TOWARD MORE DIVERSE RECOMMENDATIONS: ITEM RE-RANKING METHODS FOR RECOMMENDER SYSTEMS

Gediminas Adomavicius **YoungOk Kwon**
Department of Information and Decision Sciences
Carlson School of Management, University of Minnesota
gedas@umn.edu, kwonx052@umn.edu

Abstract

Recommender systems are becoming increasingly important to individual users and businesses for providing personalized recommendations. However, while the majority of algorithms proposed in recommender systems literature have focused on improving recommendation accuracy (as exemplified by the recent Netflix Prize competition), other important aspects of recommendation quality, such as the diversity of recommendations, have often been overlooked. In this paper, we introduce a number of item re-ranking methods that can generate substantially more diverse recommendations across all users while maintaining comparable levels of recommendation accuracy. Empirical results consistently show the diversity gains of the proposed re-ranking methods for several real-world rating datasets and different rating prediction techniques.

Keywords: recommender systems, collaborative filtering, recommendation diversity, ranking functions.

1. Introduction and Motivation

In recent years, recommender systems have become an important research topic in academia and industry (Adomavicius and Tuzhilin 2005). However, in most cases, new techniques are designed to improve the accuracy of recommendations, including the most recent algorithms from Netflix Prize competition (netflixprize.com); other aspects, such as the *diversity* of recommendations, have often been overlooked in evaluating the recommendation quality. The importance of diverse recommendations has been emphasized in several recent studies (Adomavicius and Kwon 2008, Bradley and Smyth 2001, Brynjolfsson et al. 2007, Fleder and Hosanagar 2009, Zhang and Hurley 2008, Ziegler et al. 2005). These studies argue that one of the goals of recommender systems is to provide a user with highly idiosyncratic or personalized items, and more diverse recommendations result in more opportunities for users to get recommended such items. With this motivation, some studies proposed new recommendation methods that can increase the diversity of recommendation sets for a given individual user (i.e., *individual* diversity), often measured by the average dissimilarity between all pairs of recommended items (Bradley and Smyth 2001, Zhang and Hurley 2008, Ziegler et al. 2005).

More diverse recommendations could be beneficial for some businesses as well (Brynjolfsson et al. 2007, Fleder and Hosanagar 2009). For example, it would be profitable to Netflix if their recommender system can encourage users to rent more "long-tail" type of movies (i.e., more obscure items that are located in the tail of the sales distribution) because they are typically less costly to license and acquire from distributors than new-release or highly-popular movies of big studios (Goldstein and Goldstein 2006). Few recent studies (Brynjolfsson et al. 2007, Fleder and Hosanagar 2009) started examining the impact of recommender systems on sales diversity by considering *aggregate* diversity of recommendations across all users, which will be the focus of this paper. Note that high individual diversity of recommendations does not necessarily imply high aggregate diversity. For example, while recommending to all users the same five best-selling items that are not similar to each other will result in high individual diversity, the aggregate diversity will

Table 1. Accuracy-diversity tradeoff: example

Quality Metric: / Top-1 recommendation of:	Accuracy	Diversity
Popular Item (item with the largest number of known ratings)	82%	49 distinct items
"Long-Tail" Item (item with the smallest number of known ratings)	68%	695 distinct items

Note. Recommendations for 2828 users by a standard item-based collaborative filtering technique on MovieLens data.

be very low (because only five distinct items are recommended across all users).

Higher diversity (both individual and aggregate), however, can come at the expense of accuracy. The example in Table 1 (based on the MovieLens dataset that is publicly available at grouplens.org) shows that it is possible to obtain higher diversity simply by recommending less popular items; however, the loss of recommendation accuracy in this case can be substantial. Some prior work (Adomavicius and Kwon 2008) has attempted to overcome this accuracy-diversity trade-off by filtering out less promising recommendations; as a result, however, such approaches often can offer only a fraction of possible recommendations to users. In contrast, we explore new recommendation approaches that can increase the diversity of recommendations with only a minimal (negligible) accuracy loss using different recommendation *ranking* techniques, without losing any recommendations. While the traditional recommender systems typically rank the relevant items by their predicted rating and recommend the most highly predicted item to each user, resulting in high accuracy, the proposed approaches consider additional factors, such as item popularity, when ranking the recommended item list to substantially increase recommendation diversity while maintaining comparable levels of accuracy.

2. Related Work

Recommender systems typically operate in a two dimensional space of users and items. Let U be the set of users of a recommender system, and let I be the set of all possible items that can be recommended to users. Then, the utility function that represents the preference of item $i \in I$ by user $u \in U$ is often defined as $R: U \times I \rightarrow Rating$, where *Rating* typically represents some numeric scale used by the users to evaluate each item. Also, we use the $R(u, i)$ notation to represent a known rating (i.e., the actual rating that user u gave to item i), and the $R^*(u, i)$ notation to represent the system-predicted rating for item i and user u.

Numerous metrics have been employed for measuring the accuracy of recommendations (Herlocker et al. 2004). In particular, precision is one of the most popular decision-support metrics that measures the percentage of truly "high" ratings among those that were predicted to be "high" by the recommender system. The ratings in the data that we used for our experiments are integers between 1 and 5, inclusive, and accordingly we define the items with ratings greater than 3.5 (threshold for "high" ratings, denoted by T_H) as "highly-ranked", and the ratings less than 3.5 as "non-highly-ranked." Furthermore, in real world settings, recommender systems typically recommend the most highly-ranked N items since users are usually interested in only several most relevant recommendations, and this list of N items for user u can be defined as $L_N(u) = \{i_1, \ldots, i_N\}$, where $R^*(u, i_k) \geq T_H$ for all $k \in \{1, 2, .., N\}$. Therefore, in our paper, we evaluate the recommendation accuracy based on the percentage of truly "highly-ranked" ratings, denoted by $correct(L_N(u))$, among those that were predicted to be the N most relevant "highly ranked" items for each user, i.e., using the *precision-in-top-N* metric (Herlocker et al. 2004), as written formally as follows:

$$precision\text{-}in\text{-}top\text{-}N = \sum_{u \in U} |\, correct(L_N(u)) \,| \Big/ \sum_{u \in U} |\, L_N(u) \,|,$$

where $correct(L_N(u)) = \{i \in L_N(u) \mid R(u, i) \geq T_H\}$. However, as mentioned earlier, relying on the accuracy of recommendations alone may not be enough to find the most relevant items for a user (McNee et al. 2006), and another important aspect can be measured by the *diversity* of recommendations. Since we intend to measure the recommendation quality based on the top-N recommendation lists that the system provides to its users, in this paper we use the total number of distinct items recommended across all users as an aggregate diversity measure, which we will refer to as *diversity-in-top-N* and can formally express as:

$$diversity\text{-}in\text{-}top\text{-}N = \left| \bigcup_{u \in U} L_N(u) \right|.$$

3. Improving Diversity By Re-Ranking Recommendation List

As mentioned earlier, traditional recommender systems recommend to user u a list of top-N items, $L_N(u)$, selected according to some ranking criterion. More formally, item i_x is ranked ahead of item i_y (i.e., $i_x \prec i_y$) if $rank(i_x) < rank(i_y)$, where $rank$: $I \rightarrow \mathbf{R}$ is a function representing the ranking criterion. The vast majority of current recommender systems use the predicted rating value as the ranking criterion (i.e., $rank_{\text{Standard}}$), and it shares the motivation with the widely accepted *probability ranking principle* in

information retrieval systems literature that ranks the documents in order of decreasing probability of relevance (Robertson 1997). Note that, by definition, recommending the most highly predicted items is designed to help improve recommendation accuracy, but not recommendation diversity. Therefore, new ranking criteria are needed in order to achieve diversity improvement. Recommending less popular items intuitively should have an effect towards increasing recommendation diversity, as illustrated by the example in Section 1. Following this motivation, we explore the possibility to use *item popularity* as a recommendation ranking criterion and its impact on the recommendation accuracy and diversity.

3.1 Re-Ranking Recommendation List By Item Popularity

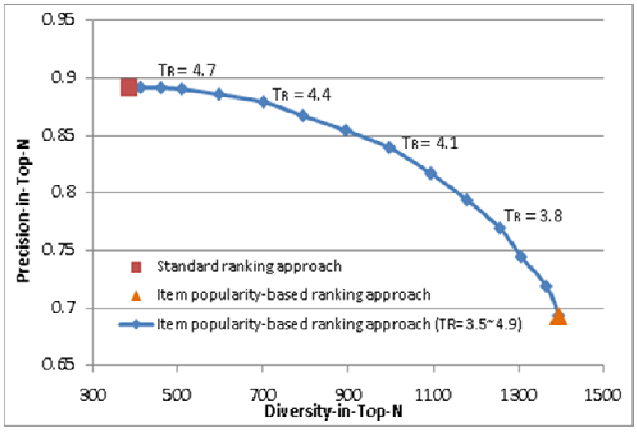
We define item popularity as the number of known ratings for each item, and item popularity-based ranking approach (denoted as $rank_{\text{ItemPop}}$) recommends the least popular items to users. The results in Figure 1 show that the item popularity-based ranking approach increased recommendation diversity from 385 to 1395 (i.e., 3.6 times); however, recommendation accuracy dropped from 89% to 69%, as compared to the standard ranking approach. Here, despite the significant diversity gain, such a significant accuracy loss (20%) would not be acceptable in most real-life personalization applications. Therefore, in the next subsection we introduce a general technique to parameterize recommendation ranking approaches, which allows to achieve significant diversity gains while controlling accuracy losses (e.g., according to how much loss is tolerable in a given application).

3.2 Controlling Accuracy-Diversity Trade-Off: Parameterized Ranking Approaches

Item popularity-based ranking approach as well as other ranking approaches proposed in this paper are parameterized with "ranking threshold" $T_R \in [T_H, T_{\max}]$ (where $T_{\max}$ is the largest possible rating on the rating scale, e.g., $T_{\max}=5$) to provide user the ability to choose a certain level of recommendation accuracy. In particular, given any ranking function $rank_X(i)$, the ranking threshold T_R is used for creating the parameterized version of this ranking function, $rank_X(i, T_R)$, which is formally defined as:

$$rank_X(i,T_R) = \begin{cases} rank_X(i), & \text{if } R^*(u,i) \in [T_R, T_{\max}] \\ \alpha_u + rank_{\text{Standard}}(i), & \text{if } R^*(u,i) \in [T_H, T_R) \end{cases}, \quad \text{where } I_u^*(T_R) = \{i \in I \mid R^*(u,i) \geq T_R\}, \; \alpha_u = \max_{i \in I_u^*(T_R)} rank_X(i)$$

Simply put, items that are predicted above ranking threshold T_R are ranked according to $rank_X(i)$, while items that are below T_R are ranked according to the standard ranking approach $rank_{\text{Standard}}(i)$. In addition, all items that are above T_R get ranked ahead of all items that are below T_R (as ensured by α_u in the above formal definition). Therefore, choosing different T_R values in-between T_H and $T_{\max}$ allows the user to set the desired balance between accuracy and diversity, as shown in Figure 1. For example, the item popularity-based ranking approach with ranking threshold 4.4 could minimize the accuracy loss to 1.32%, but still could obtain 83% diversity gain (from 385 to 703), compared to the standard ranking approach. An even higher threshold 4.7 still makes it possible to achieve 20% diversity gain (from 385 to 462) with only 0.06% of accuracy loss. Also note that, even when there are less than N items above the ranking threshold T_R, by definition, *all* the items above T_R are recommended to a user, and the remaining top-N items are selected according to the standard ranking approach. This ensures that all the ranking approaches proposed in this paper provide the same exact number of recommendations as their corresponding baseline technique, which is also very important from the experimental analysis point of view in order to have a fair performance comparison of different ranking techniques.



MovieLens data, top-5 items, item-based CF, 50 neighbors

Figure 1. Performance of item popularity-based approach with its parameterized versions

3.3 Additional Ranking Approaches

We here introduce four additional ranking approaches that can be used as alternatives to $rank_{Standard}$ to improve recommendation diversity, and the formal definitions of each ranking approach as well as standard and item popularity-based ranking approaches are provided in Figure 2. As seen from the empirical analysis on the positive relationships between average item popularity and predicted ratings in (Adomavicius and Kwon 2008), we also consistently observed that popular items, on average, are likely to have higher predicted ratings than less popular items, using different traditional recommendation techniques. Therefore, it can be suggested that recommending not as highly predicted items (but still predicted to be above T_H) likely implies recommending, on average, less popular items, potentially leading to diversity improvements; following this observation, we propose to use *predicted rating value* itself as an item ranking criterion. Based on similar empirical observations, we propose a number of alternative ranking approaches, including the ones based on *average rating*, *absolute likeability*, and *relative likeability*, as defined in Figure 2.

4. Empirical Results

The proposed recommendation ranking approaches were tested with MovieLens (grouplens.org) and Netflix (netflixprize.com) data sets. We pre-processed each dataset to include users and movies with significant rating history, which makes it possible to have sufficient number of highly-predicted items for recommendations to each user (in the test data). For each dataset, we randomly chose 60% of the ratings as training data and used them to predict the remaining 40% (i.e., test data).

4.1 Performance of Proposed Ranking Approaches

All five proposed ranking approaches were used in conjunction with one of three widely popular recommendation techniques for rating prediction, including two heuristic-based (user-based and item-based "nearest neighbor") and one model-based (matrix factorization) collaborative filtering (CF) techniques (Sarwar et al. 2001, Funk 2006), to generate top-N (N=1, 5, 10) recommendations to each user. We set predicted rating threshold as T_H = 3.5 (out of 5) to ensure that only relevant items are recommended to users, and ranking threshold T_R varied from 3.5 to 5.0. The performance of each ranking approach was measured in terms of *precision-in-top-N* and *diversity-in-top-N* and, for comparison purposes, its diversity gain and precision loss with respect to the standard ranking approach was calculated. Consistently with the accuracy-diversity tradeoff discussed in the introduction, all the proposed ranking approaches improved the diversity of recommendations by sacrificing the accuracy. However, with each ranking approach, as ranking threshold T_R increases, the accuracy loss is significantly minimized (smaller precision loss) while still exhibiting substantial diversity improvement. Therefore, with different ranking thresholds, one can obtain different diversity gains for different levels of tolerable precision loss, as compared to the standard ranking approach. Following this idea, in our experiments we compare the effectiveness (i.e., diversity gain) of different ranking techniques for various precision loss levels (0.1-10%)**.**

- ***Standard,*** i.e., ranking the candidate (highly predicted) items by their predicted rating value, from highest to lowest.
  $rank_{\text{Standard}}(i)=R^*(u, i)^{-1}$.
- ***Item Popularity,*** i.e., ranking by item popularity, from lowest to highest, where popularity is represented by the number of known ratings that each item has .
  $rank_{\text{ItemPop}}(i)= |U(i)|, U(i) = \{u \in U \mid \exists R(u, i)\}$.
- ***Reverse Predicted Rating Value***, i.e., ranking by predicted rating value, from lowest to highest.
  $rank_{\text{RevPred}}(i) = R^*(u,i)$.
- ***Item Average Rating***, i.e., ranking by an average of all known ratings for each item:
  $rank_{\text{AvgRating}}(i) = \overline{R(i)} = \sum_{u \in U(i)} R(u,i) \Big/ |U(i)|$.
- ***Item Absolute Likeability***, i.e., ranking by how many users liked the item (i.e., rated it above T_H):
  $rank_{\text{AbsLike}}(i) = |U_H(i)|, U_H(i)=\{u \in U(i) \mid R(u,i) \geq T_H\}$.
- ***Item Relative Likeability***, i.e., ranking by the percentage of the users who liked an item (among all users who rated it).
  $rank_{\text{RelLike}}(i) = |U_H(i)| / |U(i)|$.

Figure 2. Various ranking functions

Table 2. Diversity gains of proposed ranking approaches for different levels of precision loss

	Item Popularity		Reverse Prediction		Item Avg Rating		Item Abs Likeability		Item Rel Likeability	
Precision Loss	Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain	
-0.1	+800	3.078	+848	3.203	+975	3.532	+897	3.330	+937	3.434
-0.05	+594	2.543	+594	2.543	+728	2.891	+642	2.668	+699	2.816
-0.025	+411	2.068	+411	2.068	+513	2.332	+445	2.156	+484	2.257
-0.01	+270	1.701	+234	1.608	+311	1.808	+282	1.732	+278	1.722
-0.005	+189	1.491	+173	1.449	+223	1.579	+196	1.509	+199	1.517
-0.001	+93	1.242	+44	1.114	+78	1.203	+104	1.270	+96	1.249
Standard:0.892	385	1.000	385	1.000	385	1.000	385	1.000	385	1.000

(a) MovieLens dataset, top-5 items, heuristic-based technique (item-based CF, 50 neighbors)

	Item Popularity		Reverse Prediction		Item Avg Rating		Item Abs Likeability		Item Rel Likeability	
Precision Loss	Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain	
-0.1	+314	1.356	+962	2.091	+880	1.998	+732	1.830	+860	1.975
-0.05	+301	1.341	+757	1.858	+718	1.814	+614	1.696	+695	1.788
-0.025	+238	1.270	+568	1.644	+535	1.607	+464	1.526	+542	1.615
-0.01	+156	1.177	+363	1.412	+382	1.433	+300	1.340	+385	1.437
-0.005	+128	1.145	+264	1.299	+282	1.320	+247	1.280	+288	1.327
-0.001	+64	1.073	+177	1.201	+118	1.134	+89	1.101	+148	1.168
Standard:0.834	882	1.000	882	1.000	882	1.000	882	1.000	882	1.000

(b) Netflix dataset, top-5 items, model-based technique (matrix factorization CF, K=64)

Notation: Precision Loss=[Precision-in-top-N of proposed ranking approach]–[Precision-in-top-N of standard ranking approach]
Diversity Gain (column 1)=[Diversity-in-top-N of proposed ranking approach]–[Diversity-in-top-N of standard ranking approach]
Diversity Gain (column 2)=[Diversity-in-top-N of proposed ranking approach]/[Diversity-in-top-N of standard ranking approach]

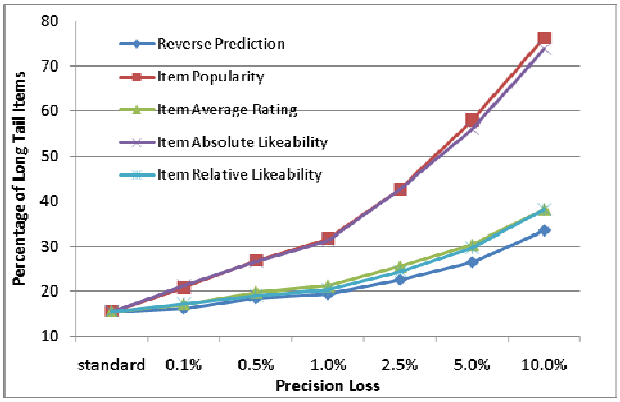
While the results were very consistent across all experiments, for illustration purposes and because of the space limitations, we present two experiments in Table 2: each using all possible ranking techniques on a different dataset and a different recommendation technique. For example, Table 2a shows the performance of the proposed ranking approaches used in conjunction with item-based CF technique to provide top-5 recommendations on the MovieLens dataset. For instance, one can observe that, with the precision loss of *only* 0.001 or 0.1% (from 0.892 of the standard ranking approach), item average rating-based ranking approach can already increase diversity by 20% (i.e., absolute gain of 78 on top of the 385 achieved by the standard ranking approach). If users can tolerate precision loss up to 1%, the diversity could be increased by 81% with the same ranking technique; and 5% precision loss can provide diversity gains up to 189% for this recommendation technique on this dataset. Substantial diversity improvements can be observed across different settings and, thus, system designers have the flexibility to choose the most desirable ranking approach based on the data in a given application.

4.2 Impact of Proposed Ranking Approaches on the Distribution of Recommended Items

Since we measure recommendation diversity as the total number of distinct items that are being recommended across all users, one could possibly argue that, while the diversity can be easily improved by recommending a few new items to some users, it may not be clear whether the proposed ranking approaches would be able to shift the overall *distribution* of recommended items towards more idiosyncratic, long-tail recommendations. Therefore, in this subsection we explore how the proposed ranking approaches change the actual distribution of recommended items in terms of their popularity. Following the popular "80-20 rule" or the Pareto principle, we define the top 20% of the most frequently rated items in the training dataset as "bestsellers" and the remaining 80% of items as "long-tail" items. We calculated the percentage of long-tail items among the items recommended across all users by the proposed ranking approaches as well as by the standard ranking approach.

For example, with the standard ranking approach, the long-tail items consist of only 16% of total recommendations (i.e., 84% recommendations were of bestsellers) when recommending top-5 items to each user using item-based CF technique on MovieLens dataset, confirming some findings in prior literature that recommender systems often gravitate towards recommending bestsellers and not long-tail items (Fleder and Hosanagar 2009). However, as shown in Figure 3, the proposed ranking approaches are able to recommend significantly more long-tail items with a small level of accuracy loss, and this distribution becomes even more skewed towards long-tail items if more accuracy loss can be tolerated.

For example, with 1% precision loss, the percentage of recommended long-tail items increased from 16% to 21% with item average rating-based ranking approach, or to 32% with item popularity-based ranking approach. And with 5% precision loss, the proportion of long-tail items can grow up to 58%, using the latter ranking approach. This analysis provides further empirical support that the proposed ranking approaches increase not just the number of distinct items recommended, but also the proportion of recommended long-tail items, thus, confirming that the proposed techniques truly contribute towards more diverse and idiosyncratic recommendations across all users.



MovieLens data, top 5 items, item-based CF, 50 Neighbors

Percentage of Long Tail Items = Percentage of recommended items that are not among top 20% most popular items

Figure 3. Proportion of long-tail items

5. Conclusions and Future Work

We believe that the diversity of recommendations should be given more weight in evaluating the recommendation quality, and more research is needed to further explore the tradeoff between accuracy and diversity in recommender systems. In this paper, we proposed new ranking approaches that can significantly increase diversity with a small amount of accuracy loss by re-ranking candidate items (that have traditionally been ranked by recommender systems according to the predicted rating values) based on several alternative item ranking criteria. The experimental results show that, at the expense of a negligible accuracy loss, recommendation ranking approaches represent effective techniques for obtaining the desired levels of diversity. As a possible direction for future research, we plan to investigate how the improvements in the recommendation diversity will affect users' behavior in e-commerce settings.

Acknowledgement

This research was supported in part by the National Science Foundation grant IIS-0546443.

References

Adomavicius, G., Y. Kwon. "Overcoming Accuracy-Diversity Tradeoff in Recommender Systems: A Variance-Based Approach," *Proceedings of the 18th Workshop on Information Technology and Systems*, 2008.

Adomavicius, G., A. Tuzhilin. "Toward the Next Generation of Recommender Systems: A Survey of the State-of-the-Art and Possible Extensions," *IEEE Trans. on Knowledge and Data Engineering* (17:6), 2005, pp. 734-749.

Bradley, K., B. Smyth. "Improving Recommendation Diversity," *Proceedings of 12th Irish Conference on Artificial Intelligence and Cognitive Science*, 2001, pp. 85-94.

Brynjolfsson, E., Y. J. Hu, D. Simester. "Goodbye Pareto Principle, Hello Long Tail: The Effect of Search Costs on the Concentration of Product Sales," *NET Institute Working Paper*, 2007.

Fleder, D., K. Hosanagar. "Blockbuster Culture's Next Rise Or Fall: The Impact of Recommender Systems on Sales Diversity," *Management Science* (55:5), 2009, pp. 697-712.

Funk, S. "Netflix Update: Try This At Home", http://sifter.org/~simon/journal/20061211.html, 2006.

Goldstein, D. G., D. C. Goldstein. "Profiting from the Long Tail," *Harvard Business Review* (84:6), 2006, pp. 24-28.

Herlocker, J. L., J. A. Konstan, L. G. Terveen, J. T. Riedl. "Evaluating Collaborative Filtering Recommender Systems," *ACM Transactions on Information Systems (TOIS)* (22:1), 2004, pp. 5-53.

McNee, S. M., J. Riedl, J. A. Konstan. "Being Accurate Is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems," *Proc. of the Conf. on Human Factors in Computing Systems*, 2006, pp. 1097-1101.

Robertson, S. E. "The Probability Ranking Principle in IR," *Readings in Information Retrieval*, 1997, pp. 281-286.

Sarwar, B., G. Karypis, J. Konstan, J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," *Proceedings of the 10th International World Wide Web Conference*, 2001, pp. 285-295.

Zhang, M., N. Hurley. "Avoiding Monotony: Improving the Diversity of Recommendation Lists," *Proceedings of the 2008 ACM Conference on Recommender Systems*, 2008, pp. 123-130.

Ziegler, C. N., S. M. McNee, J. A. Konstan, G. Lausen. "Improving Recommendation Lists through Topic Diversification," *Proceedings of the 14th International World Wide Web Conference*, 2005, pp. 22-32.