A Clustering Approach for Personalizing Diversity in Collaborative Recommender Systems

Farzad Eskandanian, Bamshad Mobasher, Robin Burke
Center for Web Intelligence
DePaul University
feskanda@depaul.edu, mobasher@cs.depaul.edu, rburke@cs.depaul.edu

Abstract

Much of the focus of recommender systems research has been on the accurate prediction of users' ratings for unseen items. Recent work has suggested that objectives such as diversity and novelty in recommendations are also important factors in the effectiveness of a recommender system. However, methods that attempt to increase diversity of recommendation lists for all users without considering each user's preference or tolerance for diversity may lead to monotony for some users and to poor recommendations for others. Our goal in this research is to evaluate the hypothesis that users' propensity towards diversity varies greatly and that the diversity of recommendation lists should be consistent with the level of user interest in diverse recommendations. We propose a pre-filtering clustering approach to group users with similar levels of tolerance for diversity. Our contributions are twofold. First, we propose a method for personalizing diversity by performing collaborative filtering independently on different segments of users based on the degree of diversity in their profiles. Secondly, we investigate the accuracy-diversity tradeoffs using the proposed method across different user segments. As part of this evaluation we propose new metrics, adapted from information retrieval, that help us measure the effectiveness of our approach in personalizing diversity. Our experimental evaluation is based on two different datasets: MovieLens movie ratings, and Yelp restaurant reviews.

1 Introduction

Recommender systems help users find useful items by tailoring recommendations to the users' tastes and preferences. From the beginning, much of the attention in recommender systems research has been devoted to generating accurate recommendations. But, focusing only on accuracy as an objective may prevent recommender systems from taking the risk of recommending items that are different from those seen by user in the past. This, in turn may lead to monotony in users' interactions with the system. Indeed, recommending too many similar or redundant items are well-known drawbacks of many of the traditional recommendation algorithms. In recent years other criteria beyond accuracy, such as diversity and novelty of recommendations, have been studied as important factors affecting the effectiveness of recommendation for users [1, 2, 3]. In particular, "diversity" is a list-wise property that may add another dimension of quality and utility to the recommendation lists generated for a user by the system. In our context, diversity refers to the distribution of categories, genres, or topical areas with a recommended list of items

Many of the approaches studied so far to incorporate the notion of diversity in recommendation have focused on increasing the diversity of recommendation lists for all users while maintaining accuracy. Early work in information retrieval, for instance, has involved re-ranking the search results in a way that would increase the topic diversity of the top results. One such approach involves a re-ranking algorithm called Maximal Marginal Relevance (MMR) [4]. *Marginal relevance* is defined as a linear combination of relevance (to the query) and dissimilarity of retrieved documents in the search results. The core ideas behind this re-ranking approach have been extended to the area of intent-aware diversification [5, 6]. Other approaches have tried to incorporate the notion of diversity into the learning to rank process [7, 8].

However, many of these methods generally assume that the user's preference or tolerance for diversity is constant across all users. This assumption, however, may not be appropriate in many situations. As an example consider the domain of movie recommendations (e.g., Netflix). One can imagine two extreme cases: one user, with very narrow movie interests, likes to receive as recommendations only science fiction movies made within the last 10 years; another user has a much broader taste in movies and would prefer a more diverse set of movies from many genres in her recommendation list. Obviously, any attempt to increase the diversity of recommendations in a uniform way for all users will result in poor recommendations for the first user. For this reason it is important to adapt the diversity of recommendations to users based on their degree of tolerance and propensity for diversity. In other words, recommender systems need not only personalize the recommendation lists based on users' preferences or tastes, but also personalize the degree of diversity in the recommendation lists.

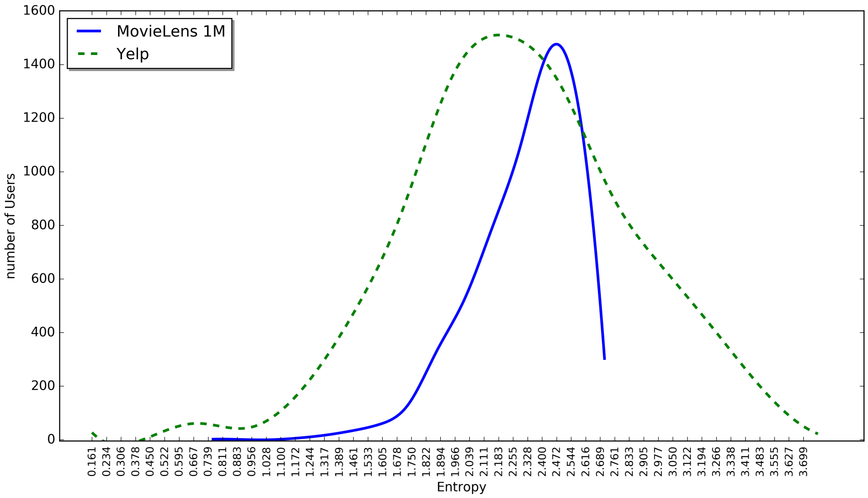

Figure 1: Entropy Distribution of user profiles.

Personalizing diversity has been discussed in some recent work. For example, one study [9] has considered a content-based approach to modeling users' interests and attempting to diversify recommendations based on the content information on items (such as item features). Another work [10] has addressed the need to increase the coverage of user propensity toward diversity and at the same time, reducing redundancy for diversity. Both of these solutions are based on a post-processing approach that accounts to a greedy re-ranking of the recommendation results.

Our goal is not only to develop recommendation algorithms that personalize the results based on past preferences, but also tailor the degree of diversity in the recommendation lists to users' tolerance for diversity. To this end, we are proposing a pre-filtering approach to personalize diversity by automatically segmenting the users based on their preferences on the predefined categories (topics, genres, etc.) associated with items. We perform this clustering using a novel approach that takes into account similarities among users based on the distribution of categories in their user profiles, and this resulting in segments of users with the same degree of interest or propensity for diversity. We then employ a user-based collaborative filtering approach to compare the recommendations in the general population (without diversity-based segmentation) with those obtained for each segment separately.

Our experimental evaluation performed on two different data sets, shows that the proposed clustering approach is effective in tailoring the diversity of recommendation lists to user's propensity for diversity.

2 Measuring Diversity

There are two primary frameworks for measuring diversity in recommender systems. The first framework is based on the pairwise similarity of items in a list called Intra-List Similarity (ILS) [2]. The other framework is called intent-aware diversification [11, 6] and has been mostly used in the Information Retrieval. The goal of intent-aware diversification is that in case of ambiguous queries the system should try to cover as many aspects or subtopics associated with the query as possible in order to increase the likelihood that the aspect corresponding to user's real intent is covered by search results.

Ideally, a diversification method should have two characteristics. First, it should diminish the amount of redundancy in the recommendation list. Secondly, it should cover all of the aspects that are interesting to user in order to personalize the level of diversity. From this perspective, our definition of diversity is similar to [10]. Both of these properties require category information for items. Fortunately, in many domains such information is usually available (e.g., genres in movies, cuisine type in restaurants, or topic categories of news stories). One goal of this research is to develop a diversity metric that can effectively capture both of the aforementioned characteristics, thus providing a uniform measure of the personalization of diversity in recommendation lists. We discuss our proposed metric, adapted from information retrieval, in Section 4.3

Assuming that a set of categories $\mathcal{C}$ already exists, the information about item categories can be used to model and measure the interest of user in each of those categories. Usually, the interaction of user u with the system is captured by the ratings $r(u, i)$ assigned to each of items $i \in I_u$, where I_u is the user profile (the collection of user's ratings on items in the set of all items I). Furthermore, we assume that each item in I belongs to one or more categories in $\mathcal{C}$. Using ratings in user profile I_u, we can define the likelihood of the interest of user u in category c_k by:

$$P(c_k|u) = \frac{\sum_{i \in I_u} r(u,i) P(i \in c_k)}{\sum_{c_j \in \mathcal{C}} \sum_{i \in I_u} r(u,i) P(i \in c_j)} \tag{1}$$

The likelihood of user being interested in a category c depends on the ratings of user on the items which belong to this category. Also the probability that each item belongs to this category $P(i \in c)$ is used as a coefficient to the ratings. A similar approach is used in [6, 12]. If we compute these values for every category in $\mathcal{C}$, we get a representation of a user's profile as a distribution over categories.

To measure the degree of tolerance for diversity for a user u we use the Shannon entropy:

$$\mathcal{H}(u) = -\sum_{c \in \mathcal{C}} P(c|u)\, log_2(P(c|u)) \tag{2}$$

High entropy in a categorical distribution of user profile represents high interest of user in diversity. Figure 1 shows the distribution of entropy values across the MovieLens and Yelp restaurant user profiles. The high variance of interest in diversity can be seen specially in the Yelp dataset. Our goal in this work is to develop an automated approach for segmenting users based on this diversity variance, so that users with with the same propensity for diversity are grouped together. We would then use standard collaborative recommendation algorithm on each segment separately. In an earlier work we showed that based on such segmentation the diversity level of users will be preserved in the recommendations [13]. But, the previous approach was based on manual segmentation of user profiles and the the metric used for diversity was Inter-List Distance (ILD) which does not capture user's preferences over genres, but only the spread of categories in the user's profile.

3 Clustering user profiles

We propose a clustering approach to automatically segment users who share the same level of diversity in their user profiles.

To segment the users based on their degree of preference in diversity we use K-Medoids clustering. K-Medoids is a partition-based clustering that uses the pair-wise distance of data points like K-Means but instead of computing the mean centroids to minimize the within-cluster sum of distances it uses the data points themselves as centroids. To use K-Medoids we need to specify an appropriate distance metric that can measure distance between two distributions. KL-divergence is a measure of difference between two distributions but it does not satisfy two properties of a distance metric, symmetry and triangle inequality. For our purposes, symmetry is a critical property. A similar and smoother measure called Jensen-Shannon divergence [14, 15] that is symmetric but still does not hold the triangle inequality can be used safely, since K-Medoids clustering does not exploit the latter property.

Therefore, in order to define the distributional similarities between categorical distributions of user profiles, we use Jensen-Shannon divergence. The distance between two users u, v is defined by:

$$D_{\mathcal{JS}}(u, v) = \frac{1}{2}\left[D_{\mathcal{KL}}(u \parallel avg_{u,v}) + D_{\mathcal{KL}}(v \parallel avg_{u,v})\right] \tag{3}$$

Where:

$$D_{\mathcal{KL}}(u \parallel v) = \sum_{c \in \mathcal{C}} P(c|u)\, log \frac{P(c|u)}{P(c|v)} \tag{4}$$

$$avg_{u,v}(c) = \frac{1}{2}(P(c|u) + P(c|v)) \tag{5}$$

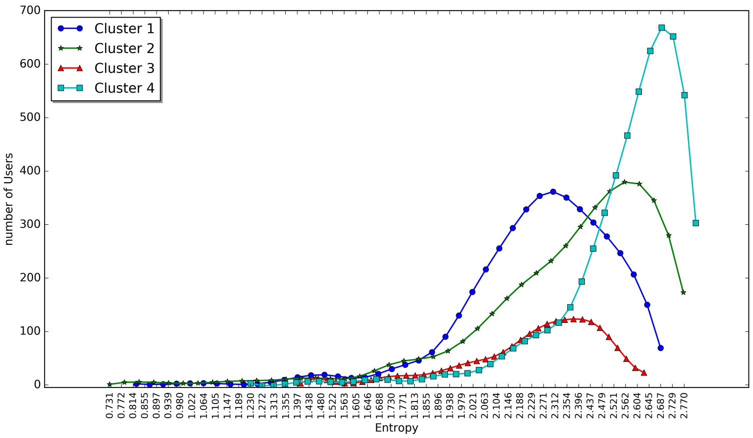

Figure 2: User clusters in MovieLens

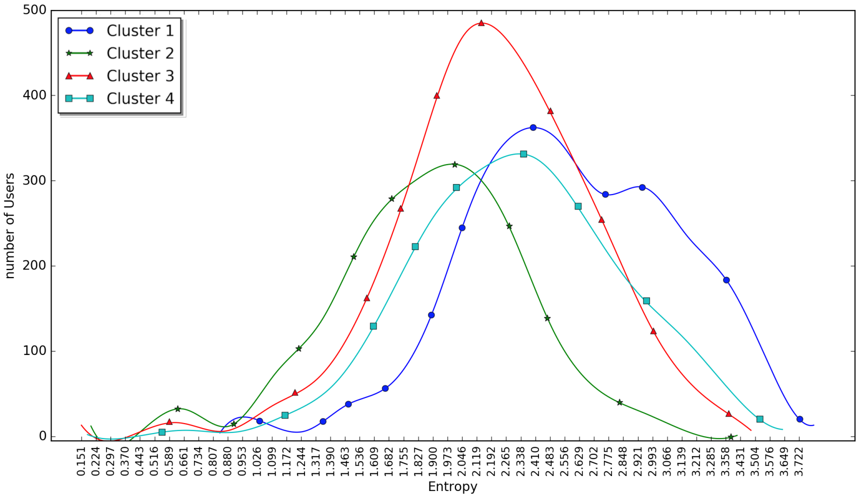

Figure 3: User clusters in Yelp.

4 Evaluation

In this section, we discuss about our experiments design and our evaluation results.

4.1 Dataset

MovieLens 1M [16] is a widely known dataset in literature. It contains about 1 million ratings from 6,040 users and 3,706 movies. There are a total of 18 binary indicated movie genres available in this dataset. The second dataset is Yelp restaurants reviews that contains 8,022 users and 5,199 businesses based on 104,576 ratings. For this dataset we have only focused on categories of restaurants (cuisine type) which number around 120. All of the ratings for both data sets are in the range of 1 to 5.

4.2 Experimental Setup

For each dataset, 80% of the ratings is used for training sets and the remaining 20% is used for testing sets. The standard collaborative recommender used in our experiments is user-based kNN (*Standard*) [17] as a baseline method. Also, we use the same recommender system for each segment separately, after clustering users (*Div-Clust*) to compare the results both in terms of accuracy and diversity. The recommendation list size in all of our experiments is fixed at 20 and the number of neighbors is fixed to 25.

4.3 Metrics

In order to evaluate the result, we use three measures of rankings. First, the standard *NDCG* metric is used to measure only the accuracy of recommendation list in terms of user's preferences on recommended items. The second, metric is *NDCG_IA* [6]; an elegant metric that measures the accuracy of results for each category based on $P(c|u)$. This metric captures the degree of personalization of diversity in the recommendation list based on the user's past tolerance for diversity. But, it also captures relevance of the results in terms of test items ranked in each category for each user. Given $P(c|u)$, For each category preference of user, we treat any item that does not belong to this category as irrelevant and compute its category-dependent rank at cutoff k as $NDCG(u, k|c)$. Therefore, *NDCG_IA* is defined as follows:

$$NDCG\_IA(u,k) = \sum_{c} P(c|u) NDCG(u,k|c) \tag{6}$$

The third metric *α-NDCG* is used only for the purpose of measuring personalization in diversity, but not relevance of the results. *α-NDCG* was originally introduced in the context of information retrieval [18] to measure both relevance and redundancy in terms of information nuggets (topics) associated with search results.

Our work heavily relies on categorical distribution of user profiles to measure the degree to which diversity is personalized for each user. To use this metric for our purposes, we need to adapt it to the recommender systems context. In our context, α is a factor to penalize the redundancy of items in a rank list. We define the *gain vector* of *α-NDCG* as follows:

$$G[k] = \sum_{c \in \mathcal{C}} P(i_k \in c) P(c|u) (1-\alpha)^{r_{c,k-1}} \tag{7}$$

Where $r_{c,0} = 0$ and

$$r_{c,k-1} = \sum_{j=1}^{k-1} P(i_j \in c), \tag{8}$$

Essentially, this represents the number of items ranked up to position $k-1$ that is known to contain category c according to $P(i \in c)$. Note that large values of α diminish the influence of personalization factor $P(c|u)$ and thus *α-NDCG* is only a measure of diversity based on categories.

4.4 Results and Discussions

In this section we discuss the results of performing collaborative filtering with and without the proposed segmentation approach.

	Size	Entropy	Method	NDCG	NDCG_IA	α-NDCG
Cluster 3	714	2.252	Div-Clust	0.203	**0.189**	**0.683**
			Standard	**0.232**	0.182	0.616
Cluster 1	1604	2.264	Div-Clust	0.285	**0.261**	**0.802**
			Standard	**0.292**	0.255	0.791
Cluster 2	1580	2.377	Div-Clust	0.227	**0.186**	**0.714**
			Standard	**0.229**	0.163	0.625
Cluster 4	2142	2.567	Div-Clust	0.195	0.137	**0.680**
			Standard	**0.231**	**0.158**	0.664

Table 1: MovieLens results for α=0.6.

	Method	α=0.1	α=0.3	α=0.6	α=0.9
Cluster 3	Div-Clust	**0.5631**	**0.6608**	**0.6826**	**0.6686**
	Standard	0.5062	0.5864	0.6159	0.6178
Cluster 1	Div-Clust	**0.7698**	**0.8014**	**0.8021**	**0.7887**
	Standard	0.7469	0.7901	0.7917	0.7806
Cluster 2	Div-Clust	**0.6098**	**0.6959**	**0.7143**	**0.7034**
	Standard	0.5358	0.6089	0.6251	0.6173
Cluster 4	Div-Clust	0.5885	0.6631	**0.6798**	**0.6721**
	Standard	**0.6204**	**0.6662**	0.6642	0.6478

Table 2: MovieLens results for α-*NDCG* with different α values.

After modeling user profile distributions based on categories in each of two datasets, we plotted the entropy distribution of users in each cluster to determine the effectiveness of clustering in capturing different levels of interest in diversity. In Figures 2 and 3, we can see various peaks for each cluster that show the density of users at different entropy values. The clusters in all of our results, are ordered based on the mean entropy values of each cluster.

The results of our experiments are shown in Tables 1 and 3 for MovieLens and Yelp datasets, respectively. In all of the tables, clusters are ordered based on the average entropy of user profiles in them. In Table 1, the results for *NDCG* does not show any significant changes for clusters 1 and 2. For the other two cluster we see a little drop in accuracy. In terms of *NDCG_IA* the results of all the clusters except cluster 4 is higher in *Div-Cluster*. This indicates that, overall, the clusters result in a fairly accurate representation of user interests across various categories. The reason for the anomalous behavior of cluster 4 is that, as indicated in Figure 2, this cluster has the highest average entropy among all other clusters. This means that there is a high degree of uncertainty about the interest of users in diversity in that cluster. In this situation, personalization based on preference distributions is ineffective.

In terms of α-*NDCG* we get improvements for all of the segments indicating that the segmentation is effective in personalizing diversity. In Table 2 we show the results of various α values to see trade-off behavior between redundancy of content and personalization of diversity. When $\alpha = 0.1$ the penalty for redundancy will be very small. Larger α values result in a higher degree of personalization based on the distribution of categories. We see this effect on cluster 4 when we look at Table 2. In this table the results of this cluster for *Div-Clust* is lower than *Standard* when α is small. As α increases the results get reversed. This shows that for this cluster redundancy in recommendations is higher for *Standard* and

	Size	Entropy	Method	NDCG	NDCG_IA	α-NDCG
Cluster 2	1592	1.884	Div-Clust	0.0066	0.2133	**0.3648**
			Standard	**0.0101**	**0.2659**	0.3628
Cluster 3	2329	2.201	Div-Clust	0.0231	**0.2513**	**0.5725**
			Standard	**0.0233**	0.2471	0.4753
Cluster 4	1765	2.323	Div-Clust	**0.0173**	**0.2068**	**0.5371**
			Standard	0.0143	0.1899	0.4609
Cluster 1	2336	2.602	Div-Clust	**0.0298**	**0.1831**	**0.6835**
			Standard	0.0266	0.1351	0.5168

Table 3: Yelp results for α=0.3.

	Method	α=0.1	α=0.3	α=0.6	α=0.9
Cluster 2	Div-Clust	**0.3052**	**0.3648**	0.4035	0.4222
	Standard	0.2973	0.3628	**0.4084**	**0.4286**
Cluster 3	Div-Clust	**0.4937**	**0.5725**	**0.5959**	**0.6016**
	Standard	0.3801	0.4753	0.5262	0.5432
Cluster 4	Div-Clust	**0.4519**	**0.5371**	**0.5777**	**0.5906**
	Standard	0.3833	0.4609	0.4942	0.5037
Cluster 1	Div-Clust	**0.5981**	**0.6835**	**0.7162**	**0.7236**
	Standard	0.4191	0.5168	0.5624	0.5738

Table 4: Yelp results for α-*NDCG* with different α values.

personalization of categories is not strong enough to increase the value of α-*NDCG* in *Div-Clust*.

The same methodology was used to evaluate the results of the Yelp dataset. These results are depicted in Tables 3 and 4. Except for cluster 2, We see better results in terms of both *NDCG_IA* and α-*NDCG* for this dataset. Even for some clusters we see higher *NDCG*. The main reason for this behavior is larger variance in the entropy distribution that we see in Figure 1 compared to MovieLens. Most probably the results of cluster 2 is due to noise in that cluster since we get low accuracy even for *Standard* recommender.

5 Conclusions

In this work, we have proposed a new approach for personalizing diversity by clustering the users based on their tolerance for diversity. In contrast to most of the work in this area that uses a greedy re-ranking approach for diversification, we use a pre-filtering approach that can be integrated into any collaborative recommender system. The adapted α-*NDCG* metric shows the effectiveness of our method.

References

[1] S. M. McNee, J. Riedl, J. A. Konstan, Being accurate is not enough: how accuracy metrics have hurt recommender systems, in: CHI'06 extended abstracts on Human factors in computing systems, ACM, 2006, pp. 1097–1101.

[2] C.-N. Ziegler, S. M. McNee, J. A. Konstan, G. Lausen, Improving recommendation lists through topic diversification, in: Proceedings of the 14th international conference on

World Wide Web, ACM, 2005, pp. 22–32.

[3] M. Zhang, N. Hurley, Avoiding monotony: improving the diversity of recommendation lists, in: Proceedings of the 2008 ACM conference on Recommender systems, ACM, 2008, pp. 123–130.

[4] J. Carbonell, J. Goldstein, The use of mmr, diversity-based reranking for reordering documents and producing summaries, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 1998, pp. 335–336.

[5] S. Vargas, P. Castells, D. Vallet, Intent-oriented diversity in recommender systems, in: Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval, ACM, 2011, pp. 1211–1212.

[6] R. Agrawal, S. Gollapudi, A. Halverson, S. Ieong, Diversifying search results, in: Proceedings of the second ACM international conference on web search and data mining, ACM, 2009, pp. 5–14.

[7] N. J. Hurley, Personalised ranking with diversity, in: Proceedings of the 7th ACM Conference on Recommender Systems, ACM, 2013, pp. 379–382.

[8] J. Wasilewski, N. Hurley, Incorporating diversity in a learning to rank recommender system., in: FLAIRS Conference, 2016, pp. 572–578.

[9] T. Di Noia, V. C. Ostuni, J. Rosati, P. Tomeo, E. Di Sciascio, An analysis of users' propensity toward diversity in recommendations, in: Proceedings of the 8th ACM Conference on Recommender Systems, ACM, 2014, pp. 285–288.

[10] S. Vargas, L. Baltrunas, A. Karatzoglou, P. Castells, Coverage, redundancy and size-awareness in genre diversity for recommender systems, in: Proceedings of the 8th ACM Conference on Recommender systems, ACM, 2014, pp. 209–216.

[11] R. L. Santos, C. Macdonald, I. Ounis, Exploiting query reformulations for web search result diversification, in: Proceedings of the 19th international conference on World wide web, ACM, 2010, pp. 881–890.

[12] S. Vargas, P. Castells, Exploiting the diversity of user preferences for recommendation, in: Proceedings of the 10th Conference on Open Research Areas in Information Retrieval, LE CENTRE DE HAUTES ETUDES INTERNATIONALES D'INFORMATIQUE DOCUMENTAIRE, 2013, pp. 129–136.

[13] F. Eskandanian, B. Mobasher, R. D. Burke, User segmentation for controlling recommendation diversity, in: Proceedings of the Poster Track of the 10th ACM Conference on Recommender Systems (RecSys 2016), Boston, USA, September 17, 2016., 2016.
URL http://ceur-ws.org/Vol-1688/paper-24.pdf

[14] J. Lin, Divergence measures based on the shannon entropy, IEEE Transactions on Information theory 37 (1) (1991) 145–151.

[15] L. Lee, Measures of distributional similarity, in: Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics, Association for Computational Linguistics, 1999, pp. 25–32.

[16] F. M. Harper, J. A. Konstan, The movielens datasets: History and context, ACM Transactions on Interactive Intelligent Systems (TiiS) 5 (4) (2016) 19.

[17] X. Ning, C. Desrosiers, G. Karypis, A comprehensive survey of neighborhood-based recommendation methods, in: Recommender systems handbook, Springer, 2015, pp. 37–76.

[18] C. L. Clarke, M. Kolla, G. V. Cormack, O. Vechtomova, A. Ashkan, S. Büttcher, I. MacKinnon, Novelty and diversity in information retrieval evaluation, in: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval, ACM, 2008, pp. 659–666.