지능정보연구 제19권 제1호 2013년 3월(pp.57~77) http://dx.doi.org/10.13088/jiis.2013.19.1.057

Clustering Method based on Genre Interest for Cold-Start Problem in Movie Recommendation*

Tithrottanak You
Department of Computer and Information Engineering, Inha University
(youtithrottanak@eslab.inha.ac.kr)

Ahmad Nurzid Rosli
Department of Computer and Information Engineering, Inha University
(nurzid@eslab.inha.ac.kr)

Inay Ha
Department of Computer and Information Engineering, Inha University
(inay@eslab.inha.ac.kr)

Geun-Sik Jo
School of Computer and Information Engineering, Inha University
(gsjo@inha.ac.kr)

Social media has become one of the most popular media in web and mobile application. In 2011, social networks and blogs are still the top destination of online users, according to a study from Nielsen Company. In their studies, nearly 4 in 5active users visit social network and blog. Social Networks and Blogs sites rule Americans' Internet time, accounting to 23 percent of time spent online. Facebook is the main social network that the U.S internet users spend time more than the other social network services such as Yahoo, Google, AOL Media Network, Twitter, Linked In and so on. In recent trend, most of the companies promote their products in the Facebook by creating the "*Facebook Page*" that refers to specific product. The "*Like*" option allows user to subscribed and received updates their interested on from the page. The film makers which produce a lot of films around the world also take part to market and promote their films by exploiting the advantages of using the "*Facebook Page*". In addition, a great number of streaming service providers allows users to subscribe their service to watch and enjoy movies and TV program. They can instantly watch movies and TV program over the internet to PCs, Macs and TVs. Netflix alone as the world's leading subscription service have more than 30 million streaming members in the United States, Latin America, the United Kingdom and the Nordics. As the matter of facts, a million of movies and TV program with different of genres are offered to the subscriber. In contrast, users need spend a lot time to find the right movies which are related to their interest genre. Recent years there are many researchers who have been propose a method to improve prediction the rating or preference that would give the most related items such as books, music or movies to the target user or the group of users that have the same interest in the particular items.

One of the most popular methods to build recommendation system is traditional Collaborative Filtering (CF). The method compute the similarity of the target user and other users, which then are

* This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MEST) (No. 2012-0005500).

cluster in the same interest on items according which items that users have been rated. The method then predicts other items from the same group of users to recommend to a group of users. Moreover, There are many items that need to study for suggesting to users such as books, music, movies, news, videos and so on. However, in this paper we only focus on movie as item to recommend to users. In addition, there are many challenges for CF task. Firstly, the "sparsity problem"; it occurs when user information preference is not enough. The recommendation accuracies result is lower compared to the neighbor who composed with a large amount of ratings. The second problem is "cold-start problem"; it occurs whenever new users or items are added into the system, which each has norating or a few rating. For instance, no personalized predictions can be made for a new user without any ratings on the record.

In this research we propose a clustering method according to the users' genre interest extracted from social network service (SNS) and user's movies rating information system to solve the "cold-start problem." Our proposed method will clusters the target user together with the other users by combining the user genre interest and the rating information. It is important to realize a huge amount of interesting and useful users'information from Facebook Graph, we can extract information from the "*Facebook Page*" which "*Like*" by them. Moreover, we use the Internet Movie Database (IMDb) as the main dataset. The IMDbis online databases that consist of a large amount of information related to movies, TV programs and including actors. This dataset not only used to provide movie information in our Movie Rating System, but also as resources to provide movie genre information which extracted from the "*Facebook Page*". Formerly, the user must login with their Facebook account to login to the Movie Rating System, at the same time our system will collect the genre interest from the "*Facebook Page*".

We conduct many experiments with other methods to see how our method performs and we also compare to the other methods. First, we compared our proposed method in the case of the normal recommendation to see how our system improves the recommendation result. Then we experiment method in case of cold-start problem. Our experiment show that our method is outperform than the other methods. In these two cases of our experimentation, we see that our proposed method produces better result in case both cases.

Received : February 22, 2013 Accepted : March 4, 2013
Type of Submission : Excellent Paper of Conference Corresponding author : Inay Ha

1. Introduction

Social media has become one of the most popular media in web and mobile application. In 2011, social networks and blogs are still the top destination of online users, according to a study[1] from Nielsen Company. In their studies, nearly 4 in 5 active users visit social network and blogs. Social Networks and Blogs sites rule Americans' Internet time, accounting to 23 percent of time spent online. Facebook[2] is the main social net-

1) http://cn.nielsen.com/documents/Nielsen-Social-Media-Report_FINAL_090911.pdf
2) http://www.facebook.com/.

work that the U.S internet users spend time more than the other social network services such as Yahoo,[3] Google,[4] AOL Media Network,[5] Twitter,[6] LinkedIn[7] and so on. In recent trend, most of the companies promote their products in the Facebook by creating the "*Facebook Page*[8]" that refers to specific product. The "*Like*" option allows user to subscribed and received updates their interested on from the page. The film makers which produce a lot of films around the world also take part to market and promote their films by exploiting the advantages of using the "*Facebook Page*". In addition, a great number of streaming service providers allows users to subscribe their service to watch and enjoy movies and TV program. They can instantly watch movies and TV program over the internet to PCs, Macs and TVs. Netflix[9] alone as the world's leading subscription service have more than 30 million streaming members in the United States, Latin America, the United Kingdom and the Nordics. As the matter of facts, a million of movies and TV program with different of genres are offered to the subscriber. In contrast, users need spend a lot time to find the right movies which are related to their interest genre. Recent years there are many researchers who have been propose a method to improve prediction the rating or preference that would give the most related items such as books, music or movies to the target user or the group of users that have the same interest in the particular items (Resnick et al., 1994; Hill et al., 1995; Shardanand and Maes, 1995; Lekakos and Giaglis, 2006).

One of the most popular methods to build recommendation system is traditional Collaborative Filtering (CF) (Herlocker, 2000). The method compute the similarity of the target user and other users, which then are cluster in the same interest on items according which items that users have been rated. The method then predicts other items from the same group of users to recommend to a group of users. Moreover, there are many need to study for suggesting to users such as books, music, movies, and videos. However, in this paper we only focus on movie as item to recommend to users. In addition, there are many challenges for CF task. Firstly, the "sparsity problem"; it occurs when user information preference is not enough. The recommendation accuracies result is lower compared to the neighbor who composed with a large amount of ratings. The second problem is "cold-start problem"; it occurs whenever new users or items areadded into the system, which each has norating or a few rating. For instance, no personalized predictions can be made for a new user without any ratings on the record.

In this research we propose a clustering method according to the users' genre interest extracted from social network service (SNS) and user's movies rating information system to solve the "cold-start problem." Our proposed method will clusters the target user together with the other users by combining the user genre interest and

3) http://www.yahoo.com/.
4) http://www.google.com/.
5) http://www.aol.com/.
6) http://twitter.com/.
7) http://www.linkedin.com/.
8) https://www.facebook.com/about/pages.
9) http://www.netflix.com/.

the rating information. It is important to realize a huge amount of interesting and useful users' information from Facebook Graph, we can extract information from the "*Facebook Page*" which "*Like*" by them. Moreover, we use the Internet Movie Database (IMDb[10]) as the main dataset. The IMDb is online databases that consist of a large amount of information related to movies, TV programs and including actors. This dataset not only used to provide movie information in our Movie Rating System, but also as resources to provide movie genre information which extracted from the "*FacebookPage*". Formerly, the user must login with their Facebook account to login to the Movie Rating System, at the same time our system will collect the genre interest from the "*Facebook Page*".

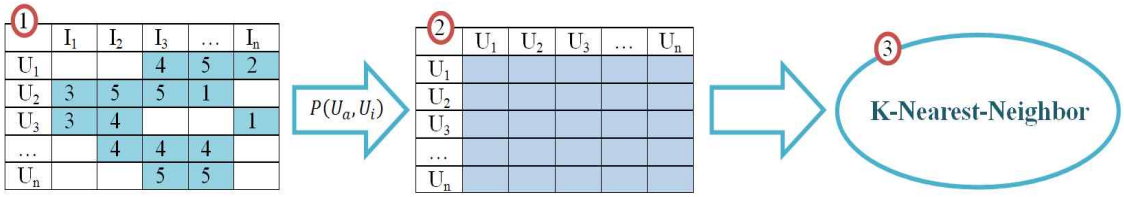
In Section 2, surveys about the related research work and the problem about CF. Then Section 3, give an overview about our propose technique. Section 4, we present our experiment results and Section 5 we conclude our work and discusses the future work.

2. Related Work

In recent years, online social networks and social media websites (e.g., Facebook, Google+, and Twitter) has change how people socialized. Jin et al. (2010) propose a system call "*LikeMiner*" to mined the power of "*Like*" in social media network. They introduce "*Like*" mining algorithms to estimate how the "*Like*" influence the object and introduce a heterogeneous network model for social media with "*Like*". The prototype system show how their propose approach is good for using the large amount of data such as Facebook.

The fundamental ideas of recommendation system are regardless of the facts that recommenders may not explicitly collaborate with the recipients and recommendations system may suggest particularly interesting items (Resnick and Varian, 1997). In typical CF scenario, there is an item which user has rated, or which also reflect their preferences. The ratings can explicit indication, and so forth, on a 1~5 scale, or implicit indication, such purchases or click through (Miller et al., 2004). In the e-commerce system, the recommendation play important role in the web for helping the sellers to recommend users the new items according to their purchased behavior or the rating on the items that they have seen previously. Kim and Moon (2011), address the recommendation system using movie genre similarity and the preferred genre. To compute genre similarity correlations they use Pearson correlation coefficient, and similar cluster is derived. The system defines genre similarity using the correlation within cluster and then, the target users are recommended with a new genre. Melville et al. (2002) implement Content-based (CB) algorithm and CF to enhance user's data to be used in rating prediction. Meanwhile, Li and Kim (2003) also proposed the clustering method which applied to the item-based CF by integrating the information content in the CF. In this work, they adjust the K-mean clustering algorithm by performing a weighted average of deviations from the neighbor's mean, then

10) http://www.imdb.com/.



<Figure 1> Traditional Collaborative Filtering

prediction is computed. On the other hands, a hybrid (CF, CB, Demographic Filtering, Utility-based and Knowledge-based) recommender system can be found in (Burke, 2002).

Let consider that two users have co-rated $I = \{i_1, i_2 \cdots i_n\}$ then it will show as the dimensional space (Adomavicius and Tuzhilin, 2005). The common CF based on user preference is generated from equation (1) to calculate Pearson correlation coefficient between user and other users (Herlocker et al., 1999).

$$P(U_a, U_i) = \frac{\sum_{k=1}^{m}(r_{a,k} - \bar{r}_a)(r_{i,k} - \bar{r}_i)}{\sqrt{\sum_{k=1}^{m}(r_{a,k} - \bar{r}_a)}\sqrt{\sum_{k=1}^{m}(r_{i,k} - \bar{r}_i)}} \quad (1)$$

Where $P(U_a, U_i)$ represent the Pearson correlation coefficient which the value is between -1 to 1. If the value equal to 1 or -1 mean that both user is similar to each other. But when the value is nearly equal to 0 it mean that these user is different preference to each other. $r_{a,k}$, $r_{i,k}$ are the rating of user U_a and U_i for the k^{th} movie. $\bar{r}_a$, $\bar{r}_i$ are the average rating of user U_a and U_i.

As shown in <Figure 1>, (1) we define the users with movies that they have been rated with the matrix. (2) Then we compute Pearson correlation with the user and then users are clustered into a group with the highest similarity. (3) We then pick users with the high similarity with the target user, which represent as a suitable candidate to recommend an item since they stay in the same cluster. Generally, traditional CF is completely based on rating on the movie rating system. The amount of rated movies is important to cluster the target user with the other users into the same group of users. In addition, the system may be suffered if new user is registered or no rated or not enough movies, the system can't determine with other users that have the same interest. To ensure they receive the higher recommendation accuracy, user need to rate the movies as many as possible. It means that traditional CF only can achieve higher recommendation accuracy, with large rated movies. This problem can only be solved if the user rated enough movies or exploiting the demographic information. Similarly when new item added, the only way to solve this is by rating those items or recommending through content analysis.

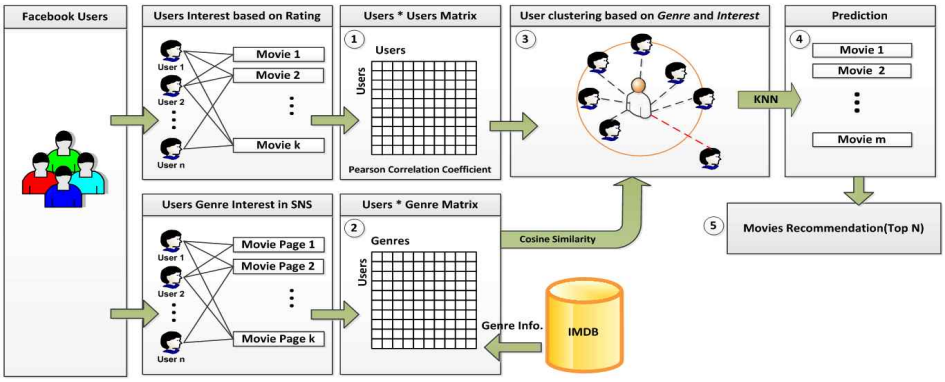
Schein et al. (2011a) have developed a method for recommending items that combines contents and collaborative data under a single probabilistic framework. Lam et al. (2008) discussed a hybrid approaches, using CF, CB and combination of these to address the cold-start problem in order to provide accurate recommendation results.

According to them, there have been a lot of works on solving item-side problems compared to user-side problems. Therefore, the authors proposed a hybrid model based on the analysis of two probabilistic aspect models using pure CF to combine with users' information. Meanwhile, Choi et al. (2010) try to propose a different CF approach based on the category content correlation. First, they compute the category correlations of contents. Second, based on the category correlation that computed before, they design a new recommendation algorithm to recommend to user with certain preference. The advantage of this new approach does not need to compute the preference similarity between users. Another similar study by Pazzani (1999), have proposed a recommendation framework which also a mixed of CF, CB and demographic filtering to recommend users information resources like Web pages or news articles. For the demographic information, they employ Winnow algorithm on HTML home pages to learn the characteristic of homepages associated with users. The efficiency of this framework cannot guarantee the reliability because they only experiment with few users and items. Furthermore, there is no details explanation about how to build the model. A hybrid recommender system using Boltzmann Machines which are the probabilistic model that combine two techniques that is CF and CB is proposed by Gunawardana and Meek (2009). This technique is used to adjust the accuracy of prediction to solve cold-start problem. In the same fashion, Cho and Bang (2011) proposed a method for the new customer recommendation by using a combined measure based on three well-used centrality measures to identify the customers who are most likely to become neighbors of the new customer.

3. Clustering Method Using Genre and Interest in SNS

3.1 System Architecture

The proposed system architecture is divided into five major steps shown in the <Figure 2>;



<Figure 2> System Architecture

(1) We build a movie rating system that allows users to rate movies according to their interest. The information rating is the most important resources to cluster users in to the same group of interest. We compute the similarity between the target user and other users by using Pearson correlation Coefficient. This system not able provide higher recommendation accuracy value if the users only rated a few movies. To address this problem, we take the advantage of SNS resources, Facebook.

(2) Users are required to login in to our movie rating system with their Facebook account. Our systems will extract the "*Facebook Page*" details, particularly in movie category that users have subscribed by "*Like*" function in order to obtain the movie titles from each page. The IMDb is the main dataset to provide movie details such as genre, release date, actors, actress and so on. In our work, we only retrieve the movie genre from each movie title obtained from "*Facebook Page*". Then we compute the similarity between the target user and other users by using the Cosine similarity based on genre interest.

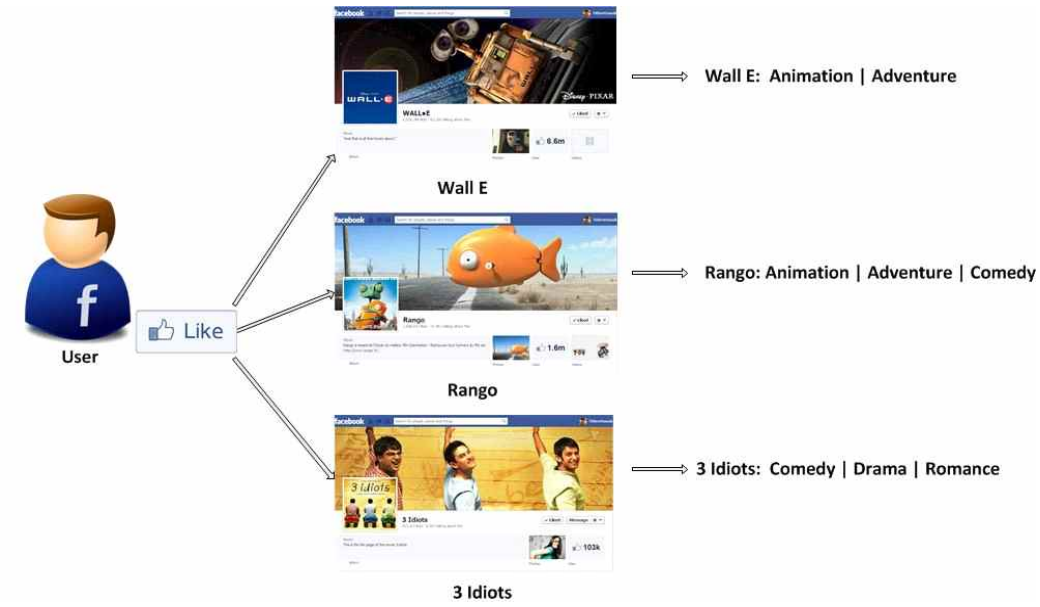
(3) After we compute the similarity values from step (1) and (2), we now drive the new similarity of the users and users obtained from equation (7). Therefore, we can easily cluster the target user and the rest of users in to the specific cluster according to their genre preference from SNS and the rating information.

(4) In this step, we predict rating of other movie that user have not seen yet, in order to recommend to them by selecting the highest prediction rating score.

(5) Finally, we listed out the Top N movies which have the high rating prediction are chosen to recommend to users.

3.2 Genre Extraction

The growing trend of Social network service such as Facebook, Google+, Twitter and other related SNS have fundamentally change our social lives, both personal and community level. Each has their own specific compelling feature to compete with each other. The Facebook has one of the killer features called "*Facebook Page*". "*Facebook Page*" is the feature that enables users to create their own pages for their purpose. The feature is similar to advertising pages where the admin or the owner can post an updates and discussions. This post information will be appeared to user's timeline who subscribed ("*Like*") the pages. There are many companies, organization, product and individuals get an advantage from this. "*Facebook Page*" are divided into various categories including; 1) Local business or places; 2) Company, organization or institution; 3) Brand or product; 4) Artists, band or public figure; 5) Entertainment and; 6) Cause or community. In our research work, we only consider movie category.



<Figure 3> Genre Interest of Facebook Pages

In "*Facebook Page*", over one billion people "*Like*" and comment an average of 3.2 billion times every day[11]. It proves that the products or items, in our case movies or film is part of the conversation which film maker has access to the most powerful kind of word-of-mouth marketing. Additionally, according to Facebook's S1 IPO filing[12], "*Facebook Page*" generates an average of 2.7 billion "*Like*" and comments per day during the three months ended December 31, 2011. The film maker also gain exposure to users who are "friends of fan", which those who are connected with user that "*Like*" pages. A recent study from ComScorefound[13], users who "*Like*" a page, a more likely to make a purchase from that retailer. A recent study by Sean et al. (2012), proposed a method to filter a user's friends for information search, which leverage the semantic video annotation and SNS. The main idea is to create a discussion and post it to the right person or the subscriber that potentially can give opinion on particular product shown on the video. This may help user to make decision to purchase the product. They construct the user profile by extracting the information from "*Facebook Page*". Then they compute the similarity between target user and friend lists to find the most common friend that may participate in the discussion for the particular items or products. Similar work by Aart et al. (2009) determines user interest from the Facebook and other SNS platform (Twitter, YouTube[14] and Last.fm[15])

11) https://www.facebook.com/business/overview.
12) http://www.cnbc.com/id/46227868.
13) http://www.comscoredatamine.com/2011/07/facebook-fans-worldwide-region-starbucks-southwest-bing/.

from NoTube[16] BeanCounter to aggregate users' activities. In the same manner, Bertini et al. (2012), also exploiting user's social graph and their "*Like*" obtained from Facebook on the purpose of generating semantic user profile.

The Facebook has become the most powerful marketing tools. It allows the marketers to target their post to segment of fans with certain genders, ages and Likes. Therefore, they can tailor market messages to a specific audience. Moreover this news feed will be shown to user's friend timeline which they subscribe or "*Like*" the page. For instance, if a user has 500 friends, all of his/her friends will be received the "*Like*" feed. Let us consider if someone "*Like*" a movie "*Facebook Page*" called "The Karate Kid". All updates regarding to the movie will posted on my timeline. At the same time, if another friend seeing this feed, they are potentially "*Like*" the same page, and this may be effect to other friends in the his/her Facebook community.

Facebook Graph APIs[17] provide a powerful way to retrieved information from registered user. In order to obtain the information, our system will require user permission to access his/her profile. We will obtain information such as user id, gender, birthdate, occupation, location, and "*FacebookPage*". The "*Facebook Page*" contains information such as title, about, picture, video and other information. In addition, the user who is the fan of the page can also participate on the page. In our work we only extract the "*Facebook Page*" title which generally represents the movie title.

In order to get movies genre, we use IMDb[18]. The IMDb is known as the world's most popular and authoritative source for movie, TV and entertainment program. As shown in <Figure 3>, we use IMDb to retrieve genres of each "*Facebook Page*".

On the other hand, some of the movies are described with three genres, but it is in different order. The order of genre describes the most content composition of the corresponding movie. According to our survey on the movies dataset, most of the movies has different composition amount of genre. While some of the movies have only one genre, others have two genres and the rest at most have three genres, and only few with more than three genres. Therefore, we define the three groups of movies according to the number of genres composition they have. Thus, the each movie isrepresents by genre composition as the following:

- M { $Genre_1$ }
- M { $Genre_1$, $Genre_2$ }
- M { $Genre_1$, $Genre_2$, $Genre_3$ }

Each movie has different genres composition according to movie characteristic. Furthermore, the genre order also shows what is the most genre composition that can describe about the movie. Therefore, we define the weight to each

14) http://www.youtube.com/.
15) http://www.last.fm/.
16) http://notube.tv/.
17) https://graph.facebook.com/.
18) http://www.imdb.com/.

genre of the movie in each group of the movies.

- M { $\text{Genre}_1 \times \alpha$}, (2)
  $\alpha = 1$
- M { $\text{Genre}_1 \times \alpha$, $\text{Genre}_2 \times \beta$}, (3)
  $\alpha + \beta = 1(\alpha = 0.6,\ \beta = 0.4)$
- M { $\text{Genre}_1 \times \alpha$, $\text{Genre}_2 \times \beta$, $\text{Genre}_3 \times \gamma$}, (4)
  $\alpha + \beta + \gamma = 1(\alpha = 0.6,\ \beta = 0.3,\ \gamma = 0.1)$

3.3 Users Genre Interest in SNS

Next, once we have defined the importance of genre in each movie, we then calculate the user interest. Each user is represents with the feature vector from different genre interest as follows:

$$\text{User}_n = \{(G_1, \text{Value}_1), (G_2, \text{Value}_2), (G_3, \text{Value}_3), \cdots, (G_{17}, \text{Value}_{17}), (G_{18}, \text{Value}_{18})\} \quad (5)$$

Where $G_1, G_2, \cdots, G_{18}$ is the specific genre and it has its value with $Value_1, Value_2, \cdots; Value_{18}$ generated from sum of genre value in the same type of each user's movies and each movie defined in equation (2), (3) and (4). Each User_n and with each genre interest value will form the matrix of user interest on genre as shown in <Table 1>.

To compute the similarity between users to other users, according to genre interest from SNS ("*Facebook Page*") we use cosine similarity and the equation is shown in equation (6) :

$$sim_G(U_a, U_i) = \frac{\sum_{j=1}^{n} G_{a,j} \times G_{i,j}}{\sqrt{\sum_{j=1}^{n} (G_{a,j})^2}\ \sqrt{\sum_{j=1}^{n} (G_{i,j})^2}} \quad (6)$$

Where $sim_G(U_a, U_i)$ is the similarity between users U_a and U_i which based on the genre interest from "*Facebook Page*". $G_{a,j}$ and $G_{a,i}$ are the value of interest for users U_a and U_i on genre *j*.

It is important to realize that we only using 18 out of 26 genres from IMDb, as shown in <Table 2>. The main reason is, some of genres (e.g.: Talk-show, Game-Show, News) are not categorized as movie genre.

<Table 2> Genre

No	Genre	No	Genre
G_1	Action	G_{10}	Film-Noir
G_2	Adventure	G_{11}	Horror
G_3	Animation	G_{12}	Musical
G_4	Children's	G_{13}	Mystery
G_5	Comedy	G_{14}	Romance
G_6	Crime	G_{15}	Sci-Fi
G_7	Documentary	G_{16}	Thriller
G_8	Drama	G_{17}	War
G_9	Fantasy	G_{18}	Western

<Table 1> Genre Interest Matrix

	G_1	G_2	G_3	G_4	G_5	G_6	G_7	G_8	G_9	⋯	G18
U_1	1.5	1.8	3.6	0	1.3	0.6	0	3.2	0.5	⋯	0
U_2	0.9	4	3	0	0.8	0.6	1	10.1	1.7	⋯	0
U_3	5.8	1.5	0.6	0	1.7	1.9	0	3.6	0.8	⋯	0.1
U_4	9.6	1.7	2.4	0	1.7	1.8	0	1	0.4	⋯	0
⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯	⋯
U_n	10.5	10.2	13	0	9.6	1.3	1	7.9	2.9	⋯	0.1

3.4 Similarity Combination

To improve recommendation system quality and to solve the cold-start problem, we combine the traditional CF method with our proposed method which exploit user's preferred genres which are mined from SNS. For this reason, we combine equation (1) and (6) by adding *a* value, therefore we get :

$$sim(U_a, U_i) = asim_G(U_a, U_i) + (1-\alpha)P(U_a, U_i) \quad (7)$$

Where $sim_G(U_a, U_i)$ is the similarity value between users U_a and U_i produced by equation (6). $P(U_a, U_i)$ is Pearson correlation coefficient between users U_a and U_i. Meanwhile, α is the value range between 0 to 1. Noted, the value of α is exchangeable in order to see how *a* effect the performance of our system. As a result, we can generate a new similarity $sim(U_a, U_i)$ value to improve the clustering quality of user in group according to their best similarity. We select the Top-N users that have the most similarity to the target user.

3.5 Predicting preference

In order to predict movies items, we then compute a weighted average of deviations from the neighbors. The top *N* rule is used to select the nearest *N* neighbors based on the similarities of users' genre interest and the preferred movies. The general equation for the prediction on the item *i* of user *a* is :

$$P_{a,i} = \bar{r}_a + \frac{\sum_{k=1}^{n}(r_{i,k} - \bar{r}_k) \times sim(u_a, u_k)}{\sum_{k=1}^{n} sim(u_a, u_k)} \quad (8)$$

where $P_{a,i}$ is the prediction for the target user *a* on the movie *i*, *n* mean all neighbors of movie *i*, $\bar{r}_a$ is the average rating of user a, $r_{i,k}$ is the rating of user k on movie *i*, $\bar{r}_k$ is the rating of user u_k on movie *i*, $sim(u_a, u_k)$ represent the similarity between the target user u_a and the user u_k.

Finally, after we compute equation 8, we get a list of movie candidates to recommend to user. The value of the prediction of each movie is sorted ascending pattern. Movies with the highest prediction rating value are selected for the target user.

4. Experimental Evaluation and Results

4.1 Prediction Performance Builder-Rating System

To improve the prediction performance, we perform explicit moving ratings analysis with rating value range from 1 to 5. The experiment involved 50 users which accessed to the rating system with Facebook login ID. It also involved 624 movies or films title released in year 2012. It is important to realize that, by accessing the system with Facebook ID, we are eventually accessed users "*Like*" information. We used Facebook Graph APIs to access "*Facebook Page*" information.

We employ the MAE (Mean Absolute Error)

metrics to evaluate the accuracy performance of our proposed prediction method. A great number of researcher used MAE to evaluate the recommendation accuracy by comparing the predicted value with user-provide values (Herlocker et al., 2004).

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|p_i - r_i| \quad (9)$$

Where n is the number of movie to predict, p_i is the predicted value, and r_i is the actual rating, the lowest value is the better prediction.

4.2 Benchmark Algorithm

It is also important to note, we also compare our proposed algorithm with other research work conduct by Chikhaoui et al. (2011). The author proposed an approach that combines rating prediction CF, CB and demographic filtering which includes the age, occupation, and the zip code. The CF and CB values are obtained using the Pearson correlation and the cosine similarity. According to Aimeur et al. (2006), demographic filtering helps to cluster the new user in to the specific group of users which have the similar demographic information. K-Nearest Neighbors value is selected to predict the movie in order to recommend to user. This combination approach successfully solved the cold-start users which have less of rating information.

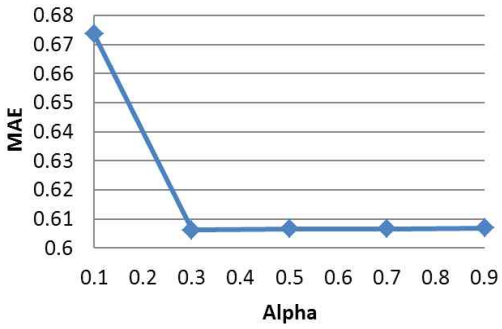
We also considered another research work conduct by Sun, Luo and Zhang (2011b). In their research work, they first cluster user that have enough preference using K-mean algorithm. Then they build the decision tree of user's to cluster the user base on demographic information such as occupation, gender, and age. Finally, movies with the high prediction value are recommended to target user. We experimented on the μ value to test with our dataset to see which value will may produce the highest value of MAE. As a results, we have selected two parameters to be used on our data set; μ = 0.2 and k = 3 for the k-mean cluster.

4.3 Experiment and Result

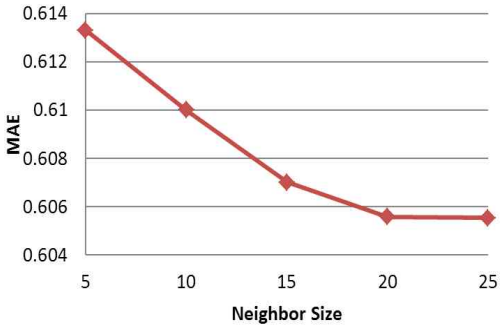
For the experiment purpose, we randomly split data sets into two set; 1) Training set and 2) Test set with the proportion 80% : 20% each. With this data set we compute the Pearson correlation coefficient for all participated users.

In order to get better value of user's genre prediction similarity and Pearson correlation coefficient we employ the α value to be employ in equation (7). It is important to realize we do compare the α value, $\alpha = \{0.1, 0.3, 0.5, 0.7, 0.9\}$ beforehand. For each target user, we predict the movie that user has not yet seen before. For each given α value, we compute the average MAE of all users to find better result from α value. As shown in <Figure 4> we can conclude, $\alpha = 0.3$ is performing better result of MAE compared to other value.

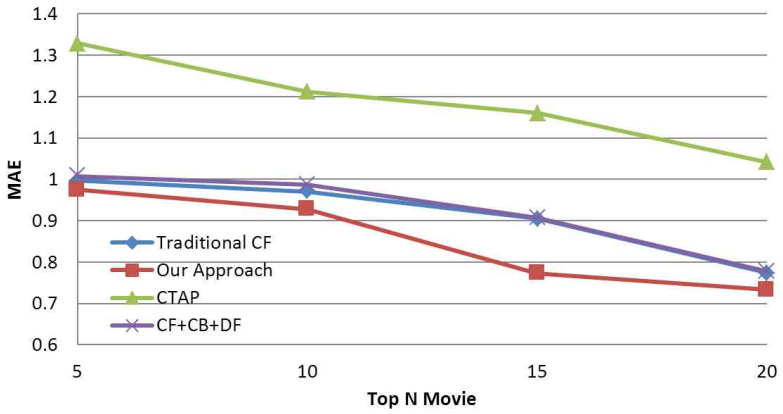
In order to examine the sensitivity of neighbor size, we perform an experiment with the nearest neighbors and computed the MAE for each groups. We employ $\alpha = 0.3$ value in the experiment resulted in the <Figure 5>. As shown,



<Figure 4> Influence of Alpha Value



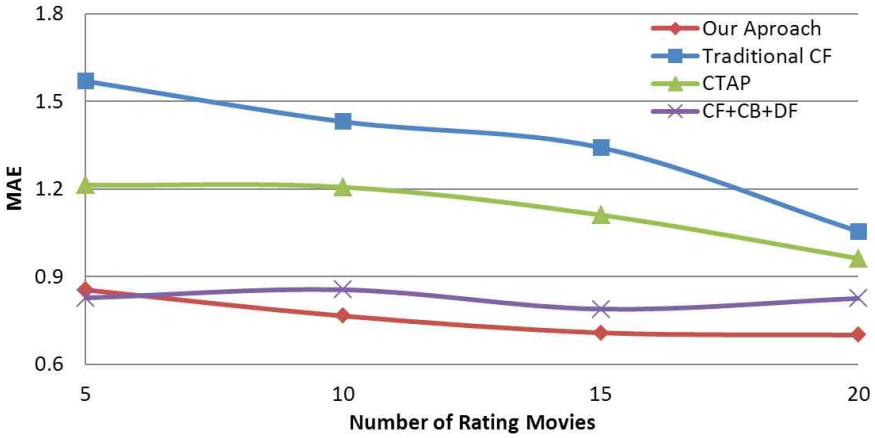
<Figure 5> Efficiency of Neighbor Size



<Figure 6> Top N Movies Recommendation

the numbers of nearest neighbors influences the performance of the system. In fact, from the experiment, we can see that the nearest neighbors with value 20 produced better result of MAE compared to the other result.

As shown in the <Figure 6> Top N movie recommendation influence to our system. We conduct several numbers of top N movies to recommend such as 5, 10, 15 and 20. In this experiment result, we can conclude that top 20 movies is likely produce better result in order to recommend to user. In short, we recommend 20 movies to each user. Furthermore, we compare the Top-N recommendation our approach with other method such as the traditional CF, CTAP and CF+CB+DF. As seen in <Figure 6>, our approach is obviously outperformed compared to other results. The combination approach, which integrate CF+CB+DF has produced comparable results with the traditional CF. This is certainly resulted from a number of movies rated by users. Moreover, such information from DF and CB will be dismissed if movie information rate is enough. Meanwhile, CTAP approach denoted poor result as we expected. The drawback are resulted from the cluster value which almost similar to centroid



<Figure 7> Cold-Start with Number of Rating Movies

value. In addition, CTAP is design not for the normal recommendation system, but instead to solve the cold-start problem. Our proposed approach produced such significant results cause by additional information from the "*Facebook Like*".

In this work, our proposed approach is designedsolve thecold-start problem. To achieve this, we select users which have small number of movies rating and then test it with the traditional CF method, combination CF+CB+DF and CTAP approach. A number of movies rated are range from 5, 10, 15 and 20 movies. We employ $\alpha = 0.3$ value obtained from <Figure 4> to combine between the similar user genres interest in SNS and rating movies information and we defined 20 neighbors obtained from <Figure 5>. Our proposed approach produced significant performance results to solve the cold-start problem (See <Figure 7>). Traditional CF approaches result lower accuracy due to the lack of information ratings. On the other hand, CTAP algorithm producess lightly better than the traditional algorithm cause by cold-start users that are clustered into the same group. As seen on <Figure 7>, we found that if users rate five movies, it produced almost similar results between our proposed approach and CF+ CB+DF approach. Basically, CF+CB+DF approach relies on movies prediction values which are obtain from movie ratings. The prediction values are adjusted according to rating information adequacy. Moreover, this algorithm requires a lot of information from each user such as DF and CB to improve the cold-start problem. Further analysis with 10, 15, and 20 rated movies clearly shown that our proposed approach is outperform compare to others. It is important to highlights our propose method is a combination approach of the user's genre interest obtained from SNS and the traditional CF. The useful resources of genre interest from "*Facebook Page*" help users in cold-start problem. With this information, users are clustered into same genre interest; therefore it produces reliable prediction which may recommend new movie for new users.

4.4 Discussion

In this section, we will discuss about the advantages and the limitation of our proposed approach. In this work, we only experiment with a small number of users (50 users) with 634 movies dataset. As seen from the experiment results, it is clearly show that our approach produce far better results compared to the three benchmark algorithms, which improve the recommendation results. First, our propose method which is combining between the traditional CF and user's genre interest from Facebook produce a better result by comparing to others three benchmark algorithms. It indicates that our method can use to improve the recommendation. Second, our method can solve the cold-start problem in recommender system. This means that we only need only few "*Facebook Page*" from the new user we can predict other movie to recommend to users. Although our proposed method is not better result compare to CF+CB+DF method, our method just add one additional information which is the genre interest in SNS, our method can produce most similar to CF+CB+DF method which the values are 0.82 and 0.85 respectively at the point that user has 5 rating. CF+CB+DF required a lot of information in order to achieve this result. In addition we summary that our propose method can save the computational coast as the result of our method required one additional information but we can achieve better result.

5. Conclusions

In this paper, we have described the design and implementation of recommendation system based ensemble information from SNS "*Facebook Page*" and movies rating system. We develop the hybrid method which employs the CF in user movies rating, together user genre interest extracted from "*Facebook Page*" on SNS. The rating information alone cannot solve the cold-start user due to lack of rating information to compute the similarity value between the target user and users. Therefore, we exploit "*Facebook Page*"features retrieve user movie interest genre which then relates to the online movie database known as IMDb. With the proposed method, the system may predict an item of movies to users that belongs into the same groups of movie genre. The user similarity is derived from the genre interest from "*Like*" page on "*Facebook Page*". This significantly contributes the quality of user clustering, since the Pearson correlation coefficient can't cluster the user in to the group if the user doesn't have enough rating data. We have conducted an experiment and evaluate the accuracy results with few benchmark algorithms (Traditional CF, CF+CB+DF and CTAP method). According to result, our proposed method improves the recommendation result by producing 10% better value than the benchmarks algorithm. In fact, we also successfully solved the cold-start problem, with significant increase in accuracy value up to 15% compared to benchmark algorithm. This again, based on our novel method to exploit "*Like*" in "*Facebook Page*". To summarize, our experiment results have demonstrate that our proposed method is outperform compared to benchmark algorithm.

In this work we mainly focus on the genre interest from SNS. However, we also realize that "*Facebook Page*", have plenty of useful information resources that we can exploit to improve our system and solve cold-start problem. In the future work, we plan to consider more about the relationship strength between users, the closeness of the users (Park and Cho, 2011) and more users' preferences such as preferred actors, actresses and authors from "*Facebook Page*" that can be considered to improve in our recommender system.

References

Aart, C. V., L. Y. Raimond, D. Brickley, G. Schreiber, M. Minno, L. Miller, D. Palmisano, M. Mostarda, R. Siebes, and V. Buser, "The No Tube Beancounter : aggregating user data for television programme recommendation", *in Proceedings of the Linked Data on the Web Workshop (LDOW 2009)*, Madrid, Spain, 2009.

Adomavicius, G. and A. Tuzhilin, "Toward the next generation of recommender system : A survey of the state-of-the-art and possible extensions", *IEEE Trans. Knowl. Data Eng.,* Vol.17, No.6(2005), 734~749.

Aimeur, E., G. Brassard, J. M. Fernandez, and F. S. Onana, "Privacy-preserving demographic filtering", *In Proceedings of the 2006 ACM symposium on Applied computing*, (2006), 872~878.

Bertini, M., A. D. Bimbo, A. Ferracani, and D. Pezzatini, "A Social Network for Video Annotation and Discovery Based on Semantic Profiling", *in Proceedings of the 21st international conference companion on World Wide Web*, Lyon, France, 2012.

Burke, R., "Hybrid recommender systems : Survey and experiments", *User modeling and user-adapted interaction*, Vol.12, No.4(2002), 331~370.

Chikhaoui, B., M. Chiazzaro, and Wang, S., "An Improved Hybrid Recommender System by Combining Predictions", *In Advanced Information Networking and Applications (WAINA), 2011 IEEE Workshops of International Conference*, (2011), 644~649.

Cho, Y. and J. Bang, "Applying Centrality Analysis to Solve the Cold-Start and Sparsity Problems in Collaborative Filtering", *Journal of Intelligence and Information Systems*, Vol.17, No.3(2011), 99~114.

Choi, S. M. and Y. S. Han, "A content recommendation system based on category correlations", *In The fifth international multi-conference on computing in the global information technology*, (2010), 1257~1260.

Choi, S. M., S. K. Ko, and Y. S. Han, "A movie recommendation algorithm based on genre correlations", *Expert Systems with Applications*, 2012.

Gunawardana, A. and C. Meek, "A unified approach to building hybrid recommender systems", *In of the third ACM conference on Recommender systems*, (2009), 117~124.

Hameed, M. A., O. Al Jadaan, and S. Ramachandram, "Collaborative Filtering Based Recommendation System : A survey", *International Journal on Computer Science and Engineering*, Vol.4, No.5(2012), 859~876.

Herlocker, J. L., J. A. Konstan, A. Borchers, and J. Riedl, "An algorithmic framework for performing collaborative filtering", *in Proceedings of the SIGIR conference, New York, NY, USA : ACM*, (1999), 230~237.

Herlocker, J. L., J. A. Konstan, L. G. Terveen,

and J. T. Riedl, "Evaluating collaborative filtering recommender systems", *ACM Transactions on Information Systems (TOIS)*, Vol. 22, No.1(2004), 5~53.

Herlocker, J. L., J. A. Konstan, and J. Riedl, "Explaining collaborative filtering recommendations", *In of the 2000 ACM conference on Computer supported cooperative work*, (2000), 241~250.

Hill, W., L. Stead, M. Rosenstein, and G. Furnas, "Recommending and evaluating choices in a virtual community of use", *Proceedings of the ACM CHI ' 95 Conference on Human Factors in Computing Systems*, (1995), 194~201.

Jin, X., C. Wang, J. Luo, X. Yu, and J. Han, "Likeminer : A system for mining the power of 'like' in social media networks", *In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, KDD*, 2011.

Jung, J. J., "Attribute selection-based recommendation framework for long-tail user group : An empirical study on movie lens dataset", *Expert Systems with Application*, 2011.

Kim, K. R. and N. Moon, "Recommender system design using movie genre similarity and preferred genres in SmartPhone", *Multimedia Tools and Applications*, (2012), 1~18.

Lam, X. N., T. Vu, T. D. Le, and Duong, A. D., "Addressing cold-start problem in recommendation systems", *In of the 2nd international conference on Ubiquitous information management and communication*, (2008), 208~211.

Lekakos, G. and G. M. Giaglis, "Improving the prediction accuracy of recommendation algorithms : Approaches anchored on human factors", *Interacting with computers*, Vol.18, No.3 (2006), 410~431.

Li, Q. and B. M. Kim, "Clustering approach for hybrid recommender system", *In Intelligence, 2003, Proceedings, IEEE/WIC International Conference*, (2003), 33~38.

Melville, P., R. J. Mooney, and R. Nagarajan, "Content-boosted collaborative filtering for improved recommendations", *In of the National Conference on Artificial Intelligence*, *Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press;* (1999), 187~192.

Miller, B. N., J. A. Konstan, and J. Riedl, "Pocket Lens : Toward a personal recommender system", *ACM Transactions on Information Systems (TOIS)*, Vol.22, No.3(2004), 437~476.

Park, J. H. K. and Y. H. Cho, "Social Network Analysis for the Effective Adoption of Recommender System", *Journal of Intelligence and Information Systems*, Vol.17, No.4(2011), 305~316.

Pazzani, M. J., "A framework for collaborative, content-based and demographic filtering", *Artificial Intelligence Review*, Vol.13, No.5(1999), 393~408.

Resnick, P. and H. R. Varian, "Recommender systems", *Communications of the ACM*, Vol.40, No.3(1997), 56~58.

Resnick, P., N. Iacovou, M. Suchak, P. Bergstrom, and J. Riedl, "Grouplens : an open architecture for collaborative filtering of netnews", *In : Proceedings of the ACM Conference on Computer Supported Cooperative Work*, (1994), 175~186.

Sean, V., K. J. Oh, and G. S. Jo, "A User Profile-based Filtering Method for Information Search in Smart TV Environment", *Journal of Intelligence and Information Systems*, Vol.18, No.3 (2012), 91~117.

Schein, A. I., A. Popescul, L. H. Ungar, and D. M. Pennock, "Methods and metrics for cold-

start recommendations", *In of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, (2002), 253~260.

Shardanand, U. and P. Maes, "Social information filtering : algorithms for automating word of mouth", *In : Proceedings of the ACM CHI'95 Conference on Human Factors in Computing Systems, Denver, Colorado*, (1995), 210~217.

Sun, D., C. Li, and Luo, Z., "A content-enhanced approach for cold-start problem in collaborative filtering", *In Intelligence, Management Science and Electronic Commerce (AIMSEC), 2011 2nd International Conference on IEEE*, (2011), 4501~4504.

Sun, D., Z. Luo, and Zhang, F., "A novel approach for collaborative filtering to alleviate the new item cold-start problem", *In Communications and Information Technologies (ISCIT), 2011 11th International Symposium on*, (2011), 402~406.

Wang, Z., M. Zhang, Y. Tan, W. Wang, Y. Zhang, and L. Chen, "Recommendation Algorithm Based on Graph-Model Considering User Background Information", *2011 Ninth International Conference on Creating, Connecting and Collaborating through Computing*, (2011), 32~39.

Abstract

영화 추천 시스템의 초기 사용자 문제를 위한 장르 선호 기반의 클러스터링 기법

유띳로따낙* · 누르지드* · 하인애* · 조근식**

소셜 미디어는 모바일 어플리케이션과 웹에서 가장 많이 사용되는 미디어 중 하나이다. Nielsen사의 보고서에 따르면 소셜 네트워크 서비스와 블로그가 온라인 사용자의 주 활동 공간으로 사용되고 있으며, 미국인 중에서 온라인 활동이 왕성한 5명의 사용자중 4명은 매일 소셜 네트워크 서비스와 블로그를 방문하고 온라인 활동 시간의 23%를 소비한다고 집계하고 있다. 미국의 인터넷 사용자들은 야후, 구글, AOL 미디어 네트워크, 트위터, 링크드인 등과 같은 소셜 네트워크 서비스중 페이스북에서 가장 많은 시간을 소비한다. 최근에는 대부분의 회사들이 자신의 특정 상품에 대하여 “페이스북 페이지(Facebook Page)”를 생성하고 상품에 대한 프로모션을 진행한다. 페이스북에서 제공되는 “좋아요” 옵션은 페이스북 페이지를 통해 자신이 관심을 가지는 상품(아이템)을 표시하고 그 상품을 지지할 수 있도록 한다. 많은 영화를 제작하는 영화 제작사들도 페이스북 페이지와 “좋아요” 옵션을 이용하여 영화 프로모션과 마케팅에 이용한다. 일반적으로 다수의 스트리밍 서비스 제공업들도 영화와 TV 프로그램을 즐기며 볼 수 있는 서비스를 사용자들에게 제공한다. 이 서비스는 일반 컴퓨터와 TV 등의 단말기에서인터넷을 통해 영화와 TV 프로그램을 즉각적으로 제공할 수 있다. 스트리밍 서비스의 선두 주자인 넷플릭스는 미국, 라틴 아메리카, 영국 그리고 북유럽 국가 등에 3천만 명 이상의 스트리밍 사용자가 가입되어 있다. 또한 넥플릭스는 다양한 장르로 구성된 수백만 개의 영화와 TV 프로그램을 보유하고 있다. 하지만 수많은 콘텐츠로 인해 사용자들은 자신이 선호하는 장르에 관련된 영화와 TV 프로그램을 찾기 위해 많은 시간을 소비해야 된다. 많은 연구자들이 이러한 사용자의 불편함을 줄이기 위해 아이템에 대한 사용자가 보지 않은 아이템에 대한 선호도를 예측하고 높은 예측값을 갖는 아이템을 사용자에게 제공하기 위한 추천 시스템을 적용하였다.

협업적 여과 방법은 추천 시스템을 구축하기 위해 가장 많이 사용되는 방법이다. 협업적 여과 시스템은 사용자들이 평가한 아이템을 기반으로 각 사용자 간의 유사도를 측정하고 목적 사용자와 유사한 성향을 가진 사용자 그룹을 결정한다. 군집된 그룹은 이웃 사용자 집단으로 불리며 이를 이용하여 특정 아이템에 대한 선호도를 예측하고, 예측 값이 높은 아이템을 목적 사용자에게 추천해 준다. 협업적 여과 방법이 적용되는 분야는 서적, 음악, 영화, 뉴스 및 비디오 등 다양하지만 논문에서는 영화에 초점을 맞춘다. 이 협업적 여과 방법이 추천 시스템 내에서 유용하게 활용되고 있지만 아직 “희박성 문제”와 “콜드 스타트

* 인하대학교 IT 공과대학 컴퓨터정보공학과
** 인하대학교 IT 공과대학 컴퓨터정보공학부

문제” 등 해결해야 할 과제가 남아있다. 희박성 문제는 아이템의 수가 증가할수록 아이템에 대한 사용자의 로그 밀도가 감소하는 것이다. 즉, 전체 아이템 수에 비해 사용자가 아이템에 대해 평가한 정보가 충분하지 않기 때문에 사용자의 성향을 파악하기 어렵고, 이로 인해 사용자가 아직 평가하지 않은 아이템에 대해서 선호도를 추측하기 어려운 것을 말한다. 이 희박성 문제가 포함된 경우 적합한 이웃 사용자 집단을 형성하는데 어려움을 겪게 되고 사용자들에게 제공되는 아이템 추천의 질이 떨어지게 된다. 콜드 스타트 문제는 시스템 내에 새로 들어온 사용자 또는 아이템으로 지금까지 한 번도 평가를 하지 않은 경우에 발생한다. 즉, 사용자가 평가한 아이템에 대한 정보가 전혀 포함되어 있지 않거나 매우 적기 때문에 이러한 경우 또한 적합한 이웃 사용자 집단을 형성하는데 어려움을 겪게 되고 사용자가 평가하지 않은 아이템에 대한 선호도 예측의 정확성이 감소되게 된다.

본 논문에서는 영화 추천 시스템에서 발생될 수 있는 초기 사용자 문제를 해결하기 위하여 사용자가 평가한 영화와 소셜 네트워크 서비스로부터 추출된 사용자 선호 장르를 활용하여 사용자 군집을 형성하고 이를 활용하는 방법을 제안한다. 소셜 네트워크 서비스로부터 사용자가 선호하는 영화 장르를 추출하기 위해 페이스북 페이지의 ‘좋아요’ 옵션을 이용하며, 이 ‘좋아요’ 정보를 분석하여 사용자의 영화 장르 관심사를 추출한다. 페이스북의 영화 페이지는 각 영화를 위한 페이스북 페이지로 구성되고 있으며, 사용자는 자신의 선호도에 따라서 “좋아요” 옵션을 선택할 수 있다. 사용자의 페이스북 정보는 페이스북 그래프 API를 활용하여 추출되고 이로부터 사용자 선호 영화를 알 수 있게 된다. 시스템에서 활용되는 영화 정보는 인터넷 영화 데이터베이스인 IMDb로부터 획득한다. IMDb는 수많은 영화와 TV 프로그램을 보유하고 있으며, 각 영화에 관련된 배우 정보, 장르 및 부가 정보들을 포함한다. 논문에서는 사용자가 “좋아요” 표시를 한 영화 페이지를 이용하여 IMDb로부터 영화 장르 정보를 가져온다. 그리고 추출된 영화 장르 선호도와 본 시스템에서 제안하는 영화 평가 항목을 이용하여 유사한 이웃 사용자 집단을 구성한 후, 사용자가 평가하지 않은 아이템에 대한 선호도를 예측하고, 높은 예측 값을 갖는 아이템을 사용자에게 추천한다.

본 논문에서 제안한 사용자의 선호 장르 기반의 사용자 군집 기법을 이용한 시스템을 평가하기 위해서 IMDb 데이터 집합을 이용하여 사용자 영화 평가 시스템을 구축하였고 참가자들의 영화 평가 정보를 획득하였다. 페이스북 영화 페이지 정보는 참가자들의 페이스북 계정과 페이스북 그래프 API를 통해 획득하였다. 사용자 영화 평가 시스템을 통해 획득된 사용자 데이터를 제안하는 방법에 적용하였고 추천 성능, 품질 및 초기 사용자 문제를 벤치마크 알고리즘과 비교하여 평가하였다. 실험 평가의 결과 제안하는 방법을 적용한 추천 시스템을 통해 추천의 품질을 10% 향상시킬 수 있었고, 초기 사용자 문제에 대해서 15% 완화시킬 수 있음을 볼 수 있었다.

Keywords : 초기 사용자 추천 시스템, 협업적 여과, 클러스터링, 장르 선호도, 소셜 네트워크

저 자 소 개

Tithrottanak You

Received a B.S. degree in Computer Science from the Royal University of Phnom Penh, Cambodia, in 2010. He is a M.S. Candidate majoring in Computer and Information Engineering at Inha University, Korea. His research interests include Recommender System, Collaborative Filtering, Social Network, Video Annotation and Augmented Reality.

Ahmad Nurzid Rosli

Received a B.S. degree in Information Technology in 2002, and M.S. degree in Information Technology in (2005) from Universiti Utara Malaysia (UUM), Malaysia. After receiving his Master degree, in 2006, he commenced his career as a lecturer at Universiti Pendidikan Sultan Idris (UPSI), Malaysia. He is currently a Ph.D. Candidate in Information Engineering at Inha University, Korea. His research interest includes Augmented Reality (AR), Semantic Web, Social Networking Service (SNS) and Collaborative Learning.

Inay Ha

Received a B.S. degree Computer Science from Suwon University, Korea, in 2005, and a M.S. degree in Information Engineering from Inha University, Korea in 2007. She is currently a Ph.D. Candidate in Information Engineering of Inha University, Korea. Herresearch interests include Information Personalization, Semantic Web, Social Network and Recommender System.

Geun-Sik Jo

Is a Professor in Computer and Information Engineering, Inha University, Korea. He received the B.S. degree in Computer Science from Inha University in 1982. He received the M.S. and the Ph.D. degrees in Computer Science from City University of New York in 1985 and 1991, respectively. He has been the General Chair and/or Technical Program Chair of more than 20 international conferences and workshops on artificial intelligence, knowledge management, and semantic applications. His research interests include knowledge-based scheduling, ontology, semantic Web, intelligent E-Commerce, constraint-directed scheduling, knowledge-based systems, decision support systems, and intelligent agents. He has authored and coauthored five books and more than 200 publications.