Identifying Representative Ratings for a New Item in Recommendation System

Sang-Min Choi
Department of Computer Science
Yonsei University, Shinchon Dong 134,
Seodaemungu, Seoul, Republic of Korea
+82-2-2123-7434
jerassi@cs.yonsei.ac.kr

Yo-Sub Han*
Department of Computer Science
Yonsei University, Shinchon Dong 134,
Seodaemungu, Seoul, Republic of Korea
+82-2-2123-7434
emmous@cs.yonsei.ac.kr

ABSTRACT

With the development of the Internet, the users share information using Web applications. Because of this reason, there is lots of information on the Web. The information includes not only high quality information, but also useless one. With the phenomena, the recommendation system appears on the Web. Existing information recommendation systems on the Web have known problems. One famous problem is cold-start. We tackle the cold-start problem for a new item in recommendation system. To alleviate cold-start for a new item, we use method for identifying representative reviewers in raters group and recommendation algorithm based on category correlations. The representative reviewers mean the users who represent their raters group. Namely, the ratings of the reviewers can represent the average ratings of other users. If there are the ratings for new items rated by the representative reviewers, then we can consider the ratings rated by many other users. We predict the ratings of these reviewers for a new item. To predict ratings, we use the recommendation algorithm based on the category correlations. This algorithm can draw the prediction results without ratings since the algorithm uses category information. We propose the prediction results of the representative reviewers as the representative ratings for a new item. We propose the algorithm to alleviate cold-start for a new item and show the reliability of our approach through tests.

Categories and Subject Descriptors

H.4 [**Information Recommendation**]: Information Recommendation

General Terms

Algorithm, Experimentation

*Corresponding author

ICUIMC(IMCOM)'13, January, 17–19, 2013, Kota Kinabalu, Malaysia

Keywords

Recommendation System, Social Group, Cold-Start Problem, Representative Reviewers

1. INTRODUCTION

In a society, we have many forms of relationships with other people from home, work and school. These relationships give rise to a social network. People in a social network receive, provide and pass lots of information. We often observe that there is a group of people who have significant influence on other people. We call these highly influential people opinion leaders. The general public accepts information not only from mass media, but also from opinion leaders. Thus, it is important and useful to identify opinion leaders in a social network. Since the late 20^{th} century, the number of Internet users has increased rapidly. Many users interact with each other in an online social network. This makes the Web community similar to real society. Thus, it is a natural task to find influential users in an online society. For example, many online articles posted by influential bloggers are used as marketing tools for companies or political advertisements for parties since these articles have huge influence on other users. Recently, the research [2] that finds representative reviewers in social-media user-network has been progressed. In this research, representative reviewers mean the users who can indicate ratings for other users.

Meanwhile, with the development of the Web, many people can use lots of information. These internet users can upload their data and download wanted information using Web platform. Because of uploading behavior, the users encounter diverse information, however, they are exposed to data that almost users do not want such as spam data. For this problem and efficiency of searching data, the recommendation system occurs on the Web.

The recommendation system has been development with various researches. The researches generally use similarities between users or items [10, 1, 5, 9], or associative methods using similarities of both [8]. Through these researches, accuracy of recommendation has risen. On the other hand, there are some problems that could not solve up to now. The problems are transparency, cold-start, and sparsity problems. The transparency is the problem that concerned with the ratings in the system. The recommendation results are ambiguity since the ratings can change with mentation of

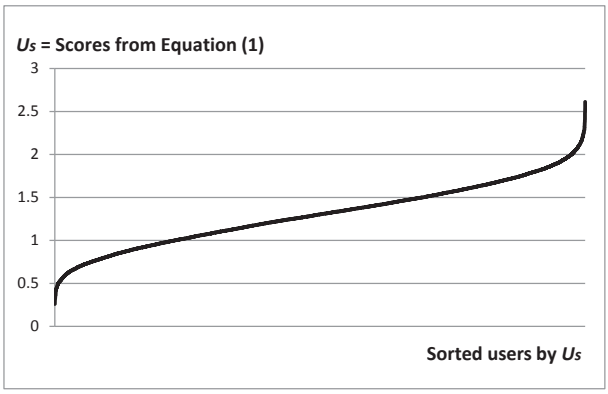

Figure 1: The distribution of scores (U_s) from Equation (1) in ascending order

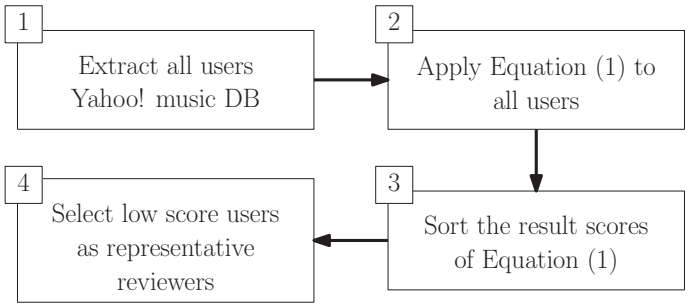

Figure 2: The procedure of identifying representative users from Yahoo! music dataset

users. The cold-star problem occurs when no information for new users of items exists to make a recommendation [12, 7]. Finally, the sparsity problem occurs when the database has small amount of information. This problem affects accuracy of recommendation. We deal with the cold-start problem for new items. To alleviate the cold-start problem, we use the method that finds representative reviewers in social-media user-network and the recommend media contents based on category correlations.

We first revisit the previous researches on finding influential users in online society and recommending the media contents for new users in Section 2. Next, we explain our approach that uses both the method of identifying representative reviewer and the recommendation algorithm based on category correlations in Section 3. Then, we show the validity of our approach using the GroupLens movie database [1]. Finally, we conclude this paper in Section 5 with future works.

2. RELATED WORKS

In this section, we first introduce the method that identifies representative reviewers in internet social media. Then, we revisit the research that alleviates the cold-start problem for users using category correlations in items.

2.1 Idetifying Representative Reviewers

There have been some researches on social networks. Han et al. [4] showed a social network based on video contents and user activities. They modified PageRank algorithm to draw user reputation in the contents-based social-network. Then they proposed an algorithm that composed the social-network from users of the video contents and derived user reputation based on uploading and subscription. Choi et al. [2] suggest an algorithm that finds representative reviewers from raters group in Yahoo music database [2]. Han et al. and Choi et al. show the representative users based on analyzing contents-based social-network. We introduce an algorithm that identifies representative reviewers in internet social media and analyzes the results for identified representative reviewers.

To extract representative reviewers, Choi et al. used Equation (1).

$$U_s = \frac{\sum_{i \in A} |(R_S(i) - R_\mu(i)|}{|A|}, \quad (1)$$

where A is a set of contents rated by user S, and $|A|$ is the cardinality of A. $R_S(i)$ is the rating of content i by user S, and $R_\mu(i)$ is the average rating of content i by user S. Note that the result of Equation (1) shows how close each users rating is to the average rating. We select those who have low scores from Equation (1) as representative reviewers. Choi et al. first extracted all users in the database, and then applied Equation (1) to all users using their ratings. Next, they sorted the result scores of Equation (1), selecting low-score users as representative reviewers. Figure 1 shows the result of Equation (1) sorted in ascending order when the Yahoo music database was applied. The previous research selected the users who were positioned with low scores as the representative reviewers. In summary, the algorithm used in this research has four steps. Figure 2 illustrates the procedure used to identify representative reviewers in the Yahoo music database. First, all users in the database were extracted, and then Equation (1) was applied to the extracted users. Finally, the results of Equation (1) were sorted in ascending order and low-score users were selected as representative reviewers.

[1]http://grouplens.org/node/12
[2]http://webscope.sandbox.yahoo.com

2.2 A Content Recommendation System based on Category Correlations

Some researchers have proposed several methods to avoid the cold-start problem [12, 7, 11]. One such approach is to use information that is reliable and available initially. Note that this type of information may not always be available. Choi et al. [1, 3] used information which is reliable and available. They proposed a movie recommendation system based on genre correlations. In their research, they considered genres as category information. The genre information is provided by movie experts such as directors. This system does not need much input data. Their system first calculates correlations for genres using the genre combination for each movie. Then, the system applies existing and input information such as genre combinations and user-preferred genres to the average rating of each movie based on genre correlations. The final step of the recommendation system is based on category correlations that comprise a recommendation list based on computed points for movies. We offer more details of the system below.

2.2.1 Calculating genre correlations

Genre correlations are calculated using genre combinations for each movie. In this approach, Choi et al. use the GroupLens movie database. There exist total of 18 genres in GroupLens movie database. Table 1 shows 18 genres in the database. Each movie in a database has at least one genre. This means that each movie has a genre combination which is composed of at least one genre. This genre information is given by movie experts such as directors, while user preferences are provided by users. The system chooses a genre and counts the number of other genres for each movie. For example, if movie A has a genre combination composed of G_1, G_3, and G_8, then G_1 is first selected as a criterion genre. The system increases the combination counting with G_3 and G_5 by 1. Second, G_2 is selected as a criterion genre and the system also increases the combination counting with only G_8 by another 1. This approach repeats the procedure for genre counting for all movies and draws the genre correlations by using percentages.

Table 1: Each Genre in the Database

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

2.2.2 *Applying genre correlations*

The next step is to apply genre correlations to average ratings of movies. In this step, the system uses Equation 1 and user preferred genres. The user preferred genres are input by each user.

$$R_p = \frac{\sum_{i \in up} (\sum_{j \in mg} R_{ij} M_\mu)}{|up|}, \quad (2)$$

In Equation (2), up refers to a set of preferred genres provided by the user, and mg refers to the set of genre combinations for each movie. R_{ij} is the genre correlation between genre i and j. M_μ is the average rating of a movie M. The result of Equation (2) is that the recommendation points are drawn by applying the genre correlation of the movie and the preferred genres of the user to the average rating of movie M. If genres i and j are the same, then R_{ij} becomes one. Thus, the system selects a criterion genre sequentially from among the user-preferred genres and each criterion genre applies as many genre correlations to the average preference for movie A as the number of genres of movie A.

3. OUR APPROACH

We tackle the cold-start problem for new items in the recommendation system. To alleviate this problem, we use two approaches that explained in Section 2. First of all, the representative reviewers can represent ratings for other users in their raters group. Therefore, the ratings for new items have possibility that can indicate the cumulative effects in the group. Generally, we can gain the results that have reliability when the ratings for new items are added in the system [6, 13]. If we have the ratings for new items by representative reviewers, we can draw the rating for the new item since the reviewers represent average ratings for other users. Namely, if we have the ratings for new items that have similarity to real ratings given by users, then we can alleviate the cold-start problem for new item. Nevertheless, there exists a problem that the representative reviewers also do not have the ratings for the new item. This case also causes the cold-start since there are activities for users. We can alleviate this problem using prediction of recommendation system. Namely, if we predict ratings of the new items for representative reviewers, we can gain the ratings of the new items that have the effectiveness which are given by many users without any input. Thus, we can consider the predicted ratings as the average rated by many users.

We first select the representative reviewers using GroupLens movie database. Then, we draw selection tendency of the users for category. Next, we calculate category preference scores based on the selection tendency. Finally, we draw the ratings for new items using category correlations, category combinations for new item, and selection tendency of the users.

3.1 Identifying Representative Reviewers using GroupLens Database

We select the representative reviewers using GroupLens movie database. Table 2 shows the database and Figure 3 shows the distribution of score from equation (1) in ascending order. In the database, there are 3,883 items and 6,040 users. In Figure 3, we consider the 10% users as representative reviewers.

Table 2: GroupLens movie database

Dataset	Attribute	Explanation
Movie Dataset	MovieID, Title, Genre	There are total of 3,883 movies
User Dataset	UserID, Gender, Age, Occupation, Zip-code	There are total of 6,040 users
Rating Dataset	UserID, MovieID, Rating, Timestamp	There are total of 1,000,209 ratings

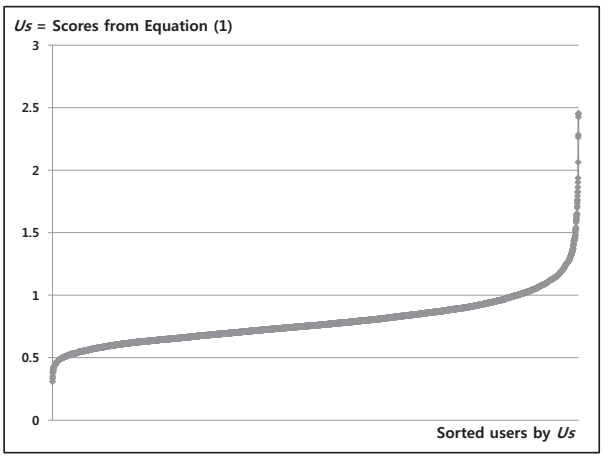

Figure 3: The distribution of score (U_s) from Equation (1) in ascending order

3.2 Drawing Selection Tendency and Category Preference of Users

First, we draw the selection tendency. Table 3 shows an example of the drawing selection tendency using movie items. User A has selected 8 movies and we consider the genres in the movie list as category combination.

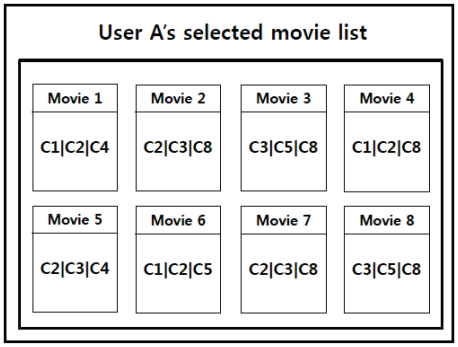

Figure 4: Selected Movie List of User A

Each movie in the list has a category combination. The first movie, called movie 1, has categories $C1$, $C2$, and $C4$. This means, for example, $C1$ is animation, $C2$ is for children, and $C4$ is comedy. All movies in the list have a similar category combination. Selection preference of A user A is studied by counting how many times each category appears in the list. For example, the selection tendency of category $C1$ is 3 since $C1$ appears total three times in movies 1, 4, and 6. The selection tendency of category 2 is 6, from movies $1, 2, 4, 5, 6$, and 7. We repeat this until the last category in the list. Then, we organize selections of user A into selected movie list. Table 3 shows selection tendencies of user A.

Table 3: Selection Tendency of User A

Category ID	Selection Tendency
C1	3
C2	6
C3	5
C4	2
C5	3
C8	5

Next, we draw the category preference using the selection tendency. The category preference is drawn with the ratings for each item and the selection tendency. Figure 5 shows the example of the process to draw the category preference.

This example uses the selection tendency of an user A in Figure 3 and shows that draw the category preference for category $C2$. In the selected movie list in Figure 4, category $C2$ appears total of six times, from movies $1, 2, 4, 5, 6$, and 7. We first check the ratings of each movie rated by user A and give ratings to each movie as the category preference for $C2$. In this example, the $C2$ of movie 1 has 5 as category preference. Then, we calculate the average score for each category preference of movies. We consider this average score as category preference.

Why we calculate the average score for each category is that the ratings of users can change according to the items. If a user prefers specific category, we cannot convince that this user not always give the high ratings to preferred category. Because of this reason, we have to consider the various preferences of a user for items that have preferred category of the user. Thus, we calculate the average score when we draw the category preference.

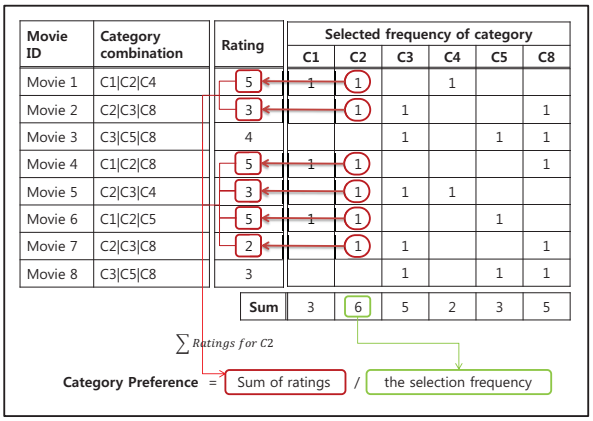

Figure 5: The Example of the Process to Draw the Category Preference

3.2.1 *Drawing the Ratings for a New Item*

A new movie item has at least one category. We first compare category combination of a new item to the categories that appeared in user selection tendency. If the user selection tendency has the category that same with the new item one, we give the category preference to this same category in the new item. If there is no same category between the new item and selection tendency, then we draw a rating using the correlations between the category in the new item and all categories in selection tendency. Figure 6 shows the example of drawing the ratings for a new item. In Figure 6,

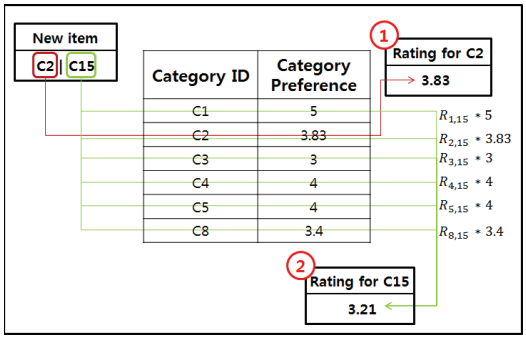

Figure 6: The Example of Drawing the Ratings for a New Item

the new item have $C2$ and $C15$ as category combination. To draw the ratings, we first check whether $C2$ exists in the selection tendency. We determine the category preference of $C2$ in selection tendency as the rating of $C2$ for the new item since there is $C2$ in the selection tendency. In Figure 6, we can show that the category $C15$ is not in the selection tendency. In this situation, to determine the rating for category $C15$ for the new item, we use the category correlations between $C15$ and all categories in selection tendency. We apply category correlation between both to category preference and calculate average for the applied scores. Then

we determine this average as the rating for $C15$ for the new item. Finally, we draw the rating for the new item using average of these two results. Equation (3) is the one for the rating of a new item when there is no same categories between category combination of a new item and the selection tendency of user.

$$C_r = \frac{\sum_{i \in ST} R_{ij} CP_i}{|ST|}, \quad (3)$$

In Equation (3), ST means the selection tendency of user and R_ij means the category correlations between category Ci and Cj. The CP_i means the category preference of Ci and $|ST|$ is the number of categories in the selection tendency.

4. TEST AND ANALYSIS

To our test, we use GroupLens movie database. We first randomly select 1,000 items of 3,883 items and apply our approach to the selected items. Then, we apply Equation (1) to all users in the database and sort the results in ascending order. in the database, the number of users are 6,040. After this, we calculate the difference between the results that drawn to our approach and real average rating for items in the database. Then, we observe these differences between top users and bottom users. Generally, top users are representative reviewers [2]. Figure 7 shows the results of top 50

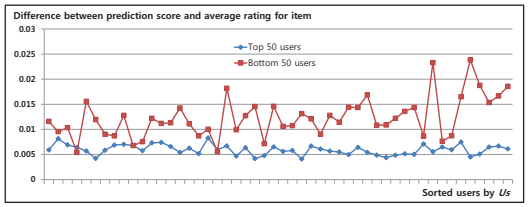

Figure 7: The Result Graph for Top 50 Users and Bottom 50 Users

users and bottom 100 users. The graph for top 50 users is stable than bottom 50 users. It means that the users in the top 50 have generally similar results. In Figure 8, the graph

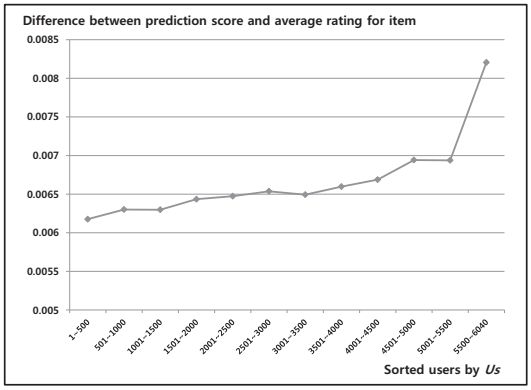

Figure 8: The Result Graph for All Users in the Database

shows that the top users have lower difference than bottom users. We can consider that the top users in Figure 3 are also representative reviewers in our approach. Thus, we can call the scores that drawn by our approach the representative ratings.

5. CONCLUSIONS

With the development of the Web, many people can use lots of information. Some users upload information and some other use these uploaded data. This uploading behaviors cause that the Web users encounter diverse information. Because of this reason, the users exposed not only their wanted information but also data that the users do not want such as spam data. For this problem and efficiency of searching data, the recommendation system occurs on the Web and many researchers have developed this system. However, the recommendation systems have some problems. One of famous problem is cold-start problem that occurs when no information for new users of items exists to make a recommendation [12, 7, 11]. We have tackled the cold-start problem for new items in the recommendation system. To alleviate this problem, we have used two approaches [2, 3]. First approach is identifying representative reviewers in the raters group and second one is a content recommendation system based on category correlations. To apply each approach, we have proposed drawing selection tendency and category preference of users. Then, we have drawn the ratings for a new item using category preference and correlations. Finally, we have shown the representative ratings for a new item using method for identifying representative reviewers. In near future, we will apply the other recommendation algorithm for ratings of representative reviewers. Then we will compare results of various recommendation algorithms and ours.

6. ACKNOWLEDGMENTS

This paper was supported by the Basic Science Research Program through NRF funded by MEST (2010-0009168), (2012R1A1A2044562).

7. REFERENCES

[1] S.-M. Choi and Y.-S. Han. A content recommendation system based on category correlations. In *Proceedings of the 2010 Fifth International Multi-conference on Computing in the Global Information*, ICCGI '10, pages 1257–1260, 2010.

[2] S.-M. Choi and Y.-S. Han. Representative reviewers for internet social media. *Expert Systems with Applications*, 40(4):1274–1282, 2013.

[3] S.-M. Choi, S.-K. Ko, and Y.-S. Han. A movie recommendation algorithm based on genre correlations. *Expert Systems with Applications*, 39(9):8079–8085, 2012.

[4] Y.-S. Han, L. Kim, and J.-W. Cha. Computing user reputation in a social network of web 2.0. *Computing and Infomatics*, 31(2):447–462, 2012.

[5] J. L. Herlocker, J. Konstan, A. Borchers, and J. Riedl. An algorithm framework for peforming collaborative filtering. In *Proceedings of the 1999 Conference on Research and Development in Information Retrieval*, pages 230–237, 1999.

[6] Z. Huang, H. Chen, and D. D. Zeng. Applying associative retrieval techniques to alleviate the sparsity problem in collaborative filtering. *ACM Transactions on Information Systems*, 22(1):116–142, 2004.

[7] M. Ishikawa, P. Géczy, N. Izumi, T. Morita, and T. Yamaguchi. Information diffusion approach to cold-start problem. In *Web Intelligence/IAT Workshops*, pages 129–132, 2007.

[8] T.-H. Kim and S.-B. Yang. An effective recommendation algorithm for improving prediction quality. In *Australian Conference on Artificial Intelligence*, pages 1288–1292, 2006.

[9] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM conference on Electronic commerce*, EC '00, pages 158–167, 2000.

[10] B. M. Sarwar, G. Karypis, J. A. Konstan, and J. Riedl. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th International World Wide Web Conference*, pages 285–295, 2001.

[11] A. I. Schein, A. Popescul, L. H. Ungar, and D. M. Pennock. Methods and metrics for cold-start recommendations. In *The 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 253–260, 2002.

[12] B. Scholz, S.-M. Choi, S.-K. Ko, H.-S. Eom, and Y.-S. Han. Analyzing category correlations for recommendation system. In *Proceedings of the 5th International Conferenece on Ubiquitous Information Management and Communication*, ICUIMC '11, pages 1:1–1:7, New York, NY, USA, 2011. ACM.

[13] D. C. Wilson, B. Smyth, and D. O'Sullivan. Sparsity reduction in collaborative recommendation: A case-based approach. *international journal of pattern recognition and artificial intelligence (ijprai)*, 17(5):863–884, 2003.