based on cosine)	Decay factor (weight of posting time)	Not support	Not support (but using direct friends info.)	Not support	Distance similarity measure (do not exist lower level of topic)	Sina Weibo (Multi-domain)
[Ma14]	Not support	Not support	Not support	Trust relations	Topic similarity (based on cosine)	Not support (do not exist higher level of topic)	Sina Weibo (Multi-domain)
[Seo17]	Frequency	Recency	Longevity	Intimacy (based on Jaccard)	Item-trust (based on JMSD)	Topic-affinity (based on KL-divergence)	MovieLens (Movie) Last.fm (Music) Twitter (Multi-domain)

However, there were few friendship strength measurements to use interaction information as shown in Table 13. [Yu13] and [Lai13] considered frequency for calculating friendship strength, but most studies using interaction information simply considered the interests of an entire social circle [Yin14] [Yin15] [Zhu14], or weighted recently posted items [Xu13] [Yu13]. Our system is different from other studies in that we considered the interaction information between users to calculate the friendship strength divided by frequency, recency, and longevity.

Further, most works did not calculate the similarity of higher level of items (i.e., topics), except one [Qian13] as shown in Tables 10 to 13. Therefore, most existing studies validated their methods using special purpose SNSs, such as movie and music. Few works [Yu13] [Liu14a] [Bhattacharya14] were conducted on the multi-domain environment. Some research [Geyer08] [Li08] [Yu13] [Ma14] measured topic similarity, but their methods did not consider a higher or lower level of topic. A few studies [Qian13] [Yin14] [Yin15] classified items and topic levels. However, [Yin14] [Yin15] did not consider the topic similarity, and [Qian13] only calculated both item and topic similarity. As shown in Table 13, we defined the higher level of items as a topic, and calculate both item and topic similarity between users.

The comparison of the usage of social big data on personalized recommender systems shown in Tables 10 to 13 shows that we considered various factors of social big data to calculate friendship strength as compared to others. Therefore, our measurement is a more appropriate measurement of friendship strength between users for personalized recommender system rather than existing research.

CHAPTER 5

Discussion

In this chapter, we briefly address the limitations of the experimental results and the applicability of the friendship strength in all recommender systems. Subsequently, in order to prove that they are not limitations in this dissertation, we explain the importance of the experimental results and the suitability of friendship strength for other social big data sets.

5.1 Limitation

In this dissertation, we have two limitations in terms of the experimental results and the applicability of friendship strength in all recommender systems. First, although we evaluated our proposed system to utilize various metrics, such as MAE, precision, recall, F1-measure, and NDCG, and the improvement in our proposed system is not significant, except that the MAE of the proposed system is higher by 15 to 20% and by 13 to 19% from that of PCC and S_PCC, respectively. Furthermore, strictly speaking, friendship strength [Seo17] is optimized for Twitter data set. It is because Twitter data set has various data to calculate all elements of friendship strength.

5.2 The Importance of the Experimental Results

The differences in the quality of recommendation are not significant due to our experimental environment. In our experiments, the number of item is smaller than other public data sets. This environment brings about small improvement in performance, because the recommended items for all recommender systems including those for the proposed system are not clearly different.

Although the differences of all metrics are not clearly different, it is difficult to improve all of them having different characteristics. Most recommender system studies validated the quality and the accuracy of their proposed system with one or two metrics. Therefore, it is a noteworthy result that our proposed system is the highest in all metrics.

5.3 The Applicability of Friendship Strength

From the viewpoint of whether friendship strength applies to other recommender systems based on social big data, our proposed friendship strength can be utilized in any system provided that interaction data among users, friend list information or items having several levels are available.

Table 14. The applicability of friendship strength in other recommender systems

Elements / Public data	I	G	P
Yelp	○	○	○
Delicious	○	×	○
Last.fm	×	○	○
MovieLens	×	×	○

Table 14 represents the applicability of friendship strength in other recommender systems. For example, Delicious has contact information among users, and Yelp has review information among users. These contact and review information can be leveraged to calculate the interaction similarity of our friendship strength. In addition, group similarity is also calculated in all recommender systems having a friend list or connection relationship, such as Last.fm and Yelp. In addition, we can calculate preference similarity via almost all public data sets. For example, in MovieLens, we can measure item-trust and topic-affinity to use movie (i.e., item) and genre (higher level of movies, i.e., topic) data.

Most recommender systems based on social big data have at least two characteristics to calculate friendship strength. Therefore, friendship strength can be used by any system.

CHAPTER 6

Conclusion and Future Work

In this chapter, we briefly summarize the entire content and contribution of this dissertation as a conclusion. Then, we describe the future work to be conducted as the next task of this dissertation.

6.1 Conclusion

In this dissertation, we proposed a friendship strength-based personalized recommender system. The proposed friendship strength considers various characteristics of social big data in order to measure the closeness between users on SNSs. Our personalized recommender system grants a weight to those users who are closely connected in their social circle based on friendship strength in order to recommend the suitable topics or interests in which users might be interested.

We conducted comparative experiments using one month's Twitter data, which is multi-domain SNS and verified the proposed algorithm using various metrics: precision, recall, F1-measure, MAE, and NDCG. The experimental results verified that the use of the information of connected users on the SNS is better than that of the information of unspecified users for social big data-based personalized recommendations. Further, and more importantly, the proposed friendship strength determines the degree of closeness between users appropriately and helps to improve personalized recommendations in a multi-domain environment as compared to other similarity measures.

6.2 Future Work

In future work, we will intend to validate that the proposed friendship strength is effective in various personalized services such as personalized retrieval, micro blog search, and semantic web.

In this dissertation, we focus mainly on the memory-based CF method. Therefore, we will apply friendship strength to the model-based CF method and validate its effectiveness by leveraging PMF (probabilistic matrix factorization) and deep neural network (i.e., deep learning).

Finally, we intended to conduct a study focusing on the accuracy of the personalized recommendations through the pre-crawled data provided by DaumSoft. Therefore, we will need to optimize the computation time of recommendations to use a filtering method such as fp-growth filtering, which allows for the efficient processing of large amounts of social big data. By using this method, we can develop personalized recommender systems to handle this data in real time.

BIBLIOGRAPHY

[Adomavicius05] Adomavicius, G., & Tuzhilin, A. (2005). 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, 17(6), 734-749.

[Baltrunas10] Baltrunas, L., Makcinskas, T., & Ricci, F. (2010). Group recommendations with rank aggregation and collaborative filtering. In Proceedings of the 4th ACM conference on Recommender systems (RecSys'10) (pp. 119-126). ACM.

[Bellogin13] Bellogin, A., Cantador, I., Diez, F., Castells, P., & Chavarriaga, E. (2013). An empirical comparison of social, collaborative filtering, and hybrid recommenders. ACM Transactions on Intelligent Systems and Technology 4(1), Article 14.

[Bhattacharya14] Bhattacharya, P., Zafar, M. B., Ganguly, N., Ghosh, S., &

Gummadi, K. P. (2014). Inferring user interests in the Twitter social network. In Proceedings of the 8th ACM Conference on Recommender Systems (RecSys'14) (pp. 357-360). ACM.

[Bobadilla10] Bobadilla, J., Serradilla, F., & Bernal J. (2010). A new collaborative filtering metric that improves the behavior of recommender systems. Knowledge-Based Systems 23(6), 520-528.

[Bobadilla11] Bobadilla, J., Ortega, F., Hernando, A., & Alcalá, J. (2011). Improving collaborative filtering recommender system results and performance using genetic algorithms. Knowledge-based systems, 24(8), 1310-1316.

[Bobadilla12] Bobadilla, J., Ortega, F., Hernando, A., & Bernal, J. (2012). A collaborative filtering approach to mitigate the new user cold start problem. Knowledge-Based Systems, 26, 225-238.

[Bobadilla13] Bobadilla, J., Ortega, F., Hernando, A., & Gutierrez, A. (2013). Recommender systems survey. Knowledge-Based Systems 46, 109-132.

[Cambria13] Cambria, E., Rajagopal, D., Olsher, D., & Das, D. (2013). Social big data Analysis. In R. Akerkar (Eds.), Big Data Computing (pp. 401-414). London: Champman and Hall/CRC.

[Chen13] Chen, L., Zeng, W., & Yuan, Q. (2013). A unified framework for recommending items, groups and friends in social media environment via mutual resource fusion. Expert Systems with Applications 40, 2889-2903.

[Dai10] Dai, N., & Davison, B. D. (2010). Freshness matters: in flowers, food and web authority. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'10) (pp. 114-121). ACM.

[Daly09] Daly, E. M., & Haahr, M. (2009). Social Network Analysis for Information Flow in Disconnected Delay-Tolerant MANETs. IEEE Transactions on Mobile Computing 8(5), 606-621.

[Deng14] Deng, S., Huang, L., & Xu, G. (2014). Social network-based service recommendation with trust enhancement. Expert Systems with Applications 41, 8075-8084.

[Ekstrand11] Ekstrand, M. D., Riedl. J. T., & Konstan. J. A. (2011). Collaborative Filtering Recommender Systems. Foundations and Trends in Human-Computer Interaction 4(2), 81-173.

[Esuli06] Esuli, A., & Sebastiani, F. (2006). SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining. In Proceedings of the

5th conference on Language Resource and Evaluation (LREC'06) (pp. 417-422).

[Esuli07] Esuli, A., & Sebastiani, F. (2007). SentiWordNet: a high-coverage lexical resource for opinion mining. Evaluation, 1-26.

[Firan07] Firan, C. S., Nejdl, W., & Paiu, R. (2007). The Benefit of Using Tag-Based Profiles. In Proceedings of the IEEE Latin American Web Conference (LA-WEB'07) (pp. 32-41). IEEE.

[Geyer08] Geyer, W., Dugan, C., Millen, D. R., Muller, M., & Freyne, J. (2008). Recommending Topics for Self-Descriptions in Online User Profiles. In Proceedings of the 2008 ACM Conference on Recommender systems (RecSys '08) (pp. 59-66). ACM.

[Golbeck06] Golbeck, J. (2006). Generating Predictive Movie Recommendations from Trust in Social Networks. In K. Stolen et al., (Eds.), Trust Management (pp. 93-104). Berlin : Springer.

[Granovetter73] Granovetter, M. S. (1973). The Strength of Weak Ties. American Journal of Sociology 78(6) 1360-1380.

[Guy13] Guy, I. (2013). Algorithms for Social Recommendation, In P. Michelucci (Eds.), Handbook of Human Computation (pp. 649-671).

New York : Springer.

[Guy10] Guy, I., Zwedling, N., Ronen, I., Carmel, D., & Uziel, E. (2010). Social Media Recommendation based on People and Tags. In Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'10) (pp. 194-201). ACM.

[He17] He, X., Liao, L., Zhang, H., Nie, L., Hu, X., & Chua, T. S. (2017). Neural collaborative filtering. In Proceedings of the 26th International Conference on World Wide Web (WWW'17) (pp. 173-182). International World Wide Web Conferences Steering Committee.

[Heckathorne10] Heckathorne, W. Speak Now or Forever Hold Your Tweets: Two in five say they aim to influence others when they express their preferences online (2010). http://multivu.prnewswire.com /mnr/harrisinteractive/44373/ Accessed June 2016

[Herlocker99] Herlocker, J. L., Konstan, J. A., Borchers, A., & Riedl, J. (1999). An algorithmic framework for performing collaborative filtering. In Proceedings of the 22nd international ACM SIGIR conference on Research and development in information retrieval (SIGIR'99) (pp.230-237). ACM.

[Jang13] Jang, H., Kim, M., & Shin, H. (2013). KOSAC: A Full-Fledged Korean

Sentiment Analysis Corpus. In PACLIC.

[Jameson07] Jameson, A., & Smyth, B. (2007). Recommendation to groups. In P. Brusilovsky, et al. (Eds.), The adaptive web (pp. 596–627). Berlin: Springer .

[Kim14] Kim, Y. (2014). Convolutional neural networks for sentence classification. arXiv preprint arXiv:1408. 5882.

[Konstas09]Konstas, I., Stathopoulos, V., & Jose, J. M. (2009). On Social Networks and Collaborative Recommendation. In Proceedings of the 32nd international ACM SIGIR conference on Research and development in Information Retrieval (SIGIR'09) (pp. 195-202). ACM.

[Koren09] Koren, Y., Bell, R., & Volinsky, C. (2009). Matrix factorization techniques for recommender systems. Computer, 42(8).

[Koutrika09] Koutrika, G., Bercovitz, B., & Garcia-Molina, H. (2009, June). FlexRecs: expressing and combining flexible recommendations. In Proceedings of the 2009 ACM SIGMOD International Conference on Management of data (pp. 745-758). ACM.

[Krulwich97] Krulwich, B. (1997). Lifestyle finder: Intelligent user profiling using large-scale demographic data. AI magazine, 18(2), 37.

[Kullback51] Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22(1), 79-86.

[Kwak10] Kwak, H., Lee, C., Park, H., & Moon, S. (2010). What is twitter, a social network of a news media?. In Proceedings of the 19th international conference on World Wide Web (WWW'10) (pp. 591-600). ACM.

[Lai13] Lai, C. –H., Liu, D. –R., & Liu, M. -L. (2013). Recommendation Based on Different Aspects of Influences in Social Media. In C. Huemer & P. Lops (Eds.), E-Commerce and Web Technologies (pp. 194-201). Berlin: Springer.

[Lee09] Lee, D. H., & Brusilovsky, P. (2009). Does Trust Influence Information Similarity?. In Proceedings of the ACM Workshop on Recommender Systems & the Social Web (RecSys'09) (pp. 71-74). ACM.

[Li08] Li, X., Guo, L., & Zhao, Y. E. (2008). Tag-Based Social Interest Discovery, In Proceedings of the 17th international conference on World Wide Web (WWW'08) (pp. 675-684). ACM.

[Lin14] Lin, C., Xie, R., Guan, X., Li, L., & Li, T. (2014). Personalized news recommendation via implicit social experts. Information Sciences 254, 1-18.

[Liu14a] Liu, H., Hu, Z., Mian, A., Tian, H., & Zhu, X. (2014a). A new user similarity model to improve the accuracy of collaborative filtering. Knowledge-Based Systems 56, 156-166.

[Liu14b] Liu, J., Li, Z., Tang, J., Jiang, Y., & Lu, H. (2014b). Personalized Geo-Specific Tag Recommendation for Photos on Social Websites. IEEE Transactions on Multimedia 16(3), 588-600.

[Liu13] Liu, X., & Aberer, K. (2013). SoCo: A Social Network Aided Context-Aware Recommender System. In Proceedings of the 22nd international conference on World Wide Web (WWW'13) (pp. 781-802). ACM.

[Ma13] Ma, H. (2013). An Experimental Study on Implicit Social Recommendation. In Proceedings of the 36th international ACM SIGIR conference on Research and development in information retrieval (SIGIR'13) (pp. 73-82). ACM.

[Ma14] Ma, Y., Yu, Z., & Ding, J. (2014). A Method of User Recommendation in Social Networks Based on Trust Relationship and Topic Similarity. In Heyan Huang et al., (Ed.), Social Media Processing (pp. 240-251). Berlin: Springer.

[Manovich11] Manovich, L. (2011). Trending: The promises and the Challenges in Social big data. In M. K. Gold (Eds.), Debates in the Digital

Humanities (pp. 460-475). Minnesota: The University of Minnesota Press.

[Nepal13] Nepal, S., Paris, C., Pour, P. A., Freyne, J., & Bista. S. K., (2013). Interaction Based Content Recommendation in Online Communities. In S. Carberry et al., (Eds.), User Modeling, Adaptation and Personalization (pp. 14-24). Berlin: Springer.

[Ogneva10] Ogneva, M. (2010). How companies can use sentiment analysis to improve their business. Mashable.

[Pang04] Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In Proceedings of the 42nd annual meeting on Association for Computational Linguistics (ACL'04) (p. 271). Association for Computational Linguistics.

[Park07] Park, M. -H., Hong, J. -H., & Cho, S. -B. (2007). Location-based recommendation system using Bayesian user's preference model in mobile devices. In International Conference on Ubiquitous Intelligence and Computing (pp. 1130-1139). Springer, Berlin, Heidelberg.

[Pazzani99] Pazzani, M. J. (1999). A framework for collaborative, content-based and demographic filtering. Artificial intelligence review, 13(5-6), 393-

408.

[Qian13] Qian, X., Feng, H., Zhao G., & Mei, T. (2013) Personalized Recommendation Combining User Interest and Social Circle. IEEE Transactions on Knowledge and Data Engineering 26(7), 1763-1777.

[Rau08] Rau, P. –L. P., Gao, Q., & Ding, Y. (2008). Relationship between the level of intimacy and lurking in online social network services. Computers in Human Behavior 24(6), 2757-2770.

[Saleem08] Saleem, M. Collaborative Filtering: Lifeblood of The Social Web (2008). http://readwrite.com/2008/06/30/collaborative_filtering_social_web/ Accessed June 2016

[Sen09] Sen, S., Vig, J., & Riedl, J. (2009). Tagommenders: connecting users to items through tags. In Proceedings of the 18th international conference on World wide web (WWW'09) (pp. 671-680). ACM.

[Seo17] Seo, Y. D., Kim, Y. G., Lee, E., & Baik, D. K. (2017). Personalized recommender system based on friendship strength in social network services. Expert Systems with Applications, 69, 135-148.

[Seo18] Seo, Y. D., Kim, Y. G., Lee, E., Seol, K. S., & Baik, D. K. (2018). An Enhanced Aggregation Method Considering Deviations for a Group

Recommendation. Expert Systems with Applications, 93, 299-312.

[Seol15] Seol, K., Kim, J. –D., & Baik, D. –K. (2015). Common neighbor similarity-based approach to support intimacy measurement in social networks. Journal of Information Science advance, DOI: 10.1177/0165551515589230.

[Servajean] Servajean, M., Pacitti, E., Gistau, M. L., Yahia, S. A., & Abbadi, A. E. (2014). Exploiting Diversification in Gossip-Based Recommendation. In A. Hameurlain et al., (Eds.), Data Management in Cloud, Grid and P2P Systems (pp. 23-36). Springer International Publishing.

[Servia14] Servia-Rodriguez, S., Diaz-Redondo, R. P., Fernandez-Vilas, A., Blanco-Fernandez, Y., & Pazos-Arias, J. J. (2014). A tie strength based to socially enhance applications and its enabling implementation: mySocialSphere. Expert Systems with Applications 41(5), 2582-2594.

[Shin16] Shin, H., Kim, M., & Park, S. (2016). Modality-based Sentiment Analysis through the Utilization of the Korean Sentiment Analysis Corpus. EONEOHAG : JOURNAL OF THE LINGUISTIC SOCIETY OF KOREA 74, 93-114.

[Stoyanov05] Stoyanov, V., Cardie, C., & Wiebe, J. (2005). Multi-perspective question answering using the OpQA corpus. In Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT'05) (pp. 923-930). Association for Computational Linguistics.

[Tang09] Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2009). ArnetMiner: Extraction and Mining of Academic Social Networks. In Proceedings of the 14th ACM SIGKDD international conference on Knowledge Discovery and Data mining (SIGKDD'09) (pp. 990-998). ACM.

[Tang12] Tang, J., Wu, S., Sun, J., & Su, H. (2012). Cross-domain Collaboration Recommendation. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge Discovery and Data mining (SIGKDD'09) (pp. 1285-1293).

[Toroslu10] Toroslu, Ġ. H. (2010). A SINGULAR VALUE DECOMPOSITION APPROACH FOR RECOMMENDATION SYSTEMS (Doctoral dissertation, MIDDLE EAST TECHNICAL UNIVERSITY).

[Victor10] Victor, P., Cock, M. D., & Cornelis, C. (2010). Trust and Recommendations, In F. Ricci et al., (Eds.) Recommender Systems Handbook (pp. 645-675). Springer US.

[Vosecky14] Vosecky, J., Leung, K. W. -T., & Ng, W. (2014). Collaborative personalized Twitter search with topic-language models. In Proceedings of the 37th international ACM SIGIR conference on Research and development in information retrieval (SIGIR'14) (pp. 53-62). ACM.

[Vozalis07] Vozalis, M. G., & Margaritis, K. G. (2007). Using SVD and demographic data for the enhancement of generalized collaborative filtering. Information Sciences, 177(15), 3017-3037.

[Wu17] Wu, C. Y., Ahmed, A., Beutel, A., Smola, A. J., & Jing, H. (2017). Recurrent recommender networks. In Proceedings of the Tenth ACM International Conference on Web Search and Data Mining (WSDM'17) (pp. 495-503). ACM.

[Xu13] Xu, C., Zhou, M., Chen, F. & Zhou, A. (2013). Detecting User Preference on Microblog. In W. Meng et al., (Eds.), Data Systems for Advanced Applications (pp. 219-227). Berlin: Springer.

[Yager03] Yager, R. R. (2003). Fuzzy logic methods in recommender systems. Fuzzy Sets and Systems, 136(2), 133-149.

[Yin14] Yin, H., Cui, B., Chen, L., Hu, Z., & Huang, Z. (2014). A temporal context-aware model for user behavior modeling in social media systems. In Proceedings of the 2014 ACM SIGMOD International

Conference on Management of Data (SIGMOD ’14) (pp.1543-1554). ACM.

[Yin15] Yin, H., Cui, B., Chen, L., Hu, Z., & Zhou, X. (2015) Dynamic User Modeling in Social Media Systems. ACM Transactions on Information Systems 33(3), 10:1-10:44.

[Yu13] Yu, J., Shen, Y., & Xie, J. (2013). Mining User Interest and Its Evolution for Recommendation on the Micro-blogging System. In J. Wang et al., (Eds.), Web Age Information Management (pp. 679-690). Berlin: Springer.

[Zanda12] Zanda, A., Eibe, S., & Menasalvas, E. (2012). SOMAR: A Social Mobile Activity Recommender. Expert systems with applications 39, 8423-8429.

[Zhen09] Zhen, L., Huang, G. Q., & Jiang, Z. (2009). Recommender system based on workflow. Decision Support Systems 48(1), 237-246.

[Zhu14] Zhu, W., Niu, K., Hu, G., & Xia, J. (2014). Predict User Interest with Respect to Global Interest Popularity. In Proceedings of the 11th international conference on Fuzzy Systems and Knowledge Discovery (FSKD’ 14) (pp. 19-21). IEEE.

ABSTRACT (KOREAN)

소셜 네트워크 서비스의 급격한 발전으로 "소셜 빅 데이터" 라고 불리는 대량의 데이터가 발생하고 있다. 소셜 빅 데이터는 방대한 양과 동시에 다양한 특성을 가지고 있기 때문에, 개인화 추천 시스템의 성능을 향상시키는데 많은 도움이 된다. 이로 인하여, 소셜 빅 데이터를 활용한 개인화 추천 시스템에 대한 연구가 국내외적으로 활발히 이루어지고 있으며, 소셜 빅 데이터 중에서도 사용자들 간의 관계 정보가 가장 많이 사용되고 있다. 하지만 기존의 개인화 추천 연구에서는 사용자들 간의 가까움 정도를 제대로 측정하지 않고 있다. 따라서, 영화나 음악과 같은 특정한 목적을 가지는 단일 도메인의 소셜 빅 데이터 셋을 활용한 연구가 주를 이루고 있다. 개인화 추천연구에서 단일 도메인 상의 실험을 할 경우는 비슷한 성향을 가지는 사용자 집합을 대상으로 실험이 진행되기 때문에, 사용자들간의 친밀도를 정확하게 측정하기 어려운 문제가 발생한다.

본 연구에서는 소셜 빅 데이터 분석을 통해 소셜 네트워크 상에서의 사용자들간의 가까움 정도를 계산하는 적합한 측정법을 제안하며, 이를 "Friendship Strength"라 한다. 또한, Friendship Strength에 기반하여, 트위터 상 사용자에게 적합한 상품, 관심사, 등을 추천해주는 개인화 추천 시스템을 제안한다. 제안하는 개인화 추천 시스템은 단일

도메인 데이터 셋 뿐만 아니라 다양한 도메인을 가지는 데이터 셋에서도 정확한 추천이 가능하다. 이를 검증하기 위해서 한 달치의 트위터 데이터를 활용하여 실험을 진행하였으며, MAE, 정확률, 재현율, F1-measure, NDCG와 같은 다양한 실험 메트릭을 활용하여 제안하는 추천 시스템의 정확성을 검증하였다. 실험 결과를 통해 제안하는 추천 시스템이 기존의 다른 추천 시스템들에 비해 높은 성능을 보임을 확인하였고, Friendship Strength가 개인화 추천 시스템에서 중요한 역할을 함을 확인하였다.