Recommender Systems using Category Correlations based on WordNet Similarity

Sang-Min Choi, Da-Jung Cho, Yo-Sub Han
Dept. of Computer Science
Yonsei University
Seoul, Republic of Korea
{jerassi, dajung, emmous}@cs.yonsei.ac.kr

Ka Lok Man, Yan Sun
Xi'an Jiaotong-Liverpool University
Suzhou, China
{ka.man, yan.sun}@xjtlu.edu.cn

Abstract— Recently, many internet users are not only information consumers but also information providers. There is lots of information on the Web and most people can search information what they want through the Web. One problem of the large number of data in the Web is that we often spend most of our time to find a correct result from search results. Thus, people start looking for a better system that can suggest relevant information instead of letting users go through all search results: We call such systems recommendation systems.

Conventional recommendation systems are based on collaborative filtering (CF) approaches. The CF approaches have two problems: sparsity and cold-start. Some researchers have studied to alleviate the problems in CF approaches. One of them is the recommendation algorithm based on category correlations. In this study, researchers utilize genre information in movie domain as category. They have drawn genre correlations using genre counting method. This approach can alleviate the user-side cold-start problems, however, there exists one problem that extensions of the approach are less likely. If a domain has singular category, then we cannot apply previous approaches. It means that we cannot draw category correlations. Because of this reason, we propose a novel approach that can draw category correlations for not only multiple categories but also singular one. We utilize word similarities provided by WordNet.

Keywords—recommendation systems; genre correlations; WordNet similarity

I. INTRODUCTION

We search and gain lots of information from the Web since the late 20th century. One big difference from the Web and the traditional content provides is that we can search what we are interested from the Web. For example, from a music magazine, it is not easy to find information about a particular song. On the other hand, in the music search site such as Yahoo! Music, we can find songs by simply typing in a song title [1]. We can also obtain information about books, shops or movies from the internet [2, 3]. However, huge amount of data do not always guarantee satisfactory outcome. Because of lots of spam data and wrong information on the Web, we often spend the most of our time to search for relevant information from search results by going through all of them. Namely, the accuracy and reliability of search results become very low.

In a recommendation system, users do not need to go through all search results. The system actively suggests items that are likely interested to users based on user information and, thus, eliminate the burden of looking at all results. The recommendation systems are often based on collaborative filtering [2, 3]. A conventional CF is as follows: First, users provide preferences for a set of items. Then, the system identifies groups of users with similar preference based on a similarity measure of preferences. When the system suggests an item to a user X, it first determine the group that X belongs to and suggest relevant items based on the preferences of users in the group [2]. This approach works well when there are enough user preference data. In other words, it is difficult to suggest good items if there are not enough data to create user groups [3].

There are some studies to overcome the problems that systems have not enough user data [4, 5]. Choi et al. [5] try to alleviate the problems using genre correlations. They compute genre correlations by utilizing content information. They count co-occurrence of genres in same contents. It means that this approach apply only the contents that have genre combination.

We try to solve this problem using WordNet similarity [6]. We first consider the genres as words. Then we compute similarity between genres (words) by addressing WordNet similarity based on hypernym of each word.

II. OUR APPROACH

The previous approach that utilizes genre correlations in movie domain has two steps to make recommendation list for a user [5]. The first step is to draw genre correlations. Choi et al. use the genre combination of each movie in the database to draw genre correlations. All movies have at least one genre. In other words, each movie has a genre combination which is composed of at least one genre. For example, the movie Toy Story has a combination of 'Animation', 'Children's' and 'Comedy'. This means that the movie Toy Story has characteristics of these three genres. The study uses this information. We select a genre and count the number of the other genres for each movie. For example, the movie *A* has genre combination of G1, G2, and G5. Then G1 is selected as a criterion genre which would lead to G2 and G5 to increase by 1. Next, G2 is selected as a criterion genre, this will further lead to only G5 to increase by another 1. In this way, we apply this to each movie in the database.

978-1-4799-1888-1/15 $31.00 © 2015 IEEE
DOI 10.1109/PlatCon.2015.26

If there is a movie with n-number of genres, the genre counting for such a movie will require us to increase the counts for each genre. We repeat this until the last genre. This figure shows the demonstration of genre counting in a movie with n-genres.

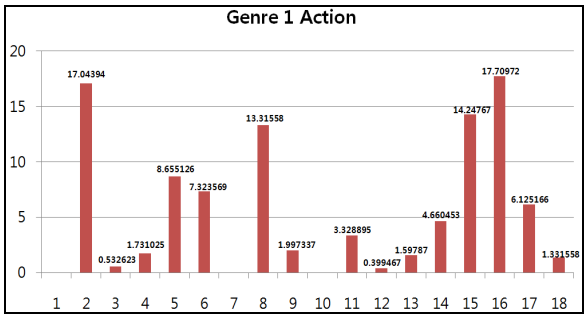

Figure 1. Genre correlations between genre 1 action and others (based on genre counting)

Figure 1 shows the percentage about genre correlations between gener1 Action genre and other genres. In this graph, y-axis is percentage of genre correlations and x-axis is genres. The action genre has a high-correlation with the genre 2 Adventure, genre 16 Thriller and genre 15 Sci-fi. That means that the action genre frequently correlates with adventure, thriller and sci-fi genre.

The next step of this approach is to apply genre correlations. In order to apply genre correlations, we need some more information. One is the user's preferred genres, and the other is the average ratings of movies. The user's preferred genres are gained through the users' inputs, and the average ratings of movies are gained through the ratings of each movie in the database.

In summary, first, we apply the user's preferred genres using genre correlations to the average rating, we calculate the new score for each movie using the average rating and genre correlations. Then we sort the scores in descending order and recommend the high position movies.

This previous approach that can make recommendation list using at least one user preferred genre can alleviate to user-side cold-start problems since conventional CF require more than 15 ratings to initial users. In the recommendation algorithm based on genre correlations, authors consider the genre information as category. The previous approach has a limitation since they utilize genre combination to calculate to genre correlations. It means that if the shape of the genre is not a combination, then we cannot draw the genre correlations. For example, in Yahoo! Music database, all music contents have only one genre. If we apply genre correlation approach to Yahoo! Music database, then we cannot gain the genre correlations. Because of this reason, we propose a novel approach that can draw genre correlations when there is no genre combination.

We use word similarity provided by WordNet [6]. WordNet provides similarity between words based on inherited hypernym. Thus, we can gain hierarchical similarity between words through WordNet. We compute similarity between 18 genres (words) in MovieLens database [7]. Figure 2 shows the percentage about genre correlations between gener1 Action genre and other genres based on our approach.

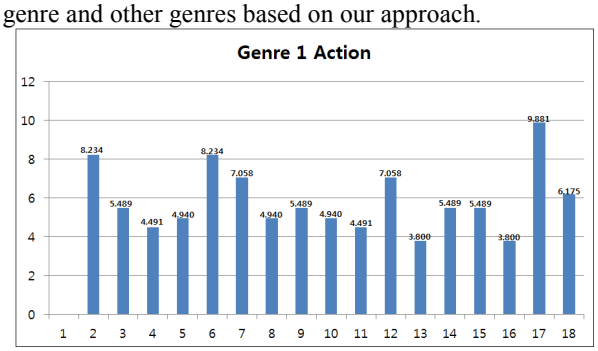

Figure 2. Genre correlations between genre 1 action and others (based on WordNet similarity)

We can also draw genre correlations by utilizing WordNet similarity. If we address these correlations and approach, we can compute genre correlations of a content that has not genre combination but one genre.

III. Conclusions

Previous approach has limitation for extensibility since the approach count co-occurrence of genres in same movie. It underlines that all genres appear on a content as combination. It means that all contents have at least one genres. Namely, we cannot apply previous approach to a domain that contents have only one genre or category. Thus, we propose a novel approach to compute genre correlations when contents have only one genre or category. We utilize WordNet similarity.

In near future, we apply the genre correlations drawn by WordNet to recommendation systems. We also test our approach to various types of domain such as movie, music and book.

Acknowledgement

The research presented in this paper is partially supported by the Basic Science Research Program through NRF funded by MEST (2012R1A1A2044562) and by Xi'an Jiaotong-Liverpool University (XJTLU) Research Development Fund RDF-13-03-18.

References

[1] http://new.music.yahoo.com

[2] Badrul Sarwar, George Karypis, Joseph Konstan, and John Rie, "Item-based Collaborative Filtering Recommendation Algorithms", Accepted for publication at the WWW10 Conference, pp. 285-295, May, 2001.

[3] Honda, K., Notsu, A., and Ichihashi, H., "Collaborative filtering by sequential extraction of user-item clusters based on structural balancing approach", Fuzzy Systems, 2009. FUZZ-IEEE 2009. IEEE International Conference on, pp. 1540-1545, 2009

[4] K. Honda, A. Notsu, and H. Ichihashi. Collaborative filtering by sequential extraction of user-item clusters based on structural balancing approach. In Fuzzy Systems, 2009., pages 1540–1545, 2009.

[5] S.-M. Choi and Y.-S. Han. A content recommendation system based on category correlations. In The Fifth ICCGI, pages 1257–1260, 2010

[6] http://wordnet.princeton.edu/

[7] http://www.grouplens.org/datasets/movielens/