Placing High-Diversity Items in Top-N Recommendation Lists

Mouzhi Ge
TU Dortmund, Germany
mouzhi.ge@tu-dortmund.de

Fatih Gedikli
TU Dortmund, Germany
fatih.gedikli@tu-dortmund.de

Dietmar Jannach
TU Dortmund, Germany
dietmar.jannach@tu-dortmund.de

Abstract

Most works in the domain of recommender systems focus on providing a list of accurate recommendations. From a user perspective, users may however feel frustrated when they are facing a monotonous recommendation list. To tackle this problem, a few recent works have proposed different algorithms to generate the recommendation list with the feature of diversity. However, little research has been done to determine the placement of the high-diversity items in a recommendation list. Therefore this paper attempts to provide a guideline to appropriately place the high-diversity items in the recommendation list. Our pilot study shows that it was easier to discover high-diversity items when arranged in a block than in the case when the high-diversity items were dispersedly positioned. In addition, providing explanations may help to improve acceptance of the high-diversity items.

1 Introduction

Over the last decade, the majority of the proposed improvements to recommender system algorithms have focused on computing more accurate predictions. Accuracy in this context means the closeness between the system's predicted rating and the user's real rating for an item [Herlocker et al. 2004]. Most recommender systems are designed to provide a list of items for which the system predicts high ratings. Accordingly, improving the accuracy means that the system can predict more reliably, which items are most probably liked by the user, which in turn is assumed to increase the overall user satisfaction.

A number of recent studies have found that beyond accuracy there are other quality factors such as diversity and novelty, which are also important to users. Only concentrating on accuracy may even negatively impact the systems [McNee et al. 2006]. Therefore, different techniques were recently proposed which compute recommendation lists that take into account alternative quality factors. Previous studies have for example argued that users may feel frustrated when there is little variance in the recommendation list [Zhang and Hurley 2008, Lathia et al. 2010]. *Diversity* therefore is an important factor that recommender systems need to take into account. Furthermore, Smyth and McClave [2001] point out that diversity is as important as accuracy and it is considered as an orthogonal measure to accuracy. In general, we can observe a trade-off between diversity and accuracy, which means that increasing the diversity of a recommendation list most probably results in a decrease of its accuracy and vice versa [Adomavicius and Kwon 2011]. Regarding this trade-off, some state-of-the-art works focus on controlling the balance between accuracy and diversity [Smyth and McClave 2001, Ziegler et al. 2005] or increasing the diversity of recommendations with a minimal loss of accuracy [Zhang and Hurley 2008, Adomavicius and Kwon 2011]. In most of the proposed algorithms, diversity is measured in terms of dissimilarity of the recommended items (intra-list similarity). Intra-list-similarity is determined by measuring the similarity between all pairs of items in the recommendation list [Ziegler et al. 2005]. This measurement can be based, for example, on the known features of the items.

Note that intra-list similarity does not depend on the ordering of the items, which means that rearranging the positions of the recommended items in the list will not affect the diversity metric [Ziegler et al. 2005]. In contrast to these previous works we however conjecture that the placement of the high-diversity items in a recommendation list may affect the perceived diversity and its utility as well as the overall quality impression by the user.

Consider a scenario in which a recommender has generated a list of ten recommendations for a user. Let us assume that three of the items were introduced by the system to increase the list's diversity. If we place these three high-diversity items in the top 3 positions in this recommendation list, the user will most probably see these items first and only then the more "accurate" items. This arrangement may confuse or disappoint the user since he could have the impression that the recommender does not understand his requirements or that the recommender's predicting ability is poor. Moreover, users may stop using the recommender system right after they viewed the first few items. We find it therefore important and valuable to study how to arrange the order of recommended items. Castells et al. [2011] point out that investigating the order of recommended items is mostly missing in recommender system research. We therefore intend to study whether and to which extent different orderings affect user satisfaction and the perceived diversity and the system's overall quality.

The final objective of this work is to propose a guideline of how to organize items in a top-N recommendation list. Accordingly, we will investigate how different placements of high-diversity items affect user satisfaction and the perceived diversity. Our results are expected to indicate where the high-diversity items should be placed in a recommendation list. As such, our contributions can also be used to improve the user interface design and the general user experience in recommender systems.

This research-in-progress paper is organized as follows. In Section 2 we shortly review papers related to diversity in recommender systems. Next, in Section 3, we propose an experimental design to study the effects of different placements of high-diversity items on user satisfaction and perceived diversity. Subsequently, we describe a pilot study and summarize our initial findings. We conclude this paper by discussing our experimental design and the identified indications of how to place the high-diversity items in a recommendation list.

2 Related Works

We propose to divide the concept of diversity into *inherent* (objective) and *perceived* (subjective) diversity. Inherent diversity refers to the diversity calculated based on the dissimilarity among the recommendations and can be further classified as *individual diversity* and *aggregate diversity*. While individual diversity, also named as intra-list diversity [Castells et al. 2011], is related to the diversity of a recommendation list for an individual user, aggregate diversity, which is also termed inter-user diversity [Zhou et al. 2010], is to address the overall diversity across all users. Considering the trade-off between accuracy and diversity, some researchers [Smyth and McClave 2001, Ziegler et al. 2005, Zhang and Hurley 2008] propose algorithms to increase the individual diversity by compromising accuracy. The goal of these works is to optimize the balance between accuracy and diversity so as to keep accuracy in a certain level when increasing diversity. To increase diversity and at the same time minimize the effect on accuracy, other researchers [Zhou et al. 2010, Adomavicius and Kwon 2011] focus on increasing the aggregate diversity to solve the dilemma between accuracy and diversity. Furthermore, Fleder and Hosanagar [2007] have shown that individual diversity and aggregate diversity are not necessarily related.

Perceived diversity refers to the diversity experienced by the user and can be divided into *current perceived diversity* and *temporal perceived diversity* [Lathia et al. 2010]. Current perceived diversity means the diversity perceived by one user at a single time. In contrast, temporal perceived diversity is the diversity perceived by the user over a period of time. It can be measured, for example, by comparing the differences between two recommendation lists provided to the same user at different times [Lathia et al. 2010]. The advantage of perceived diversity is that it can capture user's opinion towards diversity. However, since different users may perceive diversity differently, we admit that it is challenging to unify the different perceived diversities. We summarize the main focus of research of previous works related to *inherent* and *perceived* diversity in Table 1.

	Inherent diversity		Perceived diversity	
	individual diversity	aggregate diversity	current perceived diversity	temporal perceived diversity
Smyth and McClave 2001	×			
Ziegler et al. 2005	×			
Fleder and Hosanagar 2007	×	×		
Zhang and Hurley 2008	×			
Zhou et al. 2010		×		
Lathia et al. 2010				×
Castells et al. 2011	×			
Adomavicius and Kwon 2011		×		
This paper			×	

Table 1 Summary of our literature review

The selected papers are arranged in a chronological order in the table. We can observe that more recent works begin to focus on studying subjective diversity. Furthermore, we found that little work has been done to investigate the *current perceived diversity*. To bridge this gap, one of our research objectives is to study the effect of the placement of high-diversity items on the *current perceived diversity*.

3 Experimental Design

In this section, we will shortly present and review the experimental setup and measurement technique used in this study. Our general goal is to find out which placement order of high-diversity elements best suits the users' needs and is well accepted by the users. We also want to find out whether and to which extent diverse elements in a recommendation list can influence the user-perceived quality of a recommender system. Therefore we decided to conduct a user study because it is hard to simulate a user's perceptions of, for example, diversity or novelty in offline experiments. For this reason, we employ a *within subjects* user study, in which each subject is confronted with *all* variations of recommendation list tested in this work.

Our experiment consists of three different screens which were displayed for several movie genres. The first screen simulates a training or learning phase in the recommendation process. In this screen the user is provided with a list of 20 movies of one specific genre. Figure 1 shows an example list for the genre *action*. The users were asked to check the movies they have watched *and* also liked. This is to give the user the impression that there is a recommender system running in the background which is trying to learn the user's movie preferences. To support the illusion of intelligent behavior and background calculations, we showed a "Calcu-

lating" message for two seconds with the hint that the recommendations are computed after the user clicked on the "Get Recommendations" button. On the second screen, a recommendation list with 12 movies was then presented to the user. This procedure (Figure 1 and 2) is carried out for action movies, romantic movies, comedy movies and animation movies.
It is important to know that in the whole experiment we do not make use of a recommender system for computing the recommendations. Instead, we manually create a static list of movies for each genre and provide it to each user. Therefore the experiment looks exactly the same for each participant and each participant is confronted with exactly the same recommendations in the same order.

Figure 1 Screen 1 - Acquiring user preferences for action movies.

In order to design different placements of high-diversity items, in the first list of action movies we show the diverse elements in one block at the end of the list, see Figure 2. In the second list of romantic movies, the diverse elements were presented in the middle of the list, again as a block. In the third list of comedy movies, the four diverse items are respectively placed at position 3, 6, 9 and 12 in the recommendation list. Finally, we use a list of animation movies as our control group containing no diverse recommendations. For the recommendation lists of action, romantic and comedy movies, we insert four high-diversity movie recommendations that do not fit into the genre-category of the corresponding list. For example, as shown in Figure 2, we add an animation movie (Toy Story 3) for kids to the recommendation list of action movies and consider the animation movie in the list of action movies as a diverse movie recommendation. Thus in our experiment the determinant of diversity is the genre difference between movies.

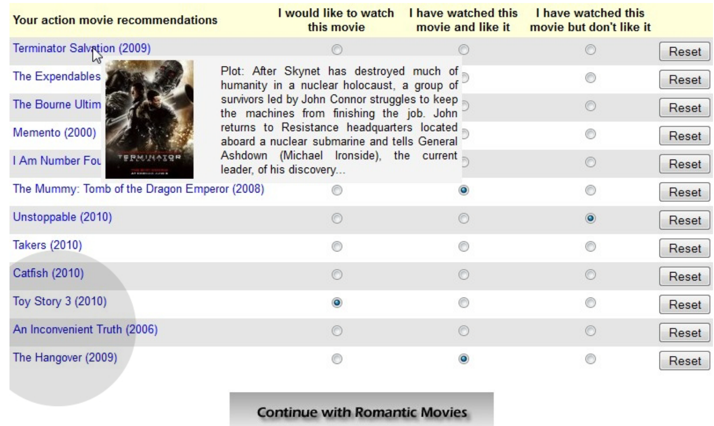

Figure 2 Screen 2 - Displaying action movie recommendations to users.

For example, Figure 2 shows the recommendation list for the category of action movies. As described above, the four movies at the bottom of the list (for example the animation movie "Toy Story 3") represent high-diversity items in the recommendation list of action movies. For each movie recommendation, the users can indicate whether they want to watch this movie; skip this movie, in case they do not like the recommendation; or indicate whether or not they like it in the case that they had already watched this movie. In the end, the user is asked to evaluate the provided recommendation list for each movie genre. On a rating scale of 1 to 5 with one point increments users could answer the following questions for each recommendation list:

- Are you satisfied with the movie recommendations? (1: not satisfied, 5: satisfied)
- Is the amount of the recommendations enough? (1: too few, 5: too many)
- Does this recommendation list surprise you? (1: not at all, 5: very surprised)
- Do you think this recommendation list is diversified? (1: not at all, 5: very diversified)

We also provided a textbox where the user could leave feedback regarding our recommendations. It took each participant between 5 and 10 minutes to complete the survey in our initial pilot study.

4 Pilot Study

As a pilot study, we invited 10 subjects to participate in the experiment. They were either working staff or students at the Technical University of Dortmund. The average age of the subjects was 29 years. 30% of the subjects were female and 70% male. All of them had little or no experience with research in recommender systems. For each subject, we supervised the whole experimental procedure. Based on this initial pilot study, we can summarize our first observations as follows:

(1) Most subjects were able to recognize that there are high-diversity items in the recommendation list. Three subjects in their feedbacks emphasized the existence of high-diversity item, for example, "*An Inconvenient Truth* is definitely not an action movie but good to know". In most cases when subjects identified high-diversity items, they further inspected the related information provided by the system (e.g. the movie plot). An interesting finding here is that some subjects were particularly interested in the high-diversity item. This indicates that although the high-diversity items may decrease the accuracy of the recommendation list, they can attract the users' attention and arouse the users' interest.
(2) For most subjects, it was easier to discover high-diversity items when arranged in a block than in the case when the high-diversity items were dispersedly positioned. This can be observed from the response of the participants. When subjects were facing a recommendation list with a block of high-diversity items, they sometimes directly told us that these items are not fitting into the recommendation list. However, this did not take place when the high-diversity items were dispersedly positioned. Therefore, we can propose a hypothesis that under the same individual diversity, a recommendation list containing a block of high-diversity items is perceived more diverse than one with dispersedly positioned high-diversity items.
(3) An explanation facility may help to increase the user acceptance of high-diversity items. In our pilot study, most subjects were particularly interested in reading the plot information and external links for the high-diversity movies. This indicates that when facing a high-diversity item, users hope to find out why the system recommends this item. Therefore we infer that if a system provides diverse recommendations but without corresponding explanations, this may decrease user satisfaction. We thus propose a hypothesis to be tested in our ongoing work that an explanation facility can increase the acceptance of high-diversity items.

5 Discussion and concluding remarks

In this research-in-progress work, we reported the preliminary observations from the pilot study of our experiment. As mentioned in the section of experimental design, we present four recommendation lists to subjects and observe their responses via changing the positions of the high-diversity items. In order to assure the effect is only from the varied positions, we need to keep the variance effect of accuracy, diversity and novelty between the four recommendation lists at a minimum. Note that our aim is to provide four recommendation lists at the same level, rather than providing high-quality recommendations. Thus, randomly choosing the popular movies for the four recommendation lists allows us to keep their accuracy at the same level. Furthermore since the total number of recommendations and the number of high-diversity recommendations are the same in the four recommendation lists, we can keep the four lists at the same diversity level. Also, considering the dependency between novelty and diversity, we control the mean and standard deviation of movie release years in the four recommendation lists at the same level. In our forthcoming work we will report more details of our experiment and data analysis results.

References

[Adomavicius and Kwon 2011] Gediminas Adomavicius, YoungOk Kwon. Improving Aggregate Recommendation Diversity Using Ranking-Based Techniques. IEEE Transactions on Knowledge and Data Engineering (forthcoming).

[Castells et al. 2011] Pablo Castells1, Saúl Vargas, Jun Wang. Novelty and Diversity Metrics for Recommender Systems: Choice, Discovery and Relevance. In *Proceedings of International Workshop on Diversity in Document Retrieval (DDR)*. Dublin, Ireland: 29-37.

[Fleder and Hosanagar 2007] Daniel Fleder, Kartik Hosanagar. Recommender Systems and their Impact on Sales Diversity. In *Proceedings of the 8th ACM Conference on Electronic Commerce*. San Diego, CA, USA: 192-199.

[Herlocker et al. 2004] Jonathan L. Herlocker, Joseph A. Konstan, Loren G. Terveen, John Riedl. Evaluating Collaborative Filtering Recommender Systems. *ACM Transactions on Information Systems* 22(1): 5-53.

[Lathia et al. 2010] Neal Lathiax, Stephen Hailesx, Licia Caprax, Xavier Amatriain. Temporal Diversity in Recommender Systems. *Proceeding of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. Geneva, Switzerland: 210-217.

[McNee et al. 2006] Sean M. McNee, John Riedl, Joseph A. Konstan. Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems. In *Proceedings of the ACM SIGCHI Conference on Human Factors in Computing Systems*. Montréal, Canada: 1097-1101.

[Smyth and McClave 2001] Barry Smyth, Paul McClave. Similarity vs. Diversity. In *Proceedings of 4th International Conference. on Case-Based Reasoning*. UK: 348-361.

[Zhou et al. 2010] Tao Zhoua, Zoltán Kuscsika, Jian-Guo Liua, Matúš Medoa, Joseph Rushton Wakelinga, Yi-Cheng Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *National Academy of Sciences of the USA.* 107(10): 4511-4515.

[Zhang and Hurley 2008] Mi Zhang, Neil Hurley. Avoiding Monotony: Improving the Diversity of Recommendation Lists. In *Proceedings of the 2nd ACM Recommender Systems*, Lausanne, Switzerland: 123-130.

[Ziegler et al. 2005] Cai-Nicolas Ziegler, Sean McNee, Joseph Konstan, Georg Lausen. Improving Recommendation Lists through Topic Diversification. In *Proceedings of the 14th WWW conference*. Chiba, Japan. 22-32.