Improving Recommendation Diversity Using Ranking-Based Techniques

Gediminas Adomavicius and 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 and explore a number of item ranking techniques that can generate substantially more diverse recommendations across all users while maintaining comparable levels of recommendation accuracy. Comprehensive empirical evaluation consistently shows the diversity gains of the proposed techniques using several real-world rating datasets and different rating prediction algorithms.

Keywords— Recommender systems, recommendation diversity, ranking functions, evaluation of recommender systems.

1 INTRODUCTION

In the current age of information overload, it is becoming increasingly harder to find relevant content. This problem is not only widespread but also alarming [27]. Over the last 10-15 years, recommender systems technologies have been introduced to help people deal with these vast amounts of information [1], [7], [9], [29], [36], [40], and they have been widely used in research as well as e-commerce applications, such as the ones used by Amazon and Netflix.

The most common formulation of the recommendation problem relies on the notion of ratings, i.e., recommender systems estimate ratings of items (or products) that are yet to be consumed by users, based on the ratings of items already consumed. Recommender systems typically try to predict the ratings of unknown items for each user, often using other users' ratings, and recommend top *N* items with the highest predicted ratings. Accordingly, there have been many studies on developing new algorithms that can improve the predictive accuracy of recommendations. However, the quality of recommendations can be evaluated along a number of dimensions, and relying on the accuracy of recommendations alone may not be enough to find the most relevant

items for each user [23], [31]. In particular, the importance of *diverse* recommendations has been previously emphasized in several studies [8], [10], [15], [32], [46], [54], [57]. 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, often measured by an average dissimilarity between all pairs of recommended items, while maintaining an acceptable level of accuracy [8], [32], [46], [54], [57]. These studies measure recommendation diversity from an individual user's perspective (i.e., *individual diversity*).

In contrast to individual diversity, which has been explored in a number of papers, some recent studies [10], [15] started examining the impact of recommender systems on sales diversity by considering *aggregate diversity* of recommendations across all users. Note that high individual diversity of recommendations does not necessarily imply high aggregate diversity. For example, if the system recommends to all users the same five best-selling items that are not similar to each other, the recommendation list for each user is diverse (i.e., high individual diversity), but only five distinct items are recommended to all users and purchased by them (i.e., resulting in low aggregate diversity or high sales concentration).

While more diverse recommendations would be helpful for individual users, they could be beneficial for some businesses as well [10], [11], [15], [20]. For example, it would be profitable to Netflix if the recommender systems can encourage users to rent "long-tail" type of movies (i.e., more obscure items that are located in the tail of the sales distribution [2]) because they are less costly to license and acquire from distributors than new-release or highly-popular movies of big studios [20]. However, the impact of recommender systems on aggregate diversity in real-world e-commerce applications has not been well-understood. For example, one study [10], using data from online clothing retailer, confirms the "long tail" phenomenon that refers to the increase in the tail of the sales distribution (i.e., the increase in aggregate diversity) attributable to the usage of the recommender system. On the other hand, another study [15] shows a contradictory finding that recommender systems actually can reduce the aggregate diversity in sales. This can be explained by the fact that the idiosyncratic items often have limited historical data and, thus, are more difficult to recommend to users; in contrast, popular items typically have more ratings and, therefore, can be recommended to more users. For example, in the context of Net-

flix Prize competition [6], [22], there is some evidence that, since recommender systems seek to find the common items (among thousands of possible movies) that two users have watched, these systems inherently tend to avoid extremes and recommend very relevant but safe recommendations to users [51].

As seen from this recent debate, there is a growing awareness of the importance of aggregate diversity in recommender systems. Furthermore, while, as mentioned earlier, there has been significant amount of work done on improving individual diversity, the issue of aggregate diversity in recommender systems has been largely untouched. Therefore, in this paper, we focus on developing algorithmic techniques for improving aggregate diversity of recommendations (which we will simply refer to as *diversity* throughout the paper, unless explicitly specified otherwise), which can be intuitively measured by the number of distinct items recommended across all users.

Higher diversity (both individual and aggregate), however, can come at the expense of accuracy. As known well, there is a tradeoff between accuracy and diversity because high accuracy may often be obtained by safely recommending to users the most popular items, which can clearly lead to the reduction in diversity, i.e., less personalized recommendations [8], [32], [46]. And conversely, higher diversity can be achieved by trying to uncover and recommend highly idiosyncratic or personalized items for each user, which often have less data and are inherently more difficult to predict, and, thus, may lead to a decrease in recommendation accuracy.

Table 1 illustrates an example of accuracy and diversity tradeoff in two extreme cases where only popular items or long-tail type items are recommended to users, using MovieLens rating dataset (datasets used in this paper are discussed in Section 5.1). In this example, we used a

TABLE 1. ACCURACY-DIVERSITY TRADEOFF: EMPIRICAL 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 (top-1 item for each user) are generated for 2828 users among the items that are predicted above the acceptable threshold 3.5 (out of 5), using a standard item-based collaborative filtering technique with 50 neighbors on the MovieLens Dataset.

popular recommendation technique, i.e., neighborhood-based collaborative filtering (CF) technique [9], to predict unknown ratings. Then, as candidate recommendations for each user, we considered only the items that were predicted above the pre-defined rating threshold to assure the acceptable level of accuracy, as is typically done in recommender systems. Among these candidate items for each user, we identified the item that was rated by most users (i.e., the item with the largest number of known ratings) as a *popular item*, and the item that was rated by least number of users (i.e., the item with the smallest number of known ratings) as a *long-tail item*. As illustrated by Table 1, if the system recommends each user the most popular item (among the ones that had a sufficiently high predicted rating), it is much more likely for many users to get the same recommendation (e.g., the best-selling item). The accuracy measured by precision-in-top-1 metric (i.e., the percentage of truly "high" ratings among those that were predicted to be "high" by the recommender system) is 82%, but only 49 popular items out of approximately 2000 available distinct items are recommended across all users. The system can improve the diversity of recommendations from 49 up to 695 (a 14-fold increase) by recommending the long-tail item to each user (i.e., the least popular item among highly-predicted items for each user) instead of the popular item. However, high diversity in this case is obtained at the significant expense of accuracy, i.e., drop from 82% to 68%.

The above example 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. In this paper, we explore new recommendation approaches that can increase the diversity of recommendations with only a minimal (negligible) accuracy loss using different recommendation *ranking* techniques. In particular, traditional recommender systems typically rank the relevant items in a descending order of their predicted ratings and recommend top N items to each user, resulting in high accuracy. In contrast, the proposed approaches consider additional factors, such as item popularity, when ranking the recommendation list to substantially increase recommendation diversity while maintaining comparable levels of accuracy. This paper provides a comprehensive empirical evaluation of the proposed approaches, where we test them with various datasets in a variety of different settings. For example, the best results show up to 20-25% diversity gain with only 0.1% accuracy loss, up to 60-80% gain with 1% accuracy loss, and even substantially higher diversity improvements (e.g., up to 250%) if some users are willing to tolerate higher accuracy loss.

In addition to providing significant diversity gains, the proposed ranking techniques have several other advantageous characteristics. In particular, these techniques are extremely *efficient*, because they are based on scalable sorting-based heuristics, as well as *parameterizable*, since the user has the control to choose the acceptable level of accuracy for which the diversity will be maximized. Also, the proposed ranking techniques provide a *flexible* solution to improving recommendation diversity, because they are applied *after* the unknown item ratings have been estimated and, thus, can achieve diversity gains in conjunction with a number of different rating prediction techniques, as illustrated in the paper.

The remainder of the paper is organized as follows. Section 2 reviews relevant literature on traditional recommendation algorithms and the evaluation of recommendation quality (including accuracy and diversity of recommendations). Section 3 describes our motivations for alternative recommendation ranking techniques, such as item popularity. We then propose several additional ranking techniques in Section 4, and the main empirical results follow in Section 5. Additional experiments are conducted to further explore the proposed ranking techniques in Section 6. Lastly, Section 7 concludes the paper by summarizing the contributions and future directions.

2 Related Work

2.1 Recommendation Techniques for Rating Prediction

Recommender systems are usually classified into three categories based on their approach to recommendation: content-based, collaborative, and hybrid approaches [1], [3]. Content-based recommender systems recommend items similar to the ones the user preferred in the past. Collaborative filtering (CF) recommender systems recommend items that users with similar preferences (i.e., "neighbors") have liked in the past. Finally, hybrid approaches can combine content-based and collaborative methods in several different ways. Recommender systems can also be classified based on the nature of their algorithmic technique into heuristic (or memory-based) and model-based approaches [1], [9]. Heuristic techniques typically calculate recommendations based directly on the previous user activities (e.g., transactional data or rating values). One of the commonly used heuristic techniques is a neighborhood-based approach that finds nearest neighbors that have tastes similar to those of the target user [9], [14], [33], [36], [41]. In contrast, model-based techniques use previous user activities to first learn a predictive model, typically using some statistical or machine-learning methods, which is then used to make recom-

mendations. Examples of such techniques include Bayesian clustering, aspect model, flexible mixture model, matrix factorization, and other methods [4], [5], [9], [24], [44], [49].

In real world settings, recommender systems generally perform the following two tasks in order to provide recommendations to each user. First, the ratings of unrated items are estimated based on the available information (typically using known user ratings and possibly also information about item content or user demographics) using some recommendation algorithm. And second, the system finds items that maximize the user's utility based on the predicted ratings, and recommends them to the user. Ranking approaches proposed in this paper are designed to improve the recommendation diversity in the second task of finding the best items for each user.

Because of the decomposition of the rating estimation and recommendation ranking tasks, our proposed ranking approaches provide a flexible solution in that they can be used in conjunction with *any* available rating estimation algorithm. In our experiments, to illustrate the broad applicability of the proposed recommendation ranking approaches, we used them in conjunction with two of the most popular and widely employed CF techniques for rating prediction: a heuristic neighborhood-based technique and a model-based matrix factorization technique. We provide a brief overview of the two techniques below.

Before we discuss the details of each technique, we introduce some notation and terminology related to recommendation problem. 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, in order to distinguish between the actual ratings and the predictions of the recommender system, 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 an unknown rating (i.e., the system-predicted rating for item i that user u has not rated before).

Neighborhood-based CF technique

There exist multiple variations of neighborhood-based CF techniques [9], [36], [41]. In this paper, to estimate $R^*(u, i)$, i.e., the rating that user u would give to item i, we first compute the similarity between user u and other users u' using a cosine similarity metric [9], [41]:

$$sim(u,u') = \frac{\sum_{i \in I(u,u')} R(u,i) \cdot R(u',i)}{\sqrt{\sum_{i \in I(u,u')} R(u,i)^2} \sqrt{\sum_{i \in I(u,u')} R(u',i)^2}}, \quad (1)$$

where $I(u, u')$ represents the set of all items rated by both user u and user u'. Based on the similarity calculation, set $N(u)$ of nearest neighbors of user u is obtained. The size of set $N(u)$ can range anywhere from 1 to $|U|$-1, i.e., all other users in the dataset. Then, $R^*(u, i)$ is calculated as the adjusted weighted sum of all known ratings $R(u', i)$, where $u' \in N(u)$ [14], [33]:

$$R^*(u,i) = \overline{R(u)} + \frac{\sum_{u' \in N(u)} sim(u,u') \cdot \left(R(u',i) - \overline{R(u')}\right)}{\sum_{u' \in N(u)} |\, sim(u,u')\,|}. \quad (2)$$

Here $\overline{R(u)}$ represents the average rating of user u.

A neighborhood-based CF technique can be user-based or item-based, depending on whether we calculate the similarity between users or items. Formulae (1) and (2) represent the user-based approach, but they can be straightforwardly rewritten for the item-based approach because of the symmetry between users and items in all neighborhood-based CF calculations [41]. In our experiments we used both user-based and item-based approaches for rating estimation.

Matrix factorization CF technique

Matrix factorization techniques have been the mainstay of numerical linear algebra dating back to the 1970s [17], [21], [26] and have recently gained popularity in recommender systems applications because of their effectiveness in improving recommendation accuracy [42], [48], [52], [55]. Many variations of matrix factorization techniques have been developed to solve the problems of data sparsity, overfitting, and convergence speed, and they turned out to be a crucial component of many well-performing algorithms in the popular Netflix Prize[1] competition [4], [5], [6], [16], [22], [28], [29]. We implemented the basic version of this technique, as presented in [16]. With the assumption that a user's rating for an item is composed of a sum of preferences about the various features of that item, this model is induced by Singular Value Decomposition (SVD) on the user-item ratings matrix. In particular, using K features (i.e., rank-K SVD), user u is associated with a user-factors vector p_u (the user's preferences for K features), and item i is associated with an item-factors vector q_i (the item's importance weights for K features). The preference of how much user u likes item i, denoted by $R^*(u, i)$, is predicted by taking an inner product of the two vectors, i.e.,

[1] More information can be found at www.netflixprize.com.

$$R^*(u,i) = p_u^T q_i \,. \tag{3}$$

All values in user- and item-factor vectors are initially assigned to arbitrary numbers and estimated with a simple gradient descent technique as described in (4). User- and item-factor vectors are iteratively updated with learning rate parameter (θ) as well as regularization parameter (λ), which is used to minimize overfitting, until the minimum improvement in predictive accuracy or a pre-defined number of iterations per feature is reached. One learning iteration is defined as:

For each rating $R(u, i)$

$$\begin{aligned} err &= R(u,i) - p_u^T q_i \\ p_u &= p_u + \theta(err \times q_i - \lambda \times p_u) \\ q_i &= q_i + \theta(err \times p_u - \lambda \times q_i) \end{aligned} \tag{4}$$

End For

Finally, unknown ratings are estimated with the final two vectors p_u and q_i as stated in (3). More details on variations of matrix factorization techniques used in recommender systems can be found in [4], [5], [29], [52], [55].

2.2 Accuracy of Recommendations

Numerous recommendation techniques have been developed over the last few years, and various metrics have been employed for measuring the accuracy of recommendations, including statistical accuracy metrics and decision-support measures [23]. As examples of statistical accuracy metrics, mean absolute error (MAE) and root mean squared error (RMSE) metrics measure how well a system can predict an exact rating value for a specific item. Examples of decision-support metrics include precision (the percentage of truly "high" ratings among those that were predicted to be "high" by the recommender system), recall (the percentage of correctly predicted "high" ratings among all the ratings known to be "high"), and F-measure, which is a harmonic mean of precision and recall. In particular, the ratings of the datasets that we used in our experiments are integers between 1 and 5, inclusive, where higher value represents a better-liked item. As commonly done in recommender systems literature, we define the items 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 popular *precision-in-top-N* metric [23]. The metric can be written formally as:

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

where $correct(L_N(u)) = \{i \in L_N(u) \mid R(u, i) \geq T_H\}$. However, relying on the accuracy of recommendations alone may not be enough to find the most relevant items for a user. It has often been suggested that recommender systems must be not only accurate, but also useful [23], [31]. For example, [31] suggests new user-centric directions for evaluating recommender systems beyond the conventional accuracy metrics. They claim that serendipity in recommendations or user experiences and expectations also should be considered in evaluating the recommendation quality. Among many different aspects that cannot be measured by accuracy metrics alone, in this paper we focus on the notion of the *diversity* of recommendations, which is discussed next.

2.3 Diversity of Recommendations

As mentioned in Section 1, the diversity of recommendations can be measured in two ways: individual and aggregate.

Most of recent studies have focused on increasing the *individual diversity*, which can be calculated from each user's recommendation list (e.g., an average dissimilarity between all pairs of items recommended to a given user) [8], [32], [46], [54], [57]. These techniques aim to avoid providing too similar recommendations for the same user. For example, some studies [8], [46], [57] used an intra-list similarity metric to determine the individual diversity. Alternatively, [54] used a new evaluation metric, item novelty, to measure the amount of additional diversity that one item brings to a list of recommendations. Moreover, the loss of accuracy, resulting from the increase in diversity, is controlled by changing the granularity of the underlying similarity metrics in the diversity-conscious algorithms [32].

On the other hand, except for some work that examined sales diversity across all users of the system by measuring a statistical dispersion of sales [10], [15], there have been few studies that explore *aggregate diversity* in recommender systems, despite the potential importance of diverse

recommendations from both user and business perspectives, as discussed in Section 1. Several metrics can be used to measure aggregate diversity, including the percentage of items that the recommender system is able to make recommendations for (often known as coverage) [23]. Since we intend to measure the recommender systems performance based on the top-*N* recommended items 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 formally define as follows:

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

Note that the diversity-in-top-N metric can also serve as an indicator of the level of personalization provided by a recommender system. For example, a very low diversity-in-top-*N* indicates that all users are being recommended the same top-*N* items (low level of personalization), whereas a very high diversity-in-top-*N* points to the fact that every user receives her own unique top-*N* items (high level of personalization).

In summary, the goal of the proposed ranking approaches is to improve the diversity of recommendations; however, as described in Section 1, there is a potential tradeoff between recommendation accuracy and diversity. Thus, in this paper, we aim to find techniques that can improve aggregate diversity of recommendations while maintaining adequate accuracy.

3 Motivations For Recommendation Re-ranking

In this section, we discuss how re-ranking of the candidate items whose predictions are above T_H can affect the accuracy-diversity tradeoff and how various item ranking factors, such as popularity-based approach, can improve the diversity of recommendations. Note that the general idea of personalized information ordering is not new; e.g., its importance has been discussed in information retrieval literature [34], [45], including some attempts to reduce redundancy and promote the diversity of retrieved results by re-ranking them [12], [39], [53].

3.1 Standard Ranking Approach

Typical recommender systems predict unknown ratings based on known ratings, using any traditional recommendation technique such as neighborhood-based or matrix factorization CF techniques, discussed in Section 2.1. Then, the predicted ratings are used to support the user's decision-making. In particular, each user *u* gets recommended a list of top-*N* items, denoted by

$L_N(u)$, 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 $rank$: $I \rightarrow$ R is a function representing the ranking criterion. The vast majority of current recommender systems use the *predicted rating value* as the ranking criterion or, more formally:

$$rank_{\text{Standard}}(i)=R^*(u, i)^{-1}.$$

The power of -1 in the above expression indicates that the items with *highest*-predicted (as opposed to lowest-predicted) ratings $R^*(u, i)$ are the ones being recommended to user. In the paper we refer to this as the *standard ranking approach*, and it shares the motivation with the widely used probability ranking principle in information retrieval literature that ranks the documents in order of decreasing probability of relevance [38].

Note that, by definition, recommending the most highly predicted items selected by the standard ranking approach is designed to help improve recommendation accuracy, but not recommendation diversity. Therefore, new ranking criteria are needed in order to achieve diversity improvement. Since recommending best-selling items to each user typically leads to diversity reduction, recommending less popular items intuitively should have an effect towards increasing recommendation diversity. And, as seen from the example in Table 1 (in Section 1), this intuition has empirical support. Following this motivation, we explore the possibility to use *item popularity* as a recommendation ranking criterion, and in the next subsection we show how this approach can affect the recommendation quality in terms of accuracy and diversity.

3.2 Proposed Approach: Item Popularity-Based Ranking

Item popularity-based ranking approach ranks items directly based on their popularity, from lowest to highest, where popularity is represented by the number of known ratings that each item has. More formally, item popularity-based ranking function can be written as follows:

$$rank_{\text{ItemPop}}(i) = |U(i)|, \text{ where } U(i) = \{u \in U \mid \exists R(u, i)\}.$$

We compared the performance of the item popularity-based ranking approach with the standard ranking approach using MovieLens dataset and item-based CF, and we present this comparison using the accuracy-diversity plot in Fig.1. In particular, the results show that, as compared to the standard ranking approach, 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%. 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).

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

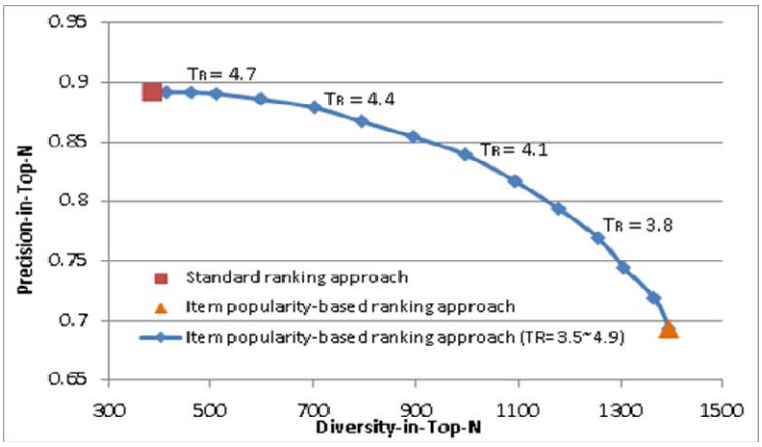
The item popularity-based ranking approach as well as all other ranking approaches proposed in this paper (to be discussed in Section 4) are parameterized with "ranking threshold" $T_R \in [T_H, T_{\max}]$ (where $T_{\max}$ is the largest possible rating on the rating scale, e.g., $T_{\max}=5$) to allow user the ability to choose a certain level of recommendation accuracy. In particular, given any ranking function $rank_X(i)$, 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:

$$rank_x(i,T_R) = \begin{cases} rank_x(i), & if\ R^*(u,i) \in [T_R, T_{\max}] \\ \alpha_u + rank_{\text{Standard}}(i), & if\ R^*(u,i) \in [T_H, T_R) \end{cases},$$

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). Thus, increasing the ranking threshold $T_R \in [T_H, T_{\max}]$ towards $T_{\max}$ would enable choosing the most highly predicted items resulting in more accuracy and less diversity (becoming increasingly similar to the standard ranking approach); in contrast, decreasing the ranking threshold $T_R \in [T_H, T_{\max}]$ towards T_H would make $rank_X(i, T_R)$ increasingly more similar to the pure ranking function $rank_X(i)$, resulting in more diversity with some accuracy loss.

Therefore, choosing different T_R values in-between the extremes allows the user to set the desired balance between accuracy and diversity. In particular, as Fig. 1 shows, the recommendation accuracy of item popularity-based ranking approach could be improved by increasing the ranking threshold. 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



MovieLens data, item-based CF (50 neighbors), top-5 item recommendation

Fig. 1. Performance of standard ranking approach and item popularity-based approach with its parameterized versions

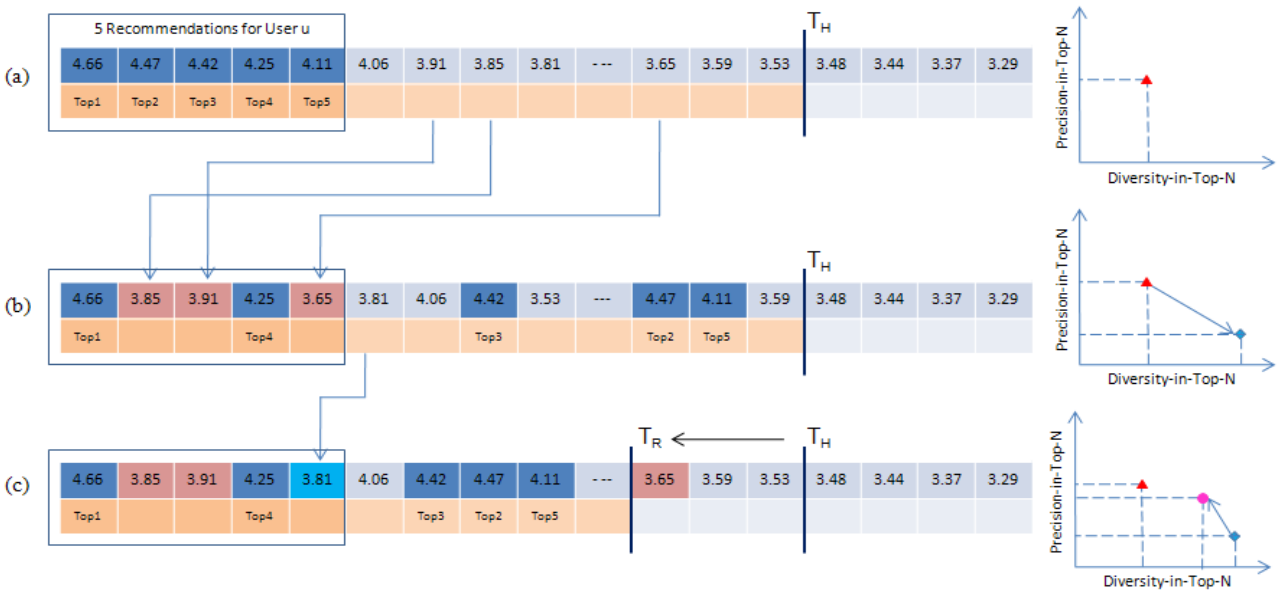
possible to achieve 20% diversity gain (from 385 to 462) with only 0.06% of accuracy loss.

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 techniques (the ones using the standard ranking approach), which is very important from the experimental analysis point of view as well in order to have a fair performance comparison of different ranking techniques.

3.4 General Steps for Recommendation Re-ranking

The item popularity-based ranking approach described above is just one example of possible ranking approaches for improving recommendation diversity, and a number of additional ranking functions, $rank_X(i)$, will be introduced in Section 4. Here, based on the previous discussion in Section 3, we summarize the general ideas behind the proposed ranking approaches, as illustrated by Fig. 2.

The first step, shown in Fig. 2a, represents the standard approach, which, for each user, ranks all the predicted items according to the predicted rating value and selects top-N candidate items, as long as they are above the highly-predicted rating threshold T_H. The recommendation quality of the overall recommendation technique is measured in terms of the precision-in-top-N and the diversity-in-top-N, as shown in the accuracy-diversity plot at the right side of the example (a).



(a) Recommending top-N highly predicted items for user u, according to standard ranking approach
(b) Recommending top-N items, according to some other ranking approach for better diversity
(c) Confining re-ranked recommendations to the items above new ranking threshold T_R (e.g., ≥ 3.8) for better accuracy

Fig. 2. General overview of ranking-based approaches for improving recommendation diversity

The second step, illustrated in Fig. 2b, shows the recommendations provided by applying one of the proposed ranking functions, $rank_X(i)$, where several different items (that are not necessarily among N most highly predicted, but are still above T_H) are recommended to the user. This way, a user can get recommended more idiosyncratic, long-tail, less frequently recommended items that may not be as widely popular, but can still be very relevant to this user (as indicated by relatively high predicted rating). Therefore, re-ranking the candidate items can significantly improve the recommendation diversity although, as discussed, this typically comes at some loss of recommendation accuracy. The performance graph of the second step (b) demonstrates this accuracy-diversity tradeoff.

The third step, shown in Fig. 2c, can significantly minimize accuracy loss by confining the re-ranked recommendations to the items above newly introduced ranking threshold T_R (e.g., 3.8 out of 5). In this particular illustration, note that the increased ranking threshold makes the fifth recommended item in the second step (b) (i.e., item with predicted rating value of 3.65) filtered out and the next possible item above the new ranking threshold (i.e. the item predicted as 3.81) is recommended to user u. Averaged across all users, this parameterization helps to make the level

of accuracy loss fairly small with still a significant diversity gain (as compared to the standard ranking approach), as shown in the performance graph of the third step (c).

We now introduce several additional item ranking functions, and provide empirical evidence that supports our motivation of using these item criteria for diversity improvement.

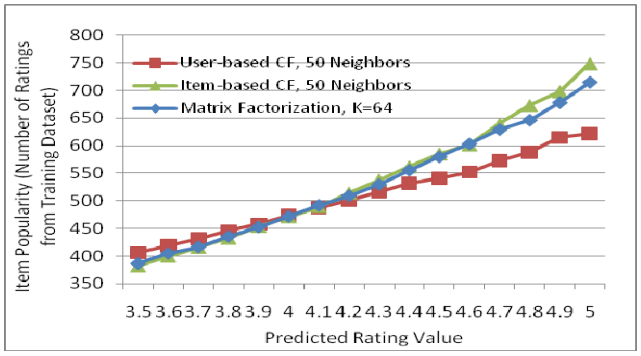
4 ADDITIONAL RANKING APPROACHES

In many personalization applications (e.g., movie or book recommendations), there often exist more highly-predicted ratings for a given user than can be put in her top-N list. This provides opportunities to have a number of alternative ranking approaches, where different sets of items can possibly be recommended to the user. In this section, we introduce six additional ranking approaches that can be used as alternatives to $rank_{\text{Standard}}$ to improve recommendation diversity, and the formal definitions of each ranking approach (provided below) are illustrated in Fig. 3 with the empirical evidence that supports the use of these item ranking criteria. Because of the space limitations, in this section we show the empirical results from MovieLens dataset; however, consistently similar patterns were found in other datasets (discussed in Section 5.1) as well.

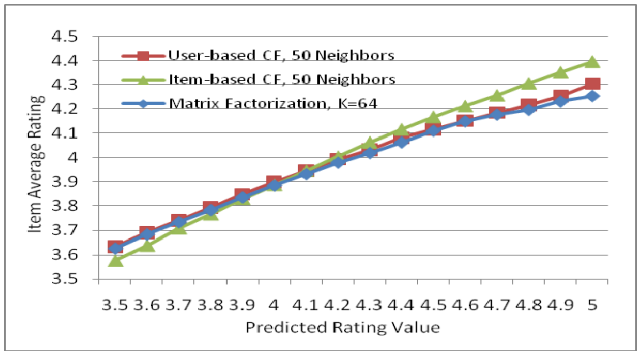
In particular, in our empirical analysis we consistently observed that popular items, on average, are likely to have higher predicted ratings than less popular items, using both heuristic- and model-based techniques for rating prediction, as shown in Fig. 3a. As discussed in Section 3, recommending less popular items helps to improve recommendation diversity; therefore, as can be immediately suggested from the monotonic relationship between average item popularity and predicted rating value, 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. Therefore, we propose to use predicted rating value itself as an item ranking criterion:

- ***Reverse Predicted Rating Value***, i.e., ranking the candidate (highly predicted) items based on their predicted rating value, from lowest to highest (as a result choosing less popular items, according to Fig. 3a). More formally:

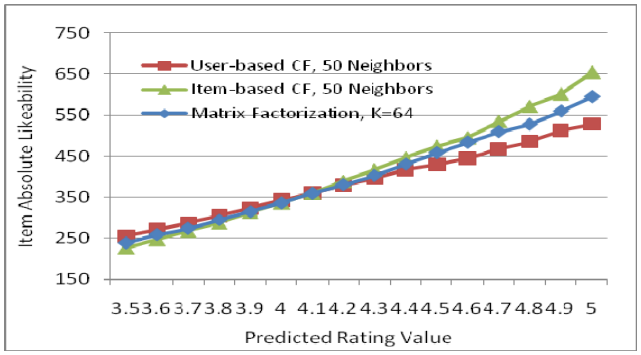
$$rank_{\text{RevPred}}(i) = R^*(u,i).$$



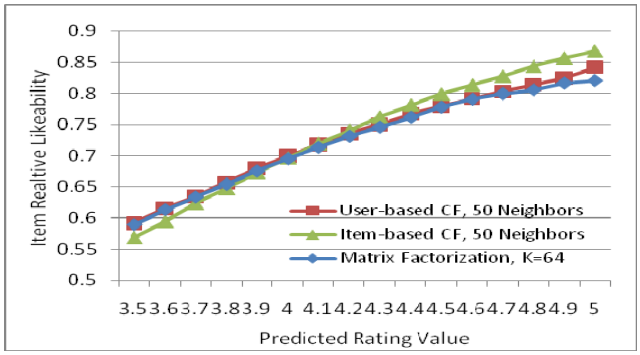
(a) Average Predicted Rating Value



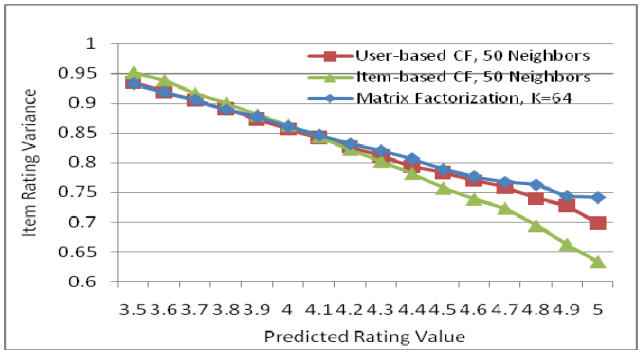
(b) Item Average Rating



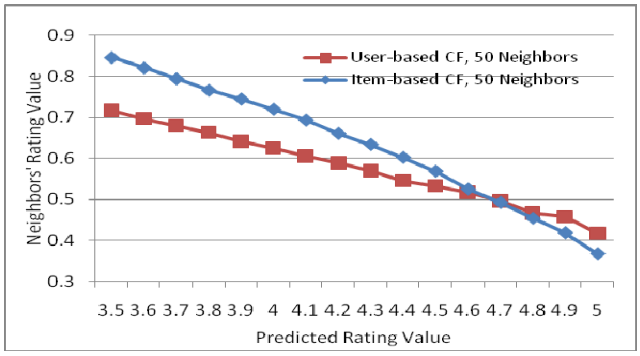
(c) Average Item Absolute Likeability



(d) Average Item Relative Likeability



(e) Average Item Rating Variance



(f) Average Neighbors' Rating Variance

Fig. 3. Relationships between various item-ranking criteria and predicted rating value, for highly-predicted ratings (MovieLens data)

We now propose several other ranking criteria that exhibit consistent relationships to predicted rating value, including average rating, absolute likeability, relative likeability, item rating variance, and neighbors' rating variance, as shown in Figures 3b-3f. In particular, the relationship between predicted rating values and the *average actual rating* of each item (as explicitly rated by users), shown in Fig. 3b, also supports a similar conjecture that items with lower average rating, on average, are more likely to have lower predicted rating values (likely representing less popular items, as shown earlier). Thus, such items could be recommended for better diversity.

- ***Item Average Rating***, i.e., ranking items according to an average of all known ratings for each item:

$$rank_{\text{AvgRating}}(i) = \overline{R(i)}, \text{ where } \overline{R(i)} = \frac{1}{|U(i)|}\sum_{u \in U(i)} R(u,i).$$

Similarly, the relationship between predicted rating values and item absolute (or relative) likeability, shown in Fig. 3c and 3d, also suggests that the items with lower likeability, on average, are more likely to have lower predicted rating values (likely representing less popular movies) and, thus, could be recommended for better diversity.

- ***Item Absolute Likeability***, i.e., ranking items according to how many users liked them (i.e., rated the item above T_H):

$$rank_{\text{AbsLike}}(i) = |U_H(i)|, \text{ where } U_H(i)=\{u \in U(i) |\ R(u,i) \geq T_H\}.$$

- ***Item Relative Likeability***, i.e., ranking items according to the percentage of the users who liked an item (among all users who rated it):

$$rank_{\text{RelLike}}(i) = |U_H(i)| \ / \ |U(i)|.$$

We can also use two different types of rating variances to improve recommendation diversity. With any traditional recommendation technique, each item's rating variance (which can be computed from known ratings submitted for that item) can be used for re-ranking candidate items. Also, if any neighborhood-based recommendation technique is used for prediction, we can use the rating variance of neighbors whose ratings are used to predict the rating for re-ranking candidate items. As shown in Fig. 3e and 3f, the relationship between the predicted rating value and each item's rating variance and the relationship between predicted rating value and 50 neighbors' rating variance obtained by using a neighborhood-based CF technique demonstrate that highly predicted items tend to be low in both item rating variance and neighbors' rating variance. In other words, among the highly-predicted ratings (i.e., above T_H) there is more user consensus for higher-predicted items than for lower-predicted ones. These findings indicate that re-ranking recommendation list by rating variance and choosing the items with higher variance could improve recommendation diversity.

- ***Item Rating Variance***, i.e., ranking items according to each item's rating variance (i.e., rating variance of users who rated the item):

$$rank_{\text{ItemVar}}(i) = \frac{1}{|U(i)|}\sum_{u\in U(i)}(R(u,i)-\overline{R(i)})^2 .$$

- ***Neighbors' Rating Variance***, i.e., ranking items according to the rating variance of neighbors of a particular user for a particular item. The closest neighbors of user u among the users who rated the particular item i, denoted by u', are chosen from the set of $U(i)\cap N(u)$.

$$rank_{\text{NeighborVar}}(i) = \frac{1}{|U(i)\cap N(u)|}\sum_{u'\in(U(i)\cap N(u))}(R(u',i)-\overline{R_u(i)})^2 \text{ , where } \overline{R_u(i)} = \frac{1}{|U(i)\cap N(u)|}\sum_{u'\in(U(i)\cap N(u))}R(u',i) .$$

In summary, there exist a number of different ranking approaches that can improve recommendation diversity by recommending items other than the ones with topmost predicted rating values to a user. In addition, as indicated in Fig. 1, the degree of improvement (and, more importantly, the degree of tolerable accuracy loss) can be controlled by the chosen ranking threshold value T_R. The next section presents comprehensive empirical results demonstrating the effectiveness and robustness of the proposed ranking techniques.

5 EMPIRICAL RESULTS

5.1 Data

The proposed recommendation ranking approaches were tested with several movie rating datasets, including MovieLens (data file available at grouplens.org), Netflix (data file available at netflixprize.com), and Yahoo! Movies (individual ratings collected from movie pages at movies.yahoo.com). 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). The basic statistical information of the resulting

TABLE 2. BASIC INFORMATION OF MOVIE RATING DATASETS

	MovieLens	Netflix	Yahoo! Movies
Number of users	2,830	3,333	1,349
Number of movies	1,919	2,092	721
Number of ratings	775,176	1,067,999	53,622
Data Sparsity	14.27%	15.32%	5.51%
Avg # of common movies between two users	64.6	57.3	4.1
Avg # of common users between two movies	85.1	99.5	6.5
Avg # of users per movie	404.0	510.5	74.4
Avg # of movies per user	274.1	320.4	39.8

datasets is summarized in Table 2. 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).

5.2 Performance of Proposed Ranking Approaches

We conducted experiments on the three datasets described in Section 5.1, using three widely popular recommendation techniques for rating prediction, including two heuristic-based (user-based and item-based CF) and one model-based (matrix factorization CF) techniques, discussed in Section 2.1. All seven proposed ranking approaches were used in conjunction with each of the three rating prediction techniques to generate top-N (N=1, 5, 10) recommendations to each user on each dataset, with the exception of neighbors' variance-based ranking of model-based predicted ratings. In particular, because there is no concept of neighbors in a pure matrix factorization technique, the ranking approach based on neighbors' rating variance was applied only with heuristic-based techniques. 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 was varied from 3.5 to 4.9. The performance of each ranking approach was measured in terms of precision-in-top-N and diversity-in-top-N (N=1, 5, 10), 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 of recommendations. 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 recommendation ranking techniques for a variety of different precision loss levels (0.1-10%).

While, as mentioned earlier, a comprehensive set of experiments was performed using every rating prediction technique in conjunction with every recommendation ranking function on every dataset for different number of top-N recommendations, the results were very consistent across all experiments and, therefore, for illustration purposes and because of the space limitations, we show only three results: each using all possible ranking techniques on a different dataset, a different recommendation technique, and a different number of recommendations. (See Table 3.)

TABLE 3. DIVERSITY GAINS OF PROPOSED RANKING APPROACHES FOR DIFFERENT LEVELS OF PRECISION LOSS

	Item Popularity		Reverse Prediction		Item Average Rating		Item Abs Likeability		Item Relative Likeability		Item Rating Variance		Neighbors' Rating Var	
Precision Loss	Diversity Gain		Diversity Gain		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	+386	2.003	+702	2.823
-0.05	+594	2.543	+594	2.543	+728	2.891	+642	2.668	+699	2.816	+283	1.735	+451	2.171
-0.025	+411	2.068	+411	2.068	+513	2.332	+445	2.156	+484	2.257	+205	1.532	+258	1.670
-0.01	+270	1.701	+234	1.608	+311	1.808	+282	1.732	+278	1.722	+126	1.327	+133	1.345
-0.005	+189	1.491	+173	1.449	+223	1.579	+196	1.509	+199	1.517	+91	1.236	+87	1.226
-0.001	+93	1.242	+44	1.114	+78	1.203	+104	1.270	+96	1.249	+21	1.055	+20	1.052
Standard: 0.892	385	1.000	385	1.000	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 Average Rating		Item Abs Likeability		Item Relative Likeability		Item Rating Variance	
Precision Loss	Diversity Gain		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	+115	1.130
-0.05	+301	1.341	+757	1.858	+718	1.814	+614	1.696	+695	1.788	+137	1.155
-0.025	+238	1.270	+568	1.644	+535	1.607	+464	1.526	+542	1.615	+110	1.125
-0.01	+156	1.177	+363	1.412	+382	1.433	+300	1.340	+385	1.437	+63	1.071
-0.005	+128	1.145	+264	1.299	+282	1.320	+247	1.280	+288	1.327	+47	1.053
-0.001	+64	1.073	+177	1.201	+118	1.134	+89	1.101	+148	1.168	+8	1.009
Standard: 0.834	882	1.000	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)

	Item Popularity		Reverse Prediction		Item Average Rating		Item Abs Likeability		Item Relative Likeability		Item Rating Variance		Neighbors' Rating Var	
Precision Loss	Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain	
-0.1	+220	1.794	+178	1.643	+149	1.538	+246	1.888	+122	1.440	+86	1.310	+128	1.462
-0.05	+198	1.715	+165	1.596	+141	1.509	+226	1.816	+117	1.422	+72	1.260	+108	1.390
-0.025	+134	1.484	+134	1.484	+103	1.372	+152	1.549	+86	1.310	+70	1.253	+98	1.354
-0.01	+73	1.264	+92	1.332	+56	1.202	+77	1.278	+58	1.209	+56	1.202	+65	1.235
-0.005	+57	1.206	+86	1.310	+38	1.137	+63	1.227	+36	1.130	+28	1.101	+51	1.184
-0.001	+42	1.152	+71	1.256	+25	1.090	+43	1.155	+30	1.110	+19	1.069	+22	1.079
Standard: 0.911	277	1.000	277	1.000	277	1.000	277	1.000	277	1.000	277	1.000	277	1.000

(c) Yahoo dataset, top-1 item, heuristic-based technique (user-based CF, 15 neighbors)

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]

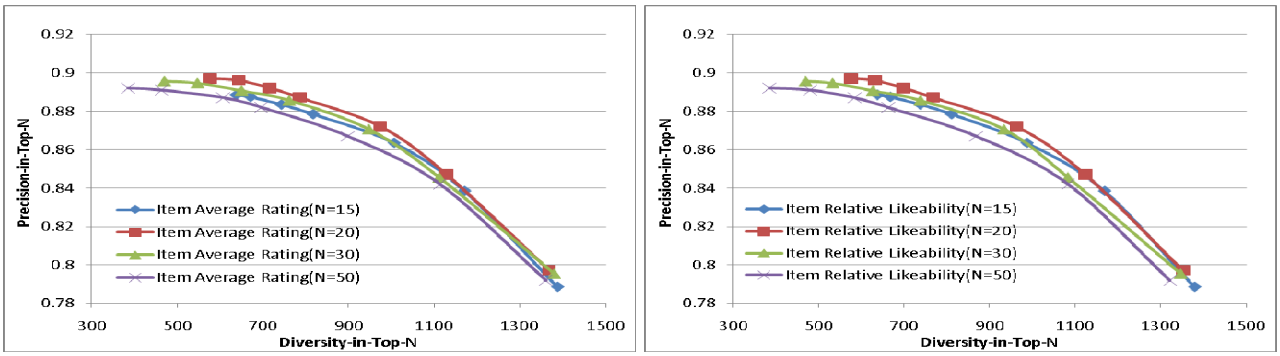
For example, Table 3a shows the performance of the proposed ranking approaches used in conjunction with item-based CF technique to provide top-5 recommendations on the MovieLens dataset. In particular, one can observe that, with the precision loss of only 0.001 or 0.1% (i.e., with precision of 0.891, down from 0.892 of the standard ranking approach), item average rating-based ranking approach can already increase recommendation diversity by 20% (i.e., absolute diversity gain of 78 on top of the 385 achieved by the standard ranking approach). If users can tolerate precision loss up to 1% (i.e., precision of 0.882 or 88.2%), the diversity could be increased by 81% with the same ranking technique; and 5% precision loss (i.e., 84.2%) can provide diversity gains up to 189% for this recommendation technique on this dataset. Substantial diversity improvements can be observed across different ranking techniques, different rating prediction techniques, and different datasets, as shown in Tables 3a, 3b, and 3c.

In general, all proposed ranking approaches were able to provide significant diversity gains, and the best-performing ranking approach may be different depending on the chosen dataset and rating prediction technique. Thus, system designers have the flexibility to choose the most desirable ranking approach based on the data in a given application. We would also like to point out that, since the proposed approaches essentially are implemented as sorting algorithms based on certain ranking heuristics, they are extremely scalable. For example, it took, on average, less than 6 seconds to rank all the predicted items and select top-N recommendations for nearly 3,000 users in our experiments with MovieLens data.

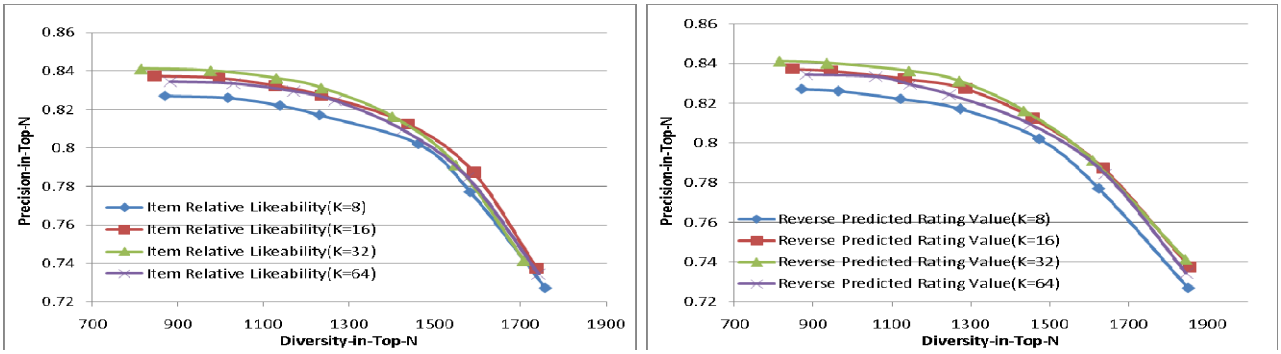
5.3 Robustness Analysis for Different Parameters

In this subsection, we present robustness analysis of the proposed techniques with respect to several parameters: number of neighbors used in heuristic-based CF, number of features used in matrix factorization CF, number of top-N recommendations provided to each user, and the value of predicted rating threshold T_H.

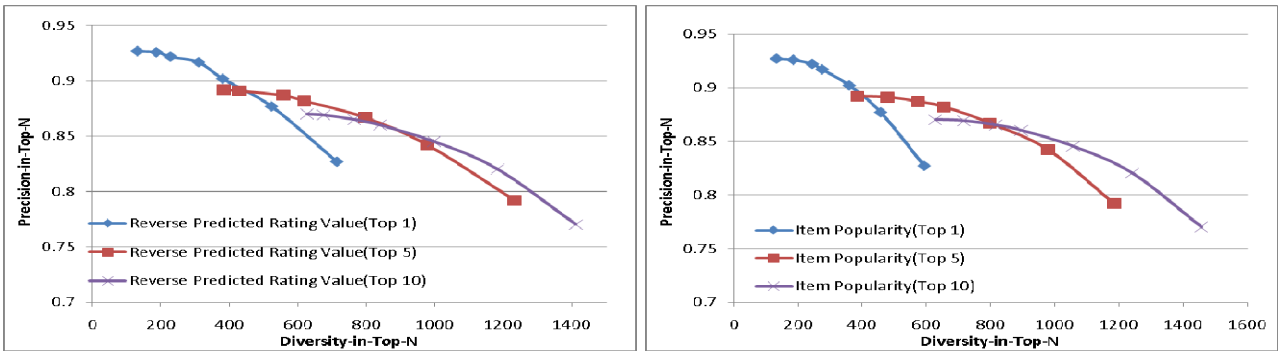
We tested the heuristic-based technique with a different number of neighbors (15, 20, 30, and 50 neighbors) and the model-based technique with a different number of features (K=8, 16, 32, and 64). For illustration purposes, Fig. 4a and 4b show how two different ranking approaches for both heuristic-based and model-based rating prediction techniques are affected by different parameter values. While different parameter values may result in slightly different performance (as is well-known in recommender systems literature), the fundamental behavior of the proposed techniques remains robust and consistent, as shown in Fig. 4a and 4b. In other words, using the



(a) Different number of neighbors (N=15, 20, 30, 50)
with MovieLens dataset, top 5 items, heuristic-based technique (item-based CF)



(b) Different number of features (K=8, 16, 32, 64)
with Netflix dataset, top 5 items, model-based technique (matrix factorization CF)



(c) Different number of recommendations (top-1, 5, 10 items)
with MovieLens dataset, heuristic-based technique (item-based CF, 50 neighbors)

Fig. 4. Performance of the proposed ranking approaches with different parameters

recommendation ranking techniques with any of the parameter values, it is possible to obtain substantial diversity improvements with only a small accuracy loss.

We also vary the number of top-N recommendations provided by the system. Note that, while it is intuitively clear that top-1, top-5, and top-10 recommendations will provide different accuracy and diversity levels (i.e., it is much easier to accurately recommend one relevant item than relevant 10 items, and it is much easier to have more aggregate diversity when you can provide

more recommendations), again we observe that, with any number of top-N recommendations, the proposed techniques exhibit robust and consistent behavior, i.e., they allow to obtain substantial diversity gains at a small accuracy loss, as shown in Fig. 4c. For example, with only 1% precision loss, we were able to increase the diversity from 133 to 311 (134% gain) using the reverse predicted rating value-based ranking approach in the top-1 recommendation task, and from 385 to 655 (70% gain) using the item-popularity-based ranking approach in the top-5 recommendation task.

Finally, our finding that the proposed ranking approaches help to improve recommendation diversity is also robust with respect to the "highly-predicted" rating threshold value T_H. In particular, with a different threshold, the baseline recommendation accuracy and diversity of the standard ranking approach could be very different, and the number of actual recommendations that are produced by the system (i.e., in case there is a limited number of items that are predicted higher than the minimum threshold) may change. However, again we observe the same consistent ability of the proposed ranking approaches to achieve substantial diversity gains with only a small accuracy loss. For example, as shown in Table 4, with a different predicted rating threshold (i.e., $T_H = 4.5$) and 1% precision loss, we could obtain 68% diversity gain by ranking the recommendations based on item average rating in top-1 recommendation task on MovieLens dataset using item-based CF for rating prediction. Similar improvements were observed for other datasets and rating prediction techniques as well.

TABLE 4 PERFORMANCE OF PROPOSED RANKING APPROACHES WITH A DIFFERENT PREDICTED RATING THRESHOLD ($T_H = 4.5$)

	Item Popularity		Reverse Prediction		Item Average Rating		Item Abs Likeability		Item Relative Likeability		Item Rating Variance		Neighbors' Rating Var	
Precision Loss	Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain		Diversity Gain	
-0.1	+187	2.928	+207	3.134	+289	3.979	+197	3.031	+262	3.701	+101	2.039	+141	2.454
-0.05	+127	2.309	+124	2.278	+189	2.948	+134	2.381	+182	2.876	+82	1.845	+83	1.856
-0.025	+72	1.742	+74	1.763	+99	2.021	+81	1.835	+101	2.041	+43	1.443	+43	1.443
-0.01	+48	1.495	+45	1.464	+66	1.680	+54	1.557	+55	1.567	+23	1.237	+18	1.186
-0.005	+41	1.420	+36	1.371	+58	1.598	+45	1.468	+47	1.485	+13	1.134	+10	1.103
-0.001	+35	1.362	+28	1.288	+52	1.536	+39	1.399	+41	1.423	+6	1.059	+4	1.039
Standard: 0.775	97	1.000	97	1.000	97	1.000	97	1.000	97	1.000	97	1.000	97	1.000

MovieLens dataset, top-1 item, heuristic-based technique (item-based CF, 50 neighbors)

6 DISCUSSION AND ADDITIONAL ANALYSIS

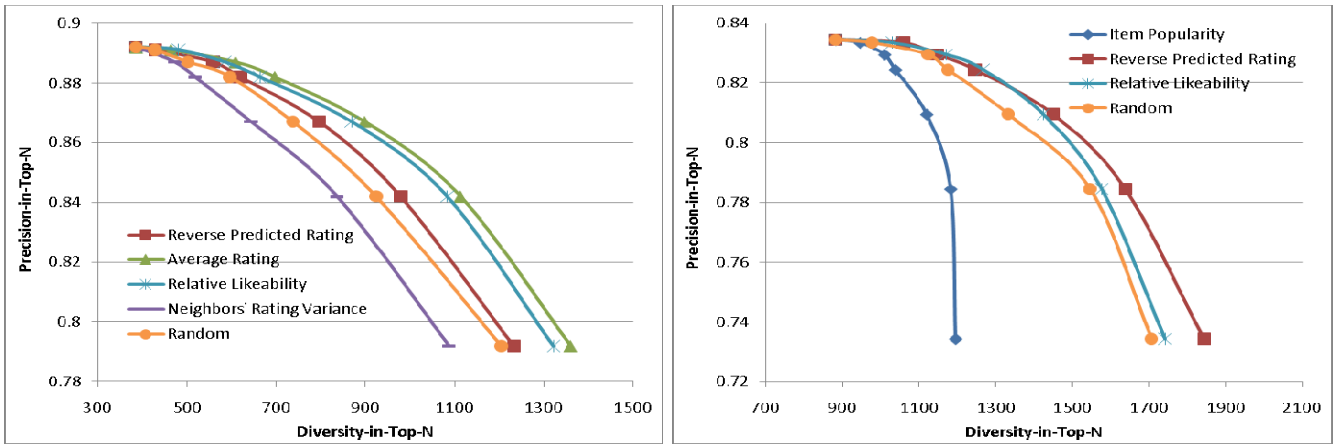
In this section, we explore and discuss several additional issues related to the proposed ranking approaches.

6.1 Random Ranking Approach

As mentioned earlier, the vast majority of traditional recommender systems adopt the standard ranking approach that ranks the candidate items according to their predicted rating values and, thus, recommends to users the topmost highly predicted items. As discussed in Section 3, since the more highly predicted items, on average, tend to be among the more popular items, using this ranking approach will often result in lower recommendation diversity. While the proposed ranking approaches improve the diversity by considering alternative item ranking functions, such as item popularity, we also found that re-ranking the candidate items even at random can provide diversity improvements as compared to the standard ranking approach. Here we defined the random ranking as:

$$Rank_{\text{Random}}(i) = \text{Random}(0,1).$$

where Random(0,1) is a function that generates uniformly distributed random numbers in the [0, 1] interval. We compare some of the proposed ranking approaches with this random ranking approach in Fig. 5. For example, as shown in Fig. 5a, the random ranking approach increased the diversity from 385 to 596 (55% gain) with 1% precision loss using heuristic-based CF technique



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

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

Fig. 5. Diversity gain of the random ranking approach with different levels of precision loss.

on MovieLens dataset. While this gain was not as big as the diversity gain of the average rating-based approach (80% gain), it actually outperformed the neighbors' rating variance-based approach (35% gain). As another example, as shown in Fig. 5b, with only 0.5% precision loss on Netflix dataset using model-based CF technique, the random ranking approach produced the results that were almost as good (27% diversity gain) as several best-performing ranking approaches (i.e., 30% gain for the reverse predicted rating-based approach or 33% gain for the relative likeability-based approach).

This provides a valuable insight that, if the goal is to improve recommendation diversity (without significant accuracy loss), even a random recommendation ranking approach can significantly outperform the traditional and widely-used standard ranking approach (based on the predicted rating value). Furthermore, as illustrated in Fig. 5a and 5b, the random ranking approach works consistently well with different datasets and in conjunction with different CF techniques.

6.2 Combining Ranking Approaches

There have been many examples in recommender systems literature where an effective technique was developed by combining or integrating several other recommendation techniques [3], [4], [5], [13], [35], [47], [52], and it has been empirically shown that the combined approaches may provide better performance and help to avoid certain limitations that individual approaches may

have. Examples of this include: hybrid recommendation techniques that combine collaborative and content-based methods [3], [35], [47]; CF approaches that combine both memory-based and model-based methods [13]; and many new "blending" or ensemble-based techniques that have become popular in recent years, stemming in part from the advances in the Netflix Prize competition [4], [5], [52].

Following this idea, we explored the possibilities to combine several of the proposed ranking approaches for the purpose of further diversity improvements. There are many possible ways to combine several ranking functions; in this paper, we use a simple technique based on a linear combination:

$$rank_{\text{Combined}}(i) = p_1 \times rank_1(i) + p_2 \times rank_2(i) + \ldots \; p_k \times rank_k(i),$$

where $p_1 + p_2 + \ldots + p_k = 1$. Many different combined ranking functions can be obtained by choosing different weights p_j, and finding the optimal combination constitutes an interesting topic for future research. For illustration purposes, in this paper we tested several combinations of two ranking approaches by using different weights (p_1= 0, 0.1, ..., 1 and $p_2 = 1\text{-}p_1$). One well-performing combination is presented in Table 5. In particular, for the model-based CF technique used in the top-5 recommendation task on the Netflix dataset, the empirical results show that the combined ranking approach (obtained by combining the item average rating and reverse prediction-based approaches) was able to produce bigger diversity gains than each of the individual approaches (for the same level of accuracy loss). For example, with 1% precision loss, while the item average rating-based or reverse predicted rating-based approaches each obtained the diversity gain of 41-43%, the combination of these two approaches was able to produce a nearly 49%

TABLE 5 PERFORMANCE OF THE COMBINED RANKING APPROACH

	Item Average Rating (R_1)		Reverse Prediction (R_2)		Combination: $0.4*R_1 + 0.6*R_2$	
Precision Loss	Diversity Gain		Diversity Gain		Diversity Gain	
-0.1	+880	1.998	+962	2.091	+1002	2.136
-0.05	+718	1.814	+757	1.858	+818	1.927
-0.025	+535	1.607	+568	1.644	+627	1.711
-0.01	+382	1.433	+363	1.412	+429	1.486
-0.005	+282	1.320	+264	1.299	+321	1.364
-0.001	+118	1.134	+177	1.201	+205	1.232
Standard: 0.834	882	1.000	882	1.000	882	1.000

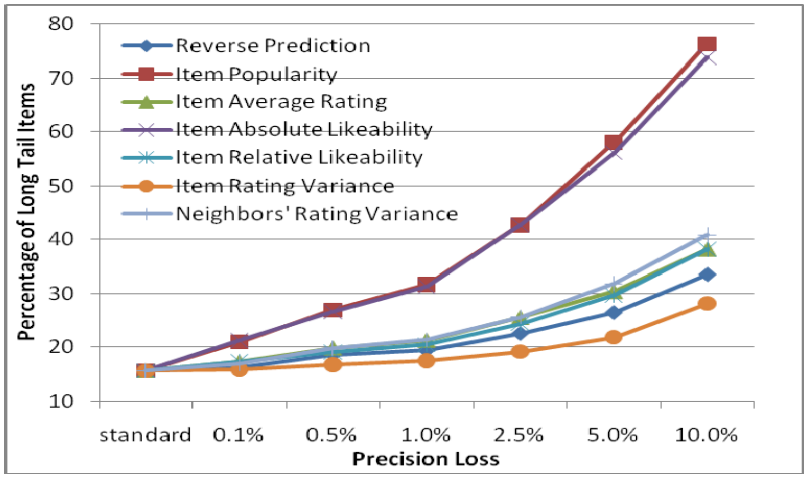
Netflix data, top-5 items, model-based technique (matrix factorization, K=64)

increase in recommendation diversity.

6.3 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. The results are shown in Fig. 6.

For example, with the standard ranking approach, the long-tail items consist of only 16% of total recommendations (i.e., 84% of 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



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

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

Fig. 6. Proportion of long-tail items among recommended items.

bestsellers and not long-tail items [15]. However, as shown in Fig. 6, 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 neighbors' rating variance-based ranking approach, or to 32% with item popularity and item absolute likeability-based approaches. And with 2.5% or 5% precision loss, the proportion of long-tail items can grow up to 43% and 58%, respectively (e.g., using item popularity ranking technique).

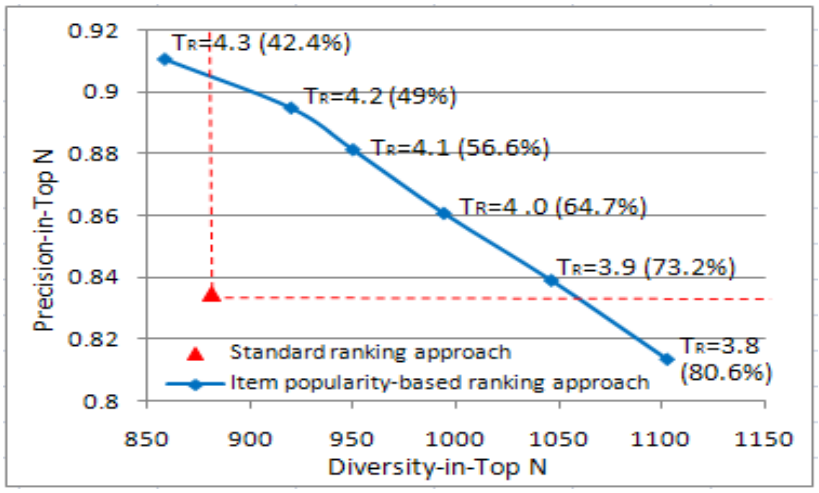
This analysis provides further empirical support to the fact that the proposed ranking approaches increase not just the number of distinct items recommended, but also the proportion of recommended long-tail items, thus, confirming that the proposed techniques truly contribute towards more diverse and idiosyncratic recommendations across all users.

6.4 Improving Both Accuracy and Diversity: Recommending Fewer Items

Empirical results in this paper consistently show that the proposed ranking approaches can obtain significant diversity gains (with a small amount of accuracy loss) as compared to the standard ranking approach that ranks recommended items based on their predicted rating value. Therefore, another interesting topic for future research would be to explore possibilities to improve *both* accuracy as well as diversity.

Based on the findings described in this paper, a possible approach to improving both the accuracy and diversity of the standard technique would be to modify the proposed recommendation re-ranking techniques, which are already known to produce diversity gains, in a way that increases their accuracy. Perhaps counter-intuitively, one of the possible ways to do this involves recommending *fewer* items. In particular, the parameterized versions of the proposed ranking techniques use threshold T_R to differentiate the items that should be ranked by the proposed technique from the ones to be ranked by the standard ranking technique, as discussed in Section 3.3. However, T_R can be used not only for ranking, but also for filtering purposes, i.e., by updating the parameterized ranking function as follows:

$$rank_x(i, T_R) = \begin{cases} rank_x(i), & if\ R^*(u,i) \in [T_R, T_{\max}] \\ \text{Remove item}, & if\ R^*(u,i) \in [T_H, T_R) \end{cases}.$$



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

Fig. 7. Improving *both* accuracy and diversity of recommendations.

This will recommend only items that are predicted to be not only above T_H, but above T_R as well (where always $T_R \geq T_H$), consequently improving the recommendation accuracy.

While the comprehensive exploration of this phenomenon is beyond the scope of this paper, in Fig. 7 we illustrate how the item popularity-based ranking approach can be modified using the above-mentioned strict filtering policy to improve upon the standard approach both in terms of accuracy and diversity. As Fig. 7 demonstrates, item popularity-based ranking approach with T_R = 4.1 (out of 5) generates only 56.6% of all possible item recommendations that could be obtained from standard ranking approach (because the recommendations with predicted rating < 4.1 were removed). Interestingly, however, despite the smaller number of recommendations, this ranking approach increased the recommendation accuracy by 4.6% (from 83.5% to 88.1%) and diversity by 70 items or 7.8% (from 881 to 951). As shown in Fig. 7, using different T_R values allows to produce different accuracy and diversity gains.

As discussed above, this approach would not be able to provide all N recommendations for each user, but it nevertheless may be useful in cases where system designers need the flexibility to apply other recommendation strategies to fill out the remaining top-N item slots. For example, some recommender systems may want to adopt "exploration-vs-exploitation" strategy [50], where some of the recommendations are tailored directly towards the user's tastes and preferences (i.e., exploitation), and the proposed ranking techniques with strict filtering can be used to fill out this part of the recommendation list for each user (providing both accuracy and diversity benefits over the standard approach). Meanwhile, the remaining recommendations can be de-

signed to learn more about the user (i.e., exploration), e.g., using *active learning* techniques [25], [56], so that the system can make better recommendations to the users in the future.

7 CONCLUSIONS AND FUTURE WORK

Recommender systems have made significant progress in recent years and many techniques have been proposed to improve the recommendation quality. However, in most cases, new techniques are designed to improve the accuracy of recommendations, whereas the recommendation diversity has often been overlooked. In particular, we showed that, while ranking recommendations according to the predicted rating values (which is a *de facto* ranking standard in recommender systems) provides good predictive accuracy, it tends to perform poorly with respect to recommendation diversity. Therefore, in this paper, we proposed a number of recommendation ranking techniques that can provide significant improvements in recommendation diversity with only a small amount of accuracy loss. In addition, these ranking techniques offer flexibility to system designers, since they are parameterizable and can be used in conjunction with different rating prediction algorithms (i.e., they do not require the designer to use only some specific algorithm). They are also based on scalable sorting-based heuristics and, thus, are extremely efficient. We provide a comprehensive empirical evaluation of the proposed techniques and obtain consistent and robust diversity improvements across multiple real-world datasets and using different rating prediction techniques.

This work gives rise to several interesting directions for future research. In particular, additional important item ranking criteria should be explored for potential diversity improvements. This may include consumer-oriented or manufacturer-oriented ranking mechanisms [19], depending on the given application domain, as well as external factors, such as social networks [30]. Also, more sophisticated techniques (such as optimization-based approaches) could be used to achieve further improvements in recommendation diversity, although these improvements may come with a (possibly significant) increase in computational complexity. Furthermore, exploration of recommendation diversity when recommending item *bundles* [18] or *sequences* [43] (instead of individual items) also constitute interesting topics for future research. In summary, we hope that this paper will stimulate further research on improving recommendation diversity and other aspects of recommendation quality.

ACKNOWLEDGMENT

The research reported in this paper was supported in part by the National Science Foundation grant IIS-0546443.

REFERENCES

[1] G. Adomavicius and 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), pp. 734-749, 2005.

[2] C. Anderson, "The Long Tail," *New York: Hyperion,* 2006.

[3] M. Balabanovic and Y. Shoham, "Fab: Content-Based, Collaborative Recommendation," Comm. ACM, 40(3), pp. 66-72, 1997.

[4] R. Bell and Y. Koren, and C. Volinsky, "The BellKor solution to the Netflix Prize," http://www.netflixprize.com/assets/ProgressPrize2007_KorBell.pdf, 2007.

[5] R. M. Bell, Y. Koren, and C. Volinsky, "The Bellkor 2008 solution to the Netflix Prize.," http://www.research.att.com/~volinsky/netflix/ProgressPrize2008 Bell KorSolution.pdf, 2008

[6] J. Bennett, and S. Lanning, "The Netflix Prize," *Proc. of KDD-Cup and Workshop at the 13th ACM SIGKDD Int'l Conf. on Knowledge and Data Mining*, 2007.

[7] D. Billsus and M. Pazzani, "Learning Collaborative Information Filters," *Proc. Int'l Conf. Machine Learning*, 1998.

[8] K. Bradley and B. Smyth, "Improving Recommendation Diversity," *Proc. of the 12th Irish Conf. on Artificial Intelligence and Cognitive Science*, 2001.

[9] S. Breese, D. Heckerman, and C. Kadie, "Empirical Analysis of Predictive Algorithms for Collaborative Filtering," *Proc. of the 14th Conf. on Uncertainty in Artificial Intelligence*, 1998.

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

[11] E. Brynjolfsson, Y. Hu, and M.D. Smith, "Consumer Surplus in the Digital Economy: Estimating the Value of Increased Product Variety at Online Booksellers," *Management Science*, 49(11), pp. 1580-1596, 2003.

[12] J. Carbonell and J. Goldstein, "The user of MMR, diversity-based reranking for reordering documents and producing summaries," *Proc. of the ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, pp. 335-336, 1998.

[13] M. Claypool, A. Gokhale, T. Miranda, P. Murnikov, D. Netes, and M. Sartin, "Combining Content-Based and Collaborative Filters in an Online Newspaper," *Proc. ACM SIGIR '99 Workshop Recommender Systems: Algorithms and Evaluation*, 1999.

[14] J. Delgado and N. Ishii, "Memory-Based Weighted-Majority Prediction for Recommender Systems," *Proc. ACM SIGIR '99 Workshop Recommender Systems: Algorithms and Evaluation*, 1999.

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

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

[17] K. R. Gabriel and S. Zamir, "Lower rank approximation of matrices by least squares with any choice of weights," *Technometrics*, 21, pp. 489–498, 1979.

[18] R. Garfinkel, R. Gopal, A. Tripathi, and F. Yin, "Design of a shopbot and recommender system for bundle purchases," Decision Support Systems, 42(3), pp. 1974-1986, 2006.

[19] A. Ghose, and P. Ipeirotis, "Designing Novel Review Ranking Systems: Predicting Usefulness and Impact of Reviews," *Proc. of the 9th Int'l Conf. on Electronic Commerce (ICEC)*, 2007.

[20] D.G. Goldstein and D.C. Goldstein, "Profiting from the Long Tail," *Harvard Business Review*, June 2006.

[21] G.H. Golub and C. Reinsche, "Singular value decomposition and least squares solution," Numer. Math., 14, pp. 403-420, 1970.

[22] K. Greene, "The $1 million Netflix challenge," Technology Review. www.tech nologyreview.com/read_article.aspx?id=17587&ch=biztech, October 6, 2006.

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

[24] T. Hofmann, "Collaborative Filtering via Gaussian Probabilistic Latent Semantic Analysis," *Proc. 26th Ann. Int'l ACM SIGIR Conf.*, 2003.

[25] Z. Huang, "Selectively Acquiring Ratings for Product Recommendation," *International Conference for Electronic Commerce*, 2007.

[26] V. Klema and A. Laub, A., "The singular value decomposition: Its computation and some applications," *IEEE Transactions on Automatic Control*, .25(2), pp. 164-176, 1980.

[27] W. Knight, "Info-mania' dents IQ more than marijuana," NewScientist.com news, 2005. URL: http://www.newscientist.com/article.ns?id=dn7298.

[28] Y. Koren, "Tutorial on recent progress in collaborative filtering," *Proc. of the 2008 ACM Conf. on recommender systems*, pp. 333-334, 2008.

[29] Y. Koren, "Collaborative filtering with temporal dynamics," *Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge discovery and data mining*, pp. 447-456, 2009.

[30] D. Lemire, S. Downes, and S. Paquet, "Diversity in open social networks," published online, 2008.

[31] S.M.. McNee, J. Riedl, J. A. Konstan, "Being Accurate is Not Enough: How Accuracy Metrics have hurt Recommender Systems," *Conf. on Human Factors in Computing Systems*, pp. 1097-1101, 2006.

[32] D. McSherry, "Diversity-Conscious Retrieval," P*roc. of the 6th European Conference on Advances in Case-Based Reasoning*, pp. 219-233, 2002.

[33] A. Nakamura and N. Abe, "Collaborative Filtering Using Weighted Majority Prediction Algorithms," *Proc. of the 15th Int'l Conf. Machine Learning*, 1998.

[34] S.T. Park and D.M. Pennock, "Applying collaborative filtering techniques to movie search for better ranking and browsing," *Proc. of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 550-559, 2007.

[35] D.M. Pennock and E. Horvitz, "Collaborative Filtering by Personality Diagnosis: A Hybrid Memory And Model-Based Approach," *Proc. Int'l Joint Conf. Artificial Intelligence Workshop: Machine Learning for Information Filtering*, Aug. 1999.

[36] P. Resnick, N. Iakovou, M. Sushak, P. Bergstrom, and J. Riedl, "GroupLens: An Open Architecture for Collaborative Filtering of Netnews," *Proc. 1994 Computer Supported Cooperative Work Conf.*, 1994.

[37] P. Resnick and H. R. Varian, "Recommender systems," *Comm. ACM*, 40(3), pp. 56–58, 1997.

[38] S.E. Robertson, "The probability ranking principles in IR," *Readings in Information Retrieval*, pp. 281-286, 1997.

[39] M. Sanderson, J. Tang, T. Arni, and P. Clough, "What Else Is There? Search Diversity Examined," *European Conf. on Information Retrieval*, pp. 562-569, 2009.

[40] B. M. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Analysis of Recommender Algorithms for E-Commerce," *ACM E-Commerce 2000 Conf.*, pp.158-167, 2000.

[41] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Item-Based Collaborative Filtering Recommendation Algorithms," *Proc. of the 10th Int'l WWW Conf.*, 2001.

[42] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl, "Application of Dimensionality Reduction in Recommender Systems—A Case Study," *Proc. ACM WebKDD Workshop*, 2000.

[43] G. Shani, D. Heckerman, and R. Brafman, "An MDP-based recommender system," *Journal of Machine Learning Research*, 6, pp. 1265-1295, 2005.

[44] L. Si and R. Jin, "Flexible Mixture Model for Collaborative Filtering," *Proc. of the 20th Int'l Conf. on Machine Learning*, 2003.

[45] B. Smyth, and K. Bradley, "Personalized Information Ordering: A Case-Study in Online Recruitment," *Journal of Knowledge-Based Systems*, 16(5-6), pp..269-275., 2003.

[46] B. Smyth and P. McClave, "Similarity vs. Diversity," *Proc. of the 4th Intl. Conf. on Case-Based Reasoning: Case-Based Reasoning Research and Development*, 2001.

[47] I. Soboroff and C. Nicholas, "Combining Content and Collaboration in Text Filtering," *Proc. Int'l Joint Conf. Artificial Intelligence Workshop: Machine Learning for Information Filtering*, Aug. 1999.

[48] N. Srebro and T. Jaakkola, "Weighted low-rank approximations," *In T. Fawcett and N. Mishra, editors, ICML*, pp. 720–727. AAAI Press, 2003.

[49] X. Su and T. M. Khoshgoftaar, "Collaborative Filtering for Multi-class Data Using Belief Nets Algorithms," *Proc. of the 8th IEEE Int'l Conf. on Tools with Artificial Intelligence*, pp. 497-504, 2006.

[50] S. ten Hagen, M. van Someren, and V. Hollink, "Exploration/exploitation in adaptive recommender systems," *Proc. of the European Symposium on Intelligent Technologies, Hybrid Systems and their implementation on Smart Adaptive Systems*, 2003.

[51] C. Thompson, "If You Liked This, You're Sure to Love That," The New York Times, http://www.nytimes.com/2008/11/23/magazine/23Netflix-t.html, November 21, 2008.

[52] M. Wu. Collaborative filtering via ensembles of matrix factorization. *In KDDCup 2007*, pp. 43-47, 2007.

[53] C. Zhai, W.W. Cohen, and J. Lafferty, "Beyond independent relevance: Methods and evaluation metrics for subtopic retrieval," *Proc. of the ACM Conf. on Research and Development in Information Retrieval (SIGIR)*, 2003.

[54] M. Zhang and N. Hurley, "Avoiding monotony: improving the diversity of recommendation lists," *Proc. of the 2008 ACM Conf. on Recommender systems*, pp. 123-130, 2008.

[55] S. Zhang, W. Wang, J. Ford, F. Makedon, and J. Pearlman, "Using singular value decomposition approximation for collaborative filtering." *Proc. of the 7th IEEE International Conf. on E-Commerce Technology (CEC'05)*, pp. 257-264, 2005.

[56] Z. Zheng and B. Padmanabhan, "Selectively Acquiring Customer Infomration: A new data acquisition problem and an Active Learning-based solution," *Management Science*, 50(5), pp. 697-712, 2006.

[57] C-N. Ziegler, S.M. McNee, J.A. Konstan, and G. Lausen, "Improving Recommendation Lists Through Topic Diversification," *Proc. of the 14th Int'l World Wide Web Conf.*, pp. 22-32, 2005.