Making Product Recommendations More Diverse

Cai-Nicolas Ziegler* Georg Lausen

Institut für Informatik, Universität Freiburg
Georges-Köhler-Allee, Gebäude 51
79110 Freiburg i. Br., Germany

{cziegler,lausen}@informatik.uni-freiburg.de

Abstract

Past research in recommender systems has mainly focused on improving accuracy, i.e., making each single recommendation get as close to the user's information need as possible. However, while this approach works well when focusing on single recommendations as atomic entities, its usefulness to the consumer appears limited when considering entire recommendation lists along with their overall utility, which often appear to provide rather an unbalanced diet to the user: Recommendation lists seldom reflect the consumer's entire spectrum of interest but rather hook on to small portions that appear particularly favorable with regard to accuracy optimization. We analyze the diversification issue in detail and present a framework that is geared towards making lists as interesting and colorful as possible, trading a minimum of accuracy in exchange for the gain in diversity. Empirical evaluations aiming for actual user satisfaction underpin the cogency of our approach.

1 Introduction

Recommender systems [10] intend to provide people with recommendations of products they will appreciate, based on their past preferences. Many of the most successful systems make use of collaborative filtering [6], and numerous commercial systems, e.g., Amazon.com's recommender [9], exploit these techniques to offer personalized recommendation lists to their customers. Though the *accuracy* of state-of-the-art collaborative filtering systems, i.e., the probability that the active user[1] will appreciate the products recommended, is excellent, some implications affecting user satisfaction have been observed in practice. Thus, on Amazon.com (*http://www.amazon.com*), many recommendations seem to be "similar" with respect to content. For instance, customers that have purchased many of Hermann Hesse's prose may happen to obtain recommendation lists where all top-5 entries contain books by that respective author only. When considering pure accuracy, all these recommendations appear excellent since the active user clearly appreciates books written by Hermann Hesse.

Bulletin of the IEEE Computer Society Technical Committee on Data Engineering

*Now working at The Boston Consulting Group, Munich. Contact at ziegler.cai-nicolas@bcg.com.

[1]The term "active user" refers to the person for whom recommendations are made.

On the other hand, assuming that the active user has several interests other than Hermann Hesse, e.g., historical novels, the recommended set of items appears poor, owing to its lack of diversity.

Traditionally, recommender system projects have focused on optimizing accuracy. Now research has reached the point where going beyond pure accuracy and toward real user experience becomes indispensable for further advances. This work looks specifically at impacts of recommendation *lists*, regarding them as entities in their own right rather than mere aggregations of single and independent suggestions.

2 On Collaborative Filtering

Collaborative filtering (CF) still represents the most commonly adopted technique in recommender systems. Its basic idea is to make recommendations based upon ratings that users have assigned to products. Ratings can either be explicit, i.e., by having the user state his opinion about a given product, or implicit, when the mere act of purchasing an item counts as an expression of appreciation. While implicit ratings are more facile to collect, their usage implies adding noise to the collected information.

User-based Collaborative Filtering has been explored in-depth during the last 10-15 years and is the most popular recommendation algorithm [6], owing to its compelling simplicity and excellent quality of recommendations. CF operates on a set of users $A = \{a_1, a_2, \ldots, a_n\}$, a set of products $B = \{b_1, b_2, \ldots, b_m\}$, and partial rating functions $r_i : B \to [-1, +1]^{\perp}$ for each user a_i. Negative values $r_i(b_k)$ denote utter dislike, while positive values express a_i's liking of product b_k. If ratings are implicit only, we represent them by set $R_i \subseteq B$, equivalent to $\{b_k \in B \mid r_i(b_k) \neq \perp\}$. The user-based CF's working process can be broken down as follows:

- **Neighborhood formation.** Assuming a_i as the active user, similarity values $c(a_i, a_j) \in [-1, +1]$ for all $a_j \in A \setminus \{a_i\}$ are computed, based upon the similarity of their respective rating functions r_i, r_j. In general, Pearson correlation [12, 4] or cosine distance [6] are used for computing $c(a_i, a_j)$. The top-M most similar users a_j become members of a_i's neighborhood, $clique(a_i) \subseteq A$.
- **Rating prediction.** Taking all the products b_k that a_i's neighbors $a_j \in clique(a_i)$ have rated and which are new to a_i, i.e., $r_i(b_k) = \perp$, a prediction of liking $w_i(b_k)$ is produced. Value $w_i(b_k)$ hereby depends on both the similarity $c(a_i, a_j)$ of a_j with $r_j(b_k) \neq \perp$, as well as the ratings $r_j(b_k)$ these a_j assigned to b_k.

Eventually, a list $P_{w_i} : \{1, 2, \ldots, N\} \to B$ of top-N recommendations is computed, based on predictions w_i. P_{w_i} is injective and reflects recommendation ranking in *descending* order, giving highest predictions first.

Item-based CF has favorable computational complexity characteristics and allows to decouple the model computation process from actual prediction making. Specifically for cases where $|A| \gg |B|$, item-based CF's computational performance has been shown superior to user-based CF [11]. Its success also extends to many commercial recommender systems, such as Amazon.com's [9].

As with user-based CF, recommendation making is based upon ratings $r_i(b_k)$ that users $a_i \in A$ provided for products $b_k \in B$. However, unlike user-based CF, similarity values c are computed for *items* rather than *users*, hence $c : B \times B \to [-1, +1]$. Roughly speaking, two items b_k, b_e are similar, i.e., have large $c(b_k, b_e)$, if users who rate one of them tend to rate the other, and if users tend to assign them identical or similar ratings. Moreover, for each b_k, its neighborhood $clique(b_k) \subseteq B$ of top-M most similar items is defined.

3 Evaluation Metrics

Evaluation metrics are essential in order to judge the performance of recommender systems, even though they are still in their infancies. Most evaluations concentrate on accuracy measurements only and neglect other fac-

tors, e.g., novelty and serendipity of recommendations, and the diversity of the recommended list's items. The following sections give a short overview of popular metrics. An extensive survey of accuracy metrics is in [7].

3.1 Accuracy Metrics

Accuracy metrics have been proposed for two major tasks: First, to judge the **accuracy of single predictions**, i.e., how much predictions $w_i(b_k)$ for products b_k deviate from a_i's actual ratings $r_i(b_k)$. These metrics are particularly suited for tasks where predictions are displayed along with the product, e.g., annotation in context [7]. Examples include the mean absolute error (MAE), and mean squared error (MSE).

Second, **decision-support metrics** evaluate the effectiveness of helping users to select high-quality items from the set of all products, generally supposing binary preferences. Typical decision-support metrics include the well-known precision and recall metrics, known from information retrieval, as well as ROC, the "receiver operating characteristic" (see, e.g., [5]).

3.2 Beyond Accuracy

Though accuracy metrics are an important facet of usefulness, there are traits of user satisfaction they are unable to capture. However, non-accuracy metrics have largely been denied major research interest so far. Among all non-accuracy evaluation metrics, **coverage** has been the most frequently used [6, 5]. Coverage measures the percentage of elements part of the problem domain for which predictions can be made.

Some recommenders produce highly accurate results that are still useless in practice, e.g., suggesting bananas to customers in grocery stores. Though being highly accurate, note that almost everybody likes and buys bananas. Hence, their recommending appears far too obvious and of little help to the shopper. **Novelty** and **serendipity** metrics thus measure the "non-obviousness" of recommendations made, avoiding "cherry-picking" [7]. For some simple measure of serendipity, take the average popularity of recommended items.

3.3 Intra-List Similarity

We present a new metric that intends to capture the diversity of a list. Hereby, diversity may refer to all kinds of features, e.g., genre, author. Based upon an arbitrary function $c_\circ : B \times B \rightarrow [-1, +1]$ measuring the similarity $c_\circ(b_k, b_e)$ between products b_k, b_e according to some custom-defined criterion, we define intra-list similarity for a_i's list P_{w_i} as follows:

$$ILS(P_{w_i}) = \frac{\sum_{b_k \in \Im P_{w_i}} \sum_{b_e \in \Im P_{w_i}, b_k \neq b_e} c_\circ(b_k, b_e)}{2} \tag{1}$$

Symbol $\Im P_{w_i}$ denotes the *image* of map P_{w_i}, i.e., all items part of the recommendation list. Higher scores of $ILS(P_{w_i})$ denote lower diversity. An interesting mathematical feature of $ILS(P_{w_i})$ we are referring to in later sections is permutation-insensitivity, i.e., let S_N be the symmetric group of all permutations on $N = |P_{w_i}|$ symbols, then $\forall \sigma_i, \sigma_j \in S_N : ILS(P_{w_i} \circ \sigma_i) = ILS(P_{w_i} \circ \sigma_j)$ holds. Hence, simply rearranging positions of recommendations in a top-N list P_{w_i} does not affect P_{w_i}'s intra-list similarity.

4 Topic Diversification

One major issue with accuracy metrics is their inability to capture the broader aspects of satisfaction, hiding blatant flaws in existing systems. For instance, suggesting a list of very similar items, e.g., with respect to the author, genre, or topic, may be of little use for the user, even though this list's average accuracy might be high. The issue has been perceived by other researchers before, coined "portfolio effect" by Ali and van Stam [1]. We believe that item-based CF systems in particular are susceptible to that effect.

procedure diversify (P_{w_i}, Θ_F) {
  $B_i \leftarrow \Im P_{w_i}$; $P_{w_i*}(1) \leftarrow P_{w_i}(1)$;
  for $z \leftarrow 2$ **to** N **do**
    set $B_i' \leftarrow B_i \setminus \{P_{w_i*}(k) \mid k \in [1, z[\,\}$;
    $\forall b \in B'$: compute $c^*(\{b\}, \{P_{w_i*}(k) \mid k \in [1, z[\,\})$;
    compute $P_{c^*} : \{1, 2, \ldots, |B_i'|\} \rightarrow B_i'$ using c^*;
    for all $b \in B_i'$ **do**
      $P_{c^*}^{\text{rev}^{-1}}(b) \leftarrow |B_i'| - P_{c^*}^{-1}(b)$;
      $w_i^*(b) \leftarrow P_{w_i}^{-1}(b) \cdot (1 - \Theta_F) + P_{c^*}^{\text{rev}^{-1}}(b) \cdot \Theta_F$;
    end do
    $P_{w_i*}(z) \leftarrow min\{w_i^*(b) \mid b \in B_i'\}$;
  end do
  return P_{w_i*};
}

Algorithm 1: Sequential topic diversification

We propose an approach we call *topic diversification* to deal with the problem at hand and make recommended lists more diverse and thus more useful. Our method represents an extension to existing recommender algorithms and is applied on top of recommendation lists.

4.1 Taxonomy-based Similarity Metric

Function $c^* : 2^B \times 2^B \rightarrow [-1, +1]$, quantifying the similarity between two product sets, forms an essential part of topic diversification. We instantiate c^* with our metric for taxonomy-driven filtering [13], though other content-based similarity measures may appear likewise suitable. Our metric computes the similarity between product sets based upon their classification. Each product belongs to one or more classes that are hierarchically arranged in classification taxonomies, describing the products in machine-readable ways.

Classification taxonomies exist for various domains. Amazon.com crafts very large taxonomies for books, DVDs, CDs, electronic goods, and apparel. Moreover, all products on Amazon.com bear content descriptions relating to these domain taxonomies. Featured topics could include author, genre, and audience.

4.2 Topic Diversification Algorithm

Algorithm 1 shows the complete topic diversification algorithm, a brief textual sketch is given in the next paragraphs.

Function P_{w_i*} denotes the new recommendation list, resulting from applying topic diversification. For every list entry $z \in [2, N]$, we collect those b from the candidate products set B_i that do not occur in positions $o < z$ in P_{w_i*} and compute their similarity with $\{P_{w_i*}(k) \mid k \in [1, z[\,\}$ containing all new recommendations preceding rank z.

Sorting all products b according to $c^*(b)$ in reverse order, we obtain the *dissimilarity rank* $P_{c^*}^{\text{rev}}$. This rank is then merged with the original recommendation rank P_{w_i} according to diversification factor Θ_F, yielding final rank P_{w_i*}. Factor Θ_F defines the *impact* that dissimilarity rank $P_{c^*}^{\text{rev}}$ exerts on the eventual overall output.

Large $\Theta_F \in [0.5, 1]$ favors diversification over a_i's original relevance order, while low $\Theta_F \in [0, 0.5[$ produces recommendation lists closer to the original rank P_{w_i}. For experimental analysis, we used factors $\Theta_F \in [0, 0.9]$.

Note that ordered input lists P_{w_i} must be considerably larger than the final top-N list. For our later experiments, we used top-50 input lists for eventual top-10 recommendations.

The effect of dissimilarity bears traits similar to that of **osmotic pressure** and selective permeability known from molecular biology; this concept allows cells (i.e, recommendation lists) to maintain their internal composition of substances (i.e., topics) at required levels (i.e., gauging accuracy versus dissimilarity) [14].

5 Empirical Analysis

We conducted offline evaluations to understand the ramifications of topic diversification on accuracy metrics, and online analysis to investigate how our method affects actual user satisfaction. We applied topic diversification with $\Theta_F \in \{0, 0.1, 0.2, \ldots 0.9\}$ to lists generated by both user-based CF and item-based CF, observing effects that occur when steadily increasing Θ_F and analyzing how both approaches respond to diversification.

We based online and offline analyses on data gathered from BookCrossing (*http://www.bookcrossing.com*). The latter community caters for book lovers exchanging books all around the world and sharing their experiences with others. We collected data on $278,858$ members of BookCrossing and $1,157,112$ ratings, both implicit and explicit, referring to $271,379$ distinct ISBNs. Invalid ISBNs were excluded from the outset. The complete BookCrossing dataset, featuring fully anonymized information, is available via the first author's homepage (*http://www.informatik.uni-freiburg.de/∼cziegler*).

Next, we mined Amazon.com's book taxonomy, comprising 13,525 distinct topics. In order to be able to apply topic diversification, we mined content information, focusing on taxonomic descriptions that relate books to taxonomy nodes from Amazon.com. Since many books on BookCrossing refer to rare, non-English books, or outdated titles not in print anymore, we were able to garner background knowledge for only $175,721$ books. In total, $466,573$ topic descriptors were found, giving an average of 2.66 topics per book.

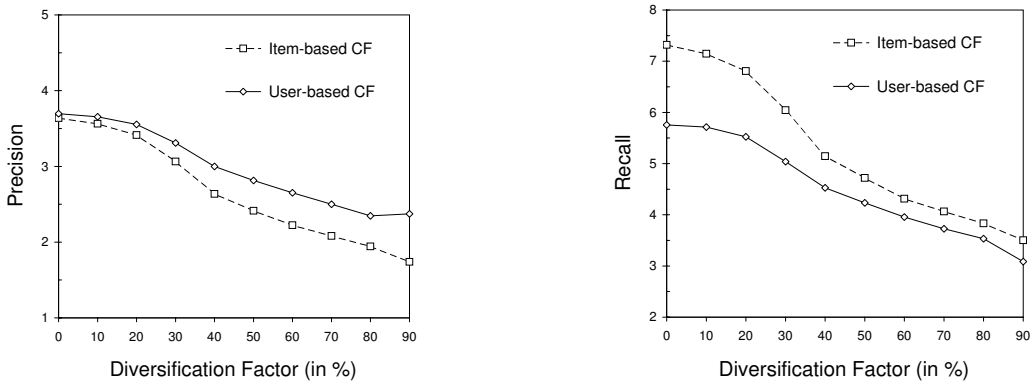
Owing to the BookCrossing dataset's extreme sparsity, we decided to further condense the set in order to obtain more meaningful results from CF algorithms when computing recommendations. Hence, we discarded all books missing taxonomic descriptions, along with all ratings referring to them. Next, we also removed book titles with fewer than 20 overall mentions. Only community members with at least 5 ratings each were kept.

The resulting dataset's dimensions were considerably more moderate, featuring $10,339$ users, $6,708$ books, and $361,349$ book ratings.

5.1 Offline Experiments

We performed offline experiments comparing precision, recall, and intra-list similarity scores for 20 different list setups. Half these recommendation lists were based upon user-based CF with different degrees of diversification, the others on item-based CF. Note that we did not compute MAE metric values since we are dealing with implicit rather than explicit ratings.

For cross-validation of precision and recall metrics of all $10,339$ users, we adopted K-folding with parameter $K = 4$. Hence, rating profiles R_i were effectively split into training sets R_i^x and test sets $T_i^x, x \in \{1, \ldots, 4\}$, at a ratio of $3 : 1$. For each of the $41,356$ different training sets, we computed 20 top-10 recommendation lists. To generate the diversified lists, we computed top-50 lists based upon pure, i.e., non-diversified, item-based CF and pure user-based CF. Next, we applied the diversification algorithm to both base cases, applying Θ_F factors ranging from 10% up to 90%. For evaluation, all lists were truncated to contain 10 books only.



(a) (b)

Figure 1: Precision (a) and recall (b) for increasing Θ_F

5.1.1 Result Analysis

We were interested in seeing how accuracy, captured by precision and recall, behaves when increasing Θ_F from 0.1 up to 0.9. Since topic diversification may make books with high predicted accuracy trickle down the list, we hypothesized that accuracy will *deteriorate* for $\Theta_F \rightarrow 0.9$. Moreover, in order to find out if our novel algorithm has any significant, positive effects on the diversity of items featured, we also applied our intra-list similarity metric. An overlap analysis for diversified lists, $\Theta_F \geq 0.1$, versus their respective non-diversified pendants indicates how many items stayed the same for increasing diversification factors.

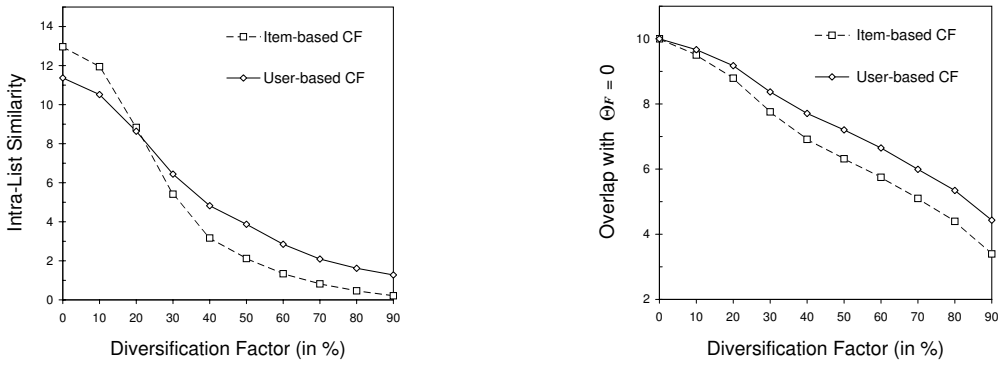
Precision and Recall. First, we analyzed precision and recall scores for both non-diversified base cases, i.e., when $\Theta_F = 0$. For item-based CF we had a precision of 3.64, and recall of 7.32. For user-based CF, the results were 3.69 and 5.76, respectively. Thus, user-based and item-based CF exhibit almost identical accuracy, indicated by precision values. Their recall values differ considerably, hinting at deviating behavior with respect to the types of users they are scoring for.

Next, we analyzed the behavior of user-based and item-based CF when steadily increasing Θ_F by increments of 10%, depicted in Figure 1. The two charts reveal that diversification has detrimental effects on both metrics and on both CF algorithms. Interestingly, corresponding precision and recall curves have almost identical shape.

The loss in accuracy is more pronounced for item-based than for user-based CF. Furthermore, for either metric and either CF algorithm, the drop is most distinctive for $\Theta_F \in [0.2, 0.4]$. For lower Θ_F, negative impacts on accuracy are marginal. We believe this last observation due to the fact that precision and recall are permutation-insensitive, i.e., the mere order of recommendations within a top-N list does not influence the metric value, as opposed to Breese score [2, 7]. However, for low Θ_F, the pressure that the dissimilarity rank exerts on the top-N list's makeup is still too weak to make many new items diffuse into the top-N list. Hence, we conjecture that rather the *positions* of current top-N items change, which does not affect either precision or recall.

Intra-List Similarity. Knowing that our diversification method exerts a significant, negative impact on accuracy metrics, we wanted to know how our approach affected the intra-list similarity measure. Similar to the precision and recall experiments, we computed values for user-based and item-based CF with $\Theta_F \in [0, 0.9]$ each. Results obtained from intra-list similarity analysis are given in Figure 2(*a*).

The topic diversification method lowers the pairwise similarity between list items, thus making top-N recommendation lists more diverse. Diversification appears to affect item-based CF stronger than its user-based counterpart, in line with our findings about precision and recall. For lower Θ_F, curves are less steep than



(a) (b)

Figure 2: Intra-list similarity behavior (a) and overlap with original list (b) for increasing Θ_F

for $\Theta_F \in [0.2, 0.4]$, which also well aligns with precision and recall analysis. Again, the latter phenomenon can be explained by one of the metric's inherent features, i.e., like precision and recall, intra-list similarity is permutation-insensitive.

Figure 2(*b*) shows the number of recommended items staying the same when increasing Θ_F with respect to the original list's content. Both curves exhibit roughly linear shapes, being less steep for low Θ_F, though. Interestingly, for factors $\Theta_F \leq 0.4$, at most 3 recommendations change on average.

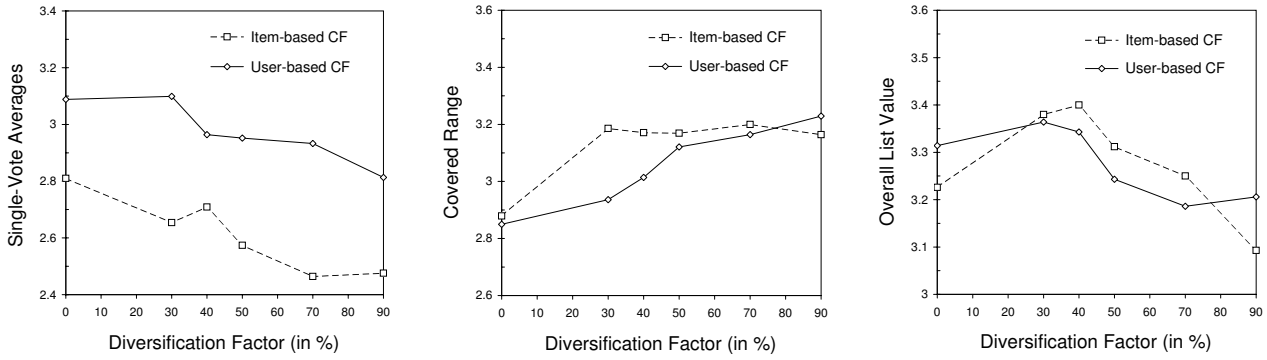
5.2 Conclusion

We found that diversification appears detrimental to both user-based and item-based CF along precision and recall metrics. In fact, this outcome aligns with our expectations, considering the nature of those two accuracy metrics and the way that the topic diversification method works. Moreover, we found that item-based CF seems more susceptible to topic diversification than user-based CF, backed by results from precision, recall and intra-list similarity metric analysis.

5.3 Online Experiments

Offline experiments helped us in understanding the implications of topic diversification on both CF algorithms. We could also observe that the effects of our approach are different on different algorithms. However, we wanted to assess actual user satisfaction for various degrees of diversification, thus necessitating an online survey. For the online study, we computed each recommendation list type anew for users in the denser BookCrossing dataset, though without K-folding. We mailed all eligible users via the community mailing system, asking them to participate in our online study. Each mail contained a personal link to our online survey pages. In order to make sure that only the users themselves would complete their survey, links contained unique, encrypted access codes. During the 3-week survey phase, $2,125$ users participated and completed the study.

The survey consisted of several screens that would tell the prospective participant about this study's nature and his task, show all his ratings used for making recommendations, and finally present a top-10 recommendation list, asking several questions thereafter. For each book, users could state their interest on a 5-point rating scale. Scales ranged from "not much" to "very much", mapped to values 1 to 4, and offered the user to indicate that he had already read the book, mapped to value 5. In order to successfully complete the study, users were not required to rate all their top-10 recommendations. Neutral values were assumed for non-votes instead. However, we required users to answer all further questions, concerning the list as a whole rather than its single recommendations, before submitting their results.



(a) (b) (c)

Figure 3: Results for single-vote averages (a), covered range of interests (b), and overall satisfaction (c)

The one top-10 list for each user was chosen among 12 candidate lists, either user-based CF or item-based with $\Theta_F \in \{0, 0.3, 0.4, 0.5, 0.7, 0.9\}$ each. The assignment of a specific list to the current user was done dynamically, at the time of the participant entering the survey, and in a round-robin fashion. Thus, we could guarantee that the number of users per list type was roughly identical.

5.3.1 Result Analysis

For the analysis of our inter-subject survey, we were mostly interested in the following three aspects: First, the average rating users gave to their 10 single recommendations. We expected results to roughly align with scores obtained from precision and recall. Second, we wanted to know if users perceived their list as well-diversified, asking them to tell whether the lists reflected rather a broad or narrow range of their interests. Referring to the intra-list similarity metric, we expected users' perceived range of topics to increase with increasing Θ_F. Third, we were curious about the overall satisfaction of users with their lists, the measure to compare performance.

Both latter-mentioned questions were answered by each user on a 5-point likert scale (higher scores denoting better performance) and we averaged the eventual results by the number of users. Statistical significance of all mean values was measured by parametric one-factor ANOVA, where $p < 0.05$ if not indicated otherwise.

Single-Vote Averages. Users perceived recommendations made by user-based CF systems on average as more accurate than those made by item-based CF systems, see Fig. 3(*a*). At each featured diversification level Θ_F, differences between the two CF types are significant, $p \ll 0.01$. Moreover, for each algorithm, higher diversification factors obviously entail lower single-vote average scores, which confirms our hypothesis stated before. The item-based CF's cusp at $\Theta_F \in [0.3, 0.5]$ appears as a notable outlier, but differences between the 3 means at $\Theta_F \in [0.3, 0.5]$ are not statistically significant, $p > 0.15$. Contrarily, differences between all factors Θ_F are significant for item-based CF, $p \ll 0.01$, and for user-based CF, $p < 0.1$.

Hence, topic diversification *negatively* correlates with pure accuracy. Besides, users perceived the performance of user-based CF as significantly better than item-based CF for all corresponding levels Θ_F.

Covered Range. Next, we analyzed if users actually *perceived* the variety-augmenting effects caused by topic diversification, illustrated before through the measurement of intra-list similarity. Users' reactions to steadily incrementing Θ_F are illustrated in Figure 3(*b*). First, between both algorithms on corresponding Θ_F levels, only the difference of means at $\Theta_F = 0.3$ shows statistical significance. Studying the trend of user-based CF for increasing Θ_F, we notice that the perceived range of reading interests covered by users' recommendation lists also increases. Hereby, the curve's first derivative maintains an approximately constant level, exhibiting slight peaks between $\Theta_F \in [0.4, 0.5]$. Statistical significance holds for user-based CF between means at $\Theta_F = 0$ and

$\Theta_F > 0.5$, and between $\Theta_F = 0.3$ and $\Theta_F = 0.9$. On the contrary, the item-based curve exhibits a drastically different behavior. While soaring at $\Theta_F = 0.3$ to 3.186, reaching a score almost identical to the user-based CF's peak at $\Theta_F = 0.9$, the curve barely rises for $\Theta_F \in [0.4, 0.9]$, remaining rather stable and showing a slight, though insignificant, upward trend. Statistical significance was shown for $\Theta_F = 0$ with respect to all other samples taken from $\Theta_F \in [0.3, 0.9]$. Hence, our online results do not perfectly align with findings obtained from offline analysis. While the intra-list similarity chart in Figure 2 indicates that diversity increases when increasing Θ_F, the item-based CF chart defies this trend, first soaring then flattening.

Overall List Value. The third feature variable, the overall value users assigned to their personal list, effectively represents the *target value* of our studies, measuring actual satisfaction. Owing to our conjecture that user satisfaction is a mere composite of accuracy and other factors, such as the list's diversity, we hypothesized that the application of topic diversification would *increase* satisfaction. At the same time, considering the downward trend of precision and recall for increasing Θ_F, in accordance with declining single-vote averages, we expected user satisfaction to drop off for large Θ_F. Hence, we supposed an arc-shaped curve for both algorithms.

Results for overall list value are given in Figure 3(c). For user-based CF we observe that the curve does *not* follow our hypothesis. Slightly improving at $\Theta_F = 0.3$ over the non-diversified case, scores drop for $\Theta_F \in [0.4, 0.7]$, culminating in a slight upturn at $\Theta_F = 0.9$. While lacking reasonable explanations, the curve's data-points de facto bear no statistical significance for $p < 0.1$. Hence, we conclude that topic diversification has a marginal, largely negligible impact on overall user satisfaction, initial positive effects eventually being offset by declining accuracy. On the contrary, for item-based CF, results look different. In compliance with our previous hypothesis, the curve's shape follows an arc, peaking at $\Theta_F = 0.4$. Taking the three data-points defining the arc, we obtain significance for $p < 0.1$. Since the endpoint's score at $\Theta_F = 0.9$ is inferior to the non-diversified case's, we find that too much diversification appears detrimental.

Eventually, for overall list value analysis, we come to conclude that topic diversification has no measurable effects on user-based CF, but significantly improves item-based CF performance for diversification factors Θ_F around 40%. In order to verify whether diversification appears as important ingredient of satisfaction, we also conducted multiple linear regresiion trials. Owing to space limitations, these trials are not reported here but can be accessed by resorting to [14].

6 Related Work

Few efforts have addressed the problem of diversifying top-N lists. Only considering literature on collaborative filtering and recommender systems in general, none have been presented before, to our best knowledge.

However, some work related to our topic diversification approach can be found in information retrieval, specifically meta-search engines. A critical aspect of meta-search engine design is the merging of several top-N lists into one single top-N list. Intuitively, this merged top-N list should reflect the highest quality ranking possible, also known as the "rank aggregation problem" [3].

More related to our idea of creating lists that represent the whole plethora of the user's topic interests, Kummamuru *et al.* [8] present their clustering scheme that groups search results into clusters of related topics. The user can then conveniently browse topic folders relevant to his search interest. The commercially available search engine NORTHERN LIGHT (*http://www.northernlight.com*) incorporates similar functionalities.

7 Conclusion

We presented topic diversification, an algorithmic framework to increase the diversity of a top-N list of recommended products. We also introduced our new intra-list similarity metric.

Contrasting precision and recall metrics for user-based and item-based CF with results obtained from a large-scale user survey, we showed that the user's overall liking of recommendation lists goes beyond accuracy and involves other factors, e.g., the users' perceived list diversity. We thus could demonstrate that lists are more than mere aggregations of single recommendations, but bear an intrinsic, added value.

Though effects of diversification were largely marginal on user-based CF, item-based CF performance improved significantly, an indication that there are some behavioral differences between both CF classes. Moreover, while pure item-based CF appeared slightly inferior to pure user-based CF in overall satisfaction, diversifying item-based CF with factors $\Theta_F \in [0.3, 0.4]$ made item-based CF outperform user-based CF. Interestingly for $\Theta_F \leq 0.4$, no more than three items tend to change with respect to the original list, shown in Figure 2. Small changes thus have high impact.

We believe our findings especially valuable for practical application scenarios, since many commercial recommender systems, e.g., Amazon.com [9] and TiVo [1], are item-based.

References

[1] ALI, K., AND VAN STAM, W. TiVo: Making show recommendations using a distributed collaborative filtering architecture. In *Proceedings of the 2004 ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Seattle, WA, USA, 2004), ACM Press, pp. 394–401.

[2] BREESE, J., HECKERMAN, D., AND KADIE, C. Empirical analysis of predictive algorithms for collaborative filtering. In *Proceedings of the Fourteenth Annual Conference on Uncertainty in Artificial Intelligence* (Madison, WI, USA, July 1998), Morgan Kaufmann, pp. 43–52.

[3] DWORK, C., KUMAR, R., NAOR, M., AND SIVAKUMAR, D. Rank aggregation methods for the Web. In *Proceedings of the Tenth International Conference on World Wide Web* (Hong Kong, China, 2001), ACM Press, pp. 613–622.

[4] GOLDBERG, D., NICHOLS, D., OKI, B., AND TERRY, D. Using collaborative filtering to weave an information tapestry. *Communications of the ACM 35*, 12 (1992), 61–70.

[5] GOOD, N., SCHAFER, B., KONSTAN, J., BORCHERS, A., SARWAR, B., HERLOCKER, J., AND RIEDL, J. Combining collaborative filtering with personal agents for better recommendations. In *Proceedings of the 16th National Conference on Artificial Intelligence and Innovative Applications of Artificial Intelligence* (Orlando, FL, USA, 1999), American Association for Artificial Intelligence, pp. 439–446.

[6] HERLOCKER, J., KONSTAN, J., BORCHERS, A., AND RIEDL, J. An algorithmic framework for performing collaborative filtering. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Berkeley, CA, USA, 1999), ACM Press, pp. 230–237.

[7] HERLOCKER, J., KONSTAN, J., TERVEEN, L., AND RIEDL, J. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems 22*, 1 (2004), 5–53.

[8] KUMMAMURU, K., LOTLIKAR, R., ROY, S., SINGAL, K., AND KRISHNAPURAM, R. A hierarchical monothetic document clustering algorithm for summarization and browsing search results. In *Proceedings of the 13th International Conference on World Wide Web* (New York, NY, USA, 2004), ACM Press, pp. 658–665.

[9] LINDEN, G., SMITH, B., AND YORK, J. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing 4*, 1 (January 2003).

[10] RESNICK, P., AND VARIAN, H. Recommender systems. *Communications of the ACM 40*, 3 (1997), 56–58.

[11] SARWAR, B., KARYPIS, G., KONSTAN, J., AND RIEDL, J. Item-based collaborative filtering recommendation algorithms. In *Proceedings of the Tenth International World Wide Web Conference* (Hong Kong, China, May 2001).

[12] SHARDANAND, U., AND MAES, P. Social information filtering: Algorithms for automating "word of mouth". In *Proceedings of the ACM CHI Conference on Human Factors in Computing Systems* (Denver, CO, USA, May 1995), ACM Press, pp. 210–217.

[13] ZIEGLER, C.-N., LAUSEN, G., AND SCHMIDT-THIEME, L. Taxonomy-driven computation of product recommendations. In *Proceedings of the 2004 ACM CIKM Conference on Information and Knowledge Management* (Washington, D.C., USA, November 2004), ACM Press, pp. 406–415.

[14] ZIEGLER, C.-N., MCNEE, S., KONSTAN, J., AND LAUSEN, G. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International World Wide Web Conference* (Chiba, Japan, May 2005), ACM Press.