Improving Recommendation Lists Through Topic Diversification

Cai-Nicolas Ziegler[1*] Sean M. McNee[2] Joseph A. Konstan[2] Georg Lausen[1]

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

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

[2] GroupLens Research, Univ. of Minnesota
4-192 EE/CS Building, 200 Union St. SE
Minneapolis, MN 55455, USA

{mcnee, konstan}@cs.umn.edu

ABSTRACT

In this work we present topic diversification, a novel method designed to balance and diversify personalized recommendation lists in order to reflect the user's complete spectrum of interests. Though being detrimental to average accuracy, we show that our method improves user satisfaction with recommendation lists, in particular for lists generated using the common item-based collaborative filtering algorithm.

Our work builds upon prior research on recommender systems, looking at properties of recommendation lists as entities in their own right rather than specifically focusing on the accuracy of individual recommendations. We introduce the intra-list similarity metric to assess the topical diversity of recommendation lists and the topic diversification approach for decreasing the intra-list similarity. We evaluate our method using book recommendation data, including offline analysis on 361,349 ratings and an online study involving more than 2,100 subjects.

Categories and Subject Descriptors

H.3.3 [**Information Storage and Retrieval**]: Information Retrieval and Search—*Information Filtering*; I.2.6 [**Artificial Intelligence**]: Learning—*Knowledge Acquisition*

General Terms

Algorithms, Experimentation, Human Factors, Measurement

Keywords

Collaborative filtering, diversification, accuracy, recommender systems, metrics

1. INTRODUCTION

Recommender systems [23] intend to provide people with recommendations of products they will appreciate, based on their past preferences, history of purchase, and demographic information. Many of the most successful systems make use of collaborative filtering [27, 8, 11], and numerous commercial systems, e.g., Amazon.com's recommender [16], 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. On the other hand, assuming that the active user has several interests other than Hermann Hesse, e.g., historical novels in general and books about world travel, the recommended set of items appears poor, owing to its lack of diversity.

Traditionally, recommender system projects have focused on optimizing accuracy using metrics such as precision/recall or mean absolute error. Now research has reached the point where going beyond pure accuracy and toward real user experience becomes indispensable for further advances [10].

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.

1.1 Contributions

We address the afore-mentioned deficiencies by focusing on techniques that are centered on real user satisfaction rather than pure accuracy. The contributions we make in this paper are the following:

- **Topic diversification.** We propose an approach towards balancing top-N recommendation lists according to the active user's full range of interests. Our novel method takes into consideration both the accuracy of suggestions made, and the user's extent of interest in specific topics. Analyses of topic diversification's implications on user-based [11, 22] and item-based [26, 5] collaborative filtering are provided.

*Researched while at GroupLens Research in Minneapolis.

WWW 2005, May 10-14, 2005, Chiba, Japan.
ACM 1-59593-046-9/05/0005.

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

- **Intra-list similarity metric.** Regarding diversity as an important ingredient to user satisfaction, metrics able to measure that characteristic feature are required. We propose the intra-list similarity metric as an efficient means for measurement, complementing existing accuracy metrics in their efforts to capture user satisfaction.

- **Accuracy versus satisfaction.** There have been several efforts in the past arguing that "accuracy does not tell the whole story" [4, 12]. Nevertheless, no evidence has been given to show that some aspects of actual user satisfaction reach beyond accuracy. We close this gap and provide analysis from large-scale online and offline evaluations, matching results obtained from accuracy metrics against actual user satisfaction and investigating interactions and deviations between both concepts.

1.2 Organization

Our paper is organized as follows. We discuss collaborative filtering and its two most prominent implementations in Section 2. The subsequent section then briefly reports on common evaluation metrics and the new intra-list similarity metric. In Section 4, we present our method for diversifying lists, describing its primary motivation and algorithmic clockwork. Section 5 reports on our offline and online experiments with topic diversification and provides ample discussion of results obtained.

2. ON COLLABORATIVE FILTERING

Collaborative filtering (CF) still represents the most commonly adopted technique in crafting academic *and* commercial [16] recommender systems. Its basic idea refers to making 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 or mentioning of an item counts as an expression of appreciation. While implicit ratings are generally more facile to collect, their usage implies adding noise to the collected information [20].

2.1 User-based Collaborative Filtering

User-based CF has been explored in-depth during the last ten years [29, 24, 14] and represents the most popular recommendation algorithm [11], 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 $\{b_k \in B \,|\, r_i(b_k) \neq \perp\}$.

The user-based CF's working process can be broken down into two major steps:

- **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 [29, 8] or cosine distance [11] 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 voters a_j with $r_j(b_k) \neq \perp$, as well as the ratings $r_j(b_k)$ these neighbors 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 upon predictions w_i. Note that function P_{w_i} is injective and reflects recommendation ranking in *descending* order, giving highest predictions first.

2.2 Item-based Collaborative Filtering

Item-based CF [13, 26, 5] has been gaining momentum over the last five years by virtue of favorable computational complexity characteristics and the ability 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 [26]. Its success also extends to many commercial recommender systems, such as Amazon.com's [16].

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.

Predictions $w_i(b_k)$ are computed as follows:

$$w_i(b_k) = \frac{\sum_{b_e \in B'_k} (c(b_k, b_e) \cdot r_i(b_e))}{\sum_{b_e \in B'_k} |c(b_k, b_e)|}, \qquad (1)$$

where

$$B'_k := \{b_e \,|\, b_e \in \text{clique}(b_k) \wedge r_i(b_k) \neq \perp\}$$

Intuitively, the approach tries to mimic real user behavior, having user a_i judge the value of an unknown product b_k by comparing the latter to known, similar items b_e and considering how much a_i appreciated these b_e.

The eventual computation of a top-N recommendation list P_{w_i} follows the user-based CF's process, arranging recommendations according to w_i in descending order.

3. EVALUATION METRICS

Evaluation metrics are essential in order to judge the quality and performance of recommender systems, even though they are still in their infancies. Most evaluations concentrate on accuracy measurements only and neglect other factors, e.g., novelty and serendipity of recommendations, and the diversity of the recommended list's items.

The following sections give an outline of popular metrics. An extensive survey of accuracy metrics is provided in [12].

3.1 Accuracy Metrics

Accuracy metrics have been defined first and foremost 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 [12].

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.

3.1.1 Predictive Accuracy Metrics

Predictive accuracy metrics measure how close predicted ratings come to true user ratings. Most prominent and widely used [29, 11, 3, 9], mean absolute error (MAE) represents an efficient means to measure the statistical accuracy of predictions $w_i(b_k)$ for sets B_i of products:

$$|\overline{E}| = \frac{\sum_{b_k \in B_i} |r_i(b_k) - w_i(b_k)|}{|B_i|} \tag{2}$$

Related to MAE, mean squared error (MSE) squares the error before summing. Hence, large errors become much more pronounced than small ones.

Very easy to implement, predictive accuracy metrics are inapt for evaluating the quality of top-N recommendation lists. Users only care about errors for high-rank products. On the other hand, prediction errors for low-rank products are unimportant, knowing that the user has no interest in them anyway. However, MAE and MSE account for both types of errors in exactly the same fashion.

3.1.2 Decision-Support Metrics

Precision and recall, both well-known from information retrieval, do not consider predictions and their deviations from actual ratings. They rather judge how *relevant* a set of ranked recommendations is for the active user.

Before using these metrics for cross-validation, K-folding is applied, dividing every user a_i's rated products $b_k \in R_i$ into K disjoint slices of preferably equal size. Hereby, $K-1$ randomly chosen slices form a_i's *training set* R_i^x. These ratings then define a_i's profile from which final recommendations are computed. For recommendation generation, a_i's residual slice $(R_i \setminus R_i^x)$ is retained and not used for prediction. This slice, denoted T_i^x, constitutes the *test set*, i.e., those products the recommenders intend to predict.

Sarwar [25] presents an adapted variant of recall, recording the percentage of test set products $b \in T_i^x$ occurring in recommendation list P_i^x with respect to the overall number of test set products $|T_i^x|$:

$$\text{Recall} = 100 \cdot \frac{|T_i^x \cap \Im P_i^x|}{|T_i^x|} \tag{3}$$

Symbol $\Im P_i^x$ denotes the *image* of map P_i^x, i.e., all items part of the recommendation list.

Accordingly, precision represents the percentage of test set products $b \in T_i^x$ occurring in P_i^x with respect to the size of the recommendation list:

$$\text{Precision} = 100 \cdot \frac{|T_i^x \cap \Im P_i^x|}{|\Im P_i^x|} \tag{4}$$

Breese *et al.* [3] introduce an interesting extension to recall, known as weighted recall or Breese score. The approach takes into account the *order* of the top-N list, penalizing incorrect recommendations less severely the further down the list they occur. Penalty decreases with exponential decay.

Other popular decision-support metrics include ROC [28, 18, 9], the "receiver operating characteristic". ROC measures the extent to which an information filtering system is able to successfully distinguish between signal and noise. Less frequently used, NDPM [2] compares two different, weakly ordered rankings.

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.

3.2.1 Coverage

Among all non-accuracy evaluation metrics, coverage has been the most frequently used [11, 19, 9]. Coverage measures the percentage of elements part of the problem domain for which predictions can be made.

3.2.2 Novelty and Serendipity

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" [12]. For some simple measure of serendipity, take the average popularity of recommended items. Lower scores obtained denote higher serendipity.

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, and other discerning characteristics. 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:

$$\text{ILS}(P_{w_i}) = \frac{\sum_{b_k \in \Im P_{w_i}} \sum_{b_e \in \Im P_{w_i},\, b_k \neq b_e} c_\circ(b_k, b_e)}{2} \tag{5}$$

Higher scores denote lower diversity. An interesting mathematical feature of $\text{ILS}(P_{w_i})$ 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:

$$\forall \sigma_i, \sigma_j \in S_N : \text{ILS}(P_{w_i} \circ \sigma_i) = \text{ILS}(P_{w_i} \circ \sigma_j) \tag{6}$$

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 user satisfaction, hiding several blatant flaws in existing systems [17]. 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. Reports from the item-based TV recommender TiVo [1], as well as personal experiences with Amazon.com's recommender, also item-based [16], back our conjecture. For instance, one of this paper's authors only gets recommendations for Heinlein's books, another complained about all his suggested books being Tolkien's writings.

Reasons for negative ramifications on user satisfaction implied by portfolio effects are well-understood and have been studied extensively in economics, termed "law of diminishing marginal returns" [30]. The law describes effects of saturation that steadily decrease the incremental utility of products p when acquired or consumed over and over again. For example, suppose you are offered your favorite drink. Let p_1 denote the price you are willing to pay for that product. Assuming your are offered a second glass of that particular drink, the amount p_2 of money you are inclined to spend will be lower, i.e., $p_1 > p_2$. Same for p_3, p_4, and so forth.

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 [33], 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. See Figure 1 for one sample taxonomy. 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 products 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 set $\{P_{w_i*}(k) \mid k \in [1, z[\}$, which contains all new recommendations preceding rank z.

Sorting all products b according to $c^*(b)$ in reverse order, we obtain the *dissimilarity rank* $P^{\text{rev}}_{c^*}$. 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 $P^{\text{rev}}_{c^*}$ 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 diversification 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.

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**
      $P^{\text{rev}^{-1}}_{c^*}(b) \leftarrow |B'_i| - P^{-1}_{c^*}(b)$;
      $w^*_i(b) \leftarrow P^{-1}_{w_i}(b) \cdot (1 - \Theta_F) + P^{\text{rev}^{-1}}_{c^*}(b) \cdot \Theta_F$;
    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

4.3 Recommendation Dependency

In order to implement topic diversification, we assume that recommended products $P_{w_i}(o)$ and $P_{w_i}(p)$, $o, p \in \mathbb{N}$, along with their content descriptions, effectively *do* exert an impact on each other, which is commonly ignored by existing approaches: usually, only relevance weight ordering $o < p \Rightarrow w_i(P_{w_i}(o)) \geq w_i(P_{w_i}(p))$ must hold for recommendation list items, no other dependencies are assumed.

In case of topic diversification, recommendation interdependence means that an item b's current dissimilarity rank with respect to preceding recommendations plays an important role and may influence the new ranking.

4.4 Osmotic Pressure Analogy

The effect of dissimilarity bears traits similar to that of osmotic pressure and selective permeability known from molecular biology [31]. Steady insertion of products b_o, taken from one specific area of interest d_o, into the recommendation list equates to the passing of molecules from one specific substance through the cell membrane into cytoplasm. With increasing concentration of d_o, owing to the membrane's selective permeability, the pressure for molecules b from other substances d rises. When pressure gets sufficiently high for one given topic d_p, its best products b_p may "diffuse" into the recommendation list, even though their original rank $P^{-1}_{w_i}(b)$ might be inferior to candidates from the prevailing domain d_o. Consequently, pressure for d_p decreases, paving the way for another domain for which pressure peaks.

Topic diversification hence resembles the membrane's selective permeability, which allows cells to maintain their internal composition of substances at required levels.

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

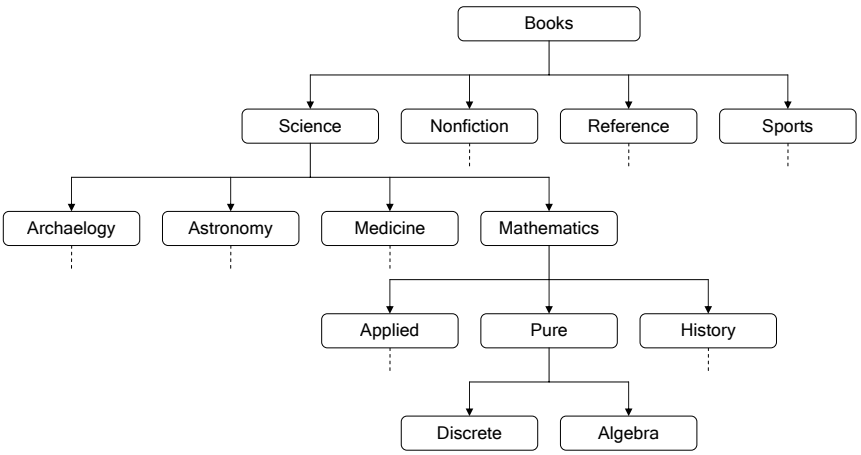


Figure 1: Fragment from the Amazon.com book taxonomy

when steadily increasing Θ_F and analyzing how both approaches respond to diversification.

5.1 Dataset Design

We based online and offline analyses on data we 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.

5.1.1 Data Collection

In a 4-week crawl, 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.

5.1.2 Condensation Steps

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.

5.2 Offline Experiments

We performed offline experiments comparing precision, recall, and intra-list similarity scores for 20 different recommendation 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.

5.2.1 Evaluation Framework Setup

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. The high-performance SUGGEST recommender engine[2] was used to compute these base case lists. 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.

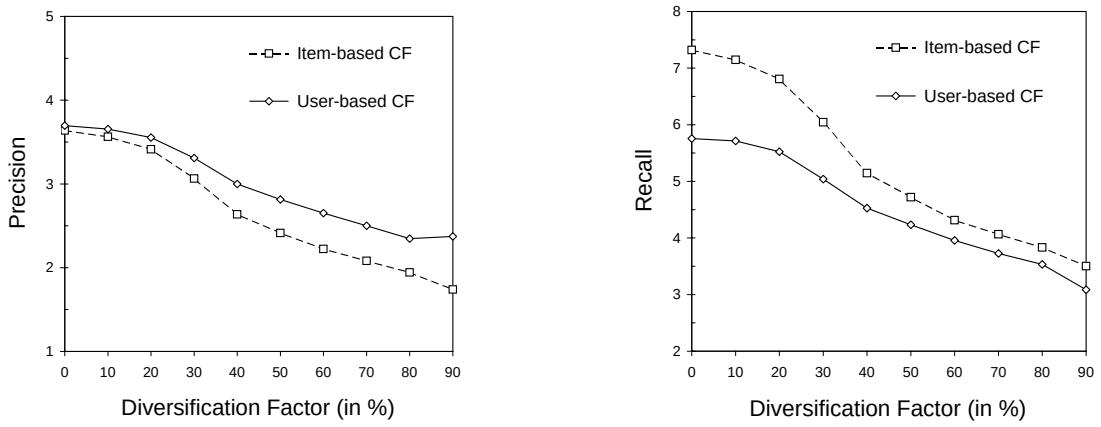
5.2.2 Result Analysis

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.

5.2.2.1 Precision and Recall.

First, we analyzed precision and recall scores for both non-diversified base cases, i.e., when $\Theta_F = 0$. Table 1 states that user-based and item-based CF exhibit almost identical accuracy, indicated by precision values. Their recall values differ

[2]Visit *http://www-users.cs.umn.edu/∼karypis/suggest/*.



(a) (b)

Figure 2: Precision (a) and recall (b) for increasing Θ_F

	Item-based CF	User-based CF
Precision	3.64	3.69
Recall	7.32	5.76

Table 1: Precision/recall for non-diversified CF

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 2. 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 [3, 12]. 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.

5.2.2.2 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 metric values for user-based and item-based CF with $\Theta_F \in [0, 0.9]$ each. Hereby, we instantiated the intra-list similarity metric function $c_\circ$ with our taxonomy-driven metric c^*. Results obtained from intra-list similarity analysis are given in Figure 3(a).

The topic diversification method considerably 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 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.

5.2.2.3 Original List Overlap.

Figure 3(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.

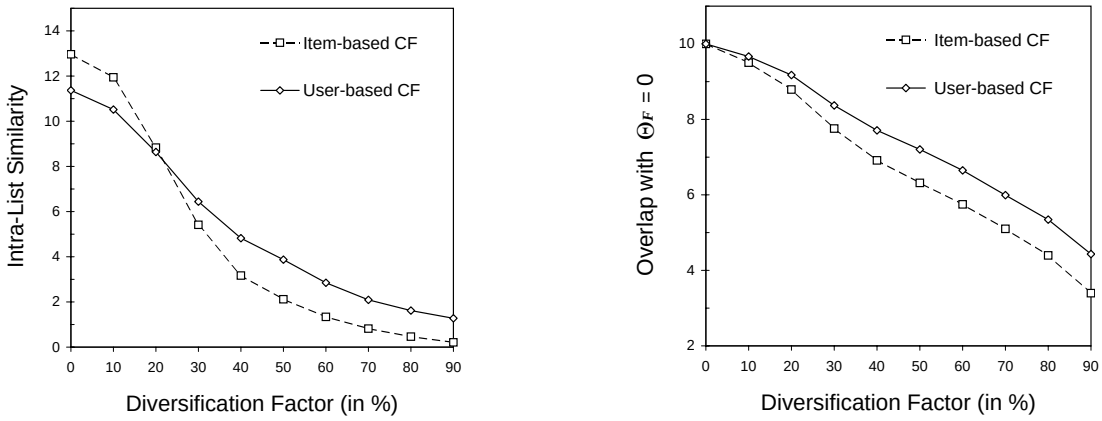
5.2.2.4 Conclusion.

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, knowing about the deficiencies of accuracy metrics, 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,



(a) (b)

Figure 3: Intra-list similarity behavior (a) and overlap with original list (b) for increasing Θ_F

though without K-folding. In cooperation with BookCrossing, we mailed all eligible users via the community mailing system, asking them to participate in our online study. Each mail contained a personal link that would direct the user 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.

5.3.1 Survey Outline and Setup

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. We embedded those questions we were actually keen about knowing into ones of lesser importance, in order to conceal our intentions and not bias users.

The one top-10 recommendation 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. We opted for those 12 instead of all 20 list types in order to acquire enough users completing the survey for each slot. 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.

5.3.2 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, owing to the very nature of these metrics. 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 reading interests. Referring to the intra-list similarity metric, we expected users' perceived range of topics, i.e., the list's diversity, to increase with increasing Θ_F. Third, we were curious about the overall satisfaction of users with their recommendation lists in their entirety, 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.

5.3.2.1 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, as depicted in Figure 4(a). At each featured diversification level Θ_F, differences between the two CF types are statistically 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, opposed to the trend, 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.

5.3.2.2 Covered Range.

Next, we analyzed whether users actually *perceived* the variety-augmenting effects caused by topic diversification, illustrated before through the measurement of intra-list sim-

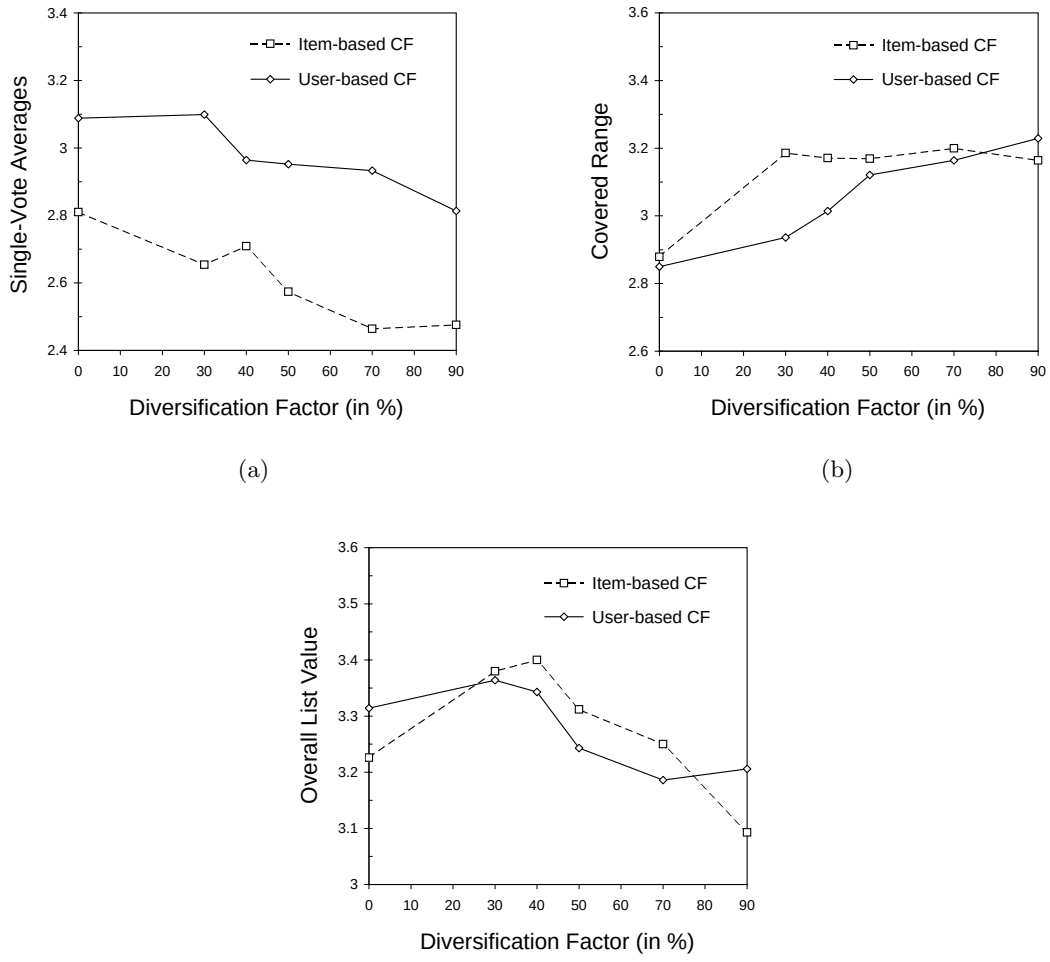


Figure 4: Results for single-vote averages (a), covered range of interests (b), and overall satisfaction (c)

ilarity. Users' reactions to steadily incrementing Θ_F are illustrated in Figure 4(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 3 indicates that diversity increases when increasing Θ_F, the item-based CF chart defies this trend, first soaring then flattening. We conjecture that the following three factors account for these peculiarities:

- **Diversification factor impact.** Our offline analysis of the intra-list similarity already suggested that the effect of topic diversification on item-based CF is much stronger than on user-based CF. Thus, the item-based CF's user-perceived interest coverage is significantly higher at $\Theta_F = 0.3$ than the user-based CF's.

- **Human perception.** We believe that human perception can capture the level of diversification inherent to a list only to some extent. Beyond that point, increasing diversity remains unnoticed. For the application scenario at hand, Figure 4 suggests this point around score value 3.2, reached by user-based CF only at $\Theta_F = 0.9$, and approximated by item-based CF already at $\Theta_F = 0.3$.

- **Interaction with accuracy.** Analyzing results obtained, bear in mind that covered range scores are *not*

fully independent from single-vote averages. When accuracy is poor, i.e., the user feels unable to identify recommendations that are interesting to him, chances are high his discontentment will also negatively affect his diversity rating. For $\Theta_F \in [0.5, 0.9]$, single-vote averages are remarkably low, which might explain why perceived coverage scores do not improve for increasing Θ_F.

However, we may conclude that users do perceive the application of topic diversification as an overly positive effect on reading interest coverage.

5.3.2.3 *Overall List Value.*

The third feature variable we were evaluating, the overall value users assigned to their personal recommendation list, effectively represents the *target value* of our studies, measuring actual user satisfaction. Owing to our conjecture that user satisfaction is a mere composite of accuracy and other influential 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 4(c). Analyzing 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 $\Theta_F \in [0.4, 0.7]$, eventually culminating in a slight but visible upturn at $\Theta_F = 0.9$. While lacking reasonable explanations and being opposed to our hypothesis, 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 obtained look different. In compliance with our previous hypothesis, the curve's shape roughly follows an arc, peaking at $\Theta_F = 0.4$. Taking the three data-points defining the arc, we obtain statistical significance for $p < 0.1$. Since the endpoint's score at $\Theta_F = 0.9$ is inferior to the non-diversified case's, we observe that too much diversification appears detrimental, perhaps owing to substantial interactions with accuracy.

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%.

5.4 Multiple Linear Regression

Results obtained from analyzing user feedback along various feature axes already indicated that users' overall satisfaction with recommendation lists not only depends on accuracy, but also on the range of reading interests covered. In order to more rigidly assess that indication by means of statistical methods, we applied multiple linear regression to our survey results, choosing the overall list value as dependent variable. As independent input variables, we provided single-vote averages and covered range, both appearing as first-order and second-order polynomials, i.e., SVA and CR, and SVA^2 and CR^2, respectively. We also tried several other, more complex models, without achieving significantly better model fitting.

	Estimate	**Error**	***t*-Value**	$Pr(>\|t\|)$
(const)	3.27	0.023	139.56	$< 2e-16$
SVA	12.42	0.973	12.78	$< 2e-16$
SVA^2	-6.11	0.976	-6.26	$4.76e-10$
CR	19.19	0.982	19.54	$< 2e-16$
CR^2	-3.27	0.966	-3.39	0.000727

Multiple R^2: 0.305, adjusted R^2: 0.303

Table 2: Multiple linear regression results

Analyzing multiple linear regression results, shown in Table 2, confidence values $Pr(>|t|)$ clearly indicate that statistically significant correlations for accuracy and covered range with user satisfaction exist. Since statistical significance also holds for their respective second-order polynomials, i.e., CR^2 and SVA^2, we conclude that these relationships are non-linear and more complex, though.

As a matter of fact, linear regression delivers a strong indication that the intrinsic utility of a list of recommended items is more than just the average value of accuracy votes for all single items, but also depends on the perceived diversity.

5.5 Limitations

There are some limitations to the study, notably referring to the way topic diversification was implemented. Though the Amazon.com taxonomies were human-created, there may still be some mismatch between what the topic diversification algorithm perceives as "diversified" and what humans do. The issue is effectively inherent to the taxonomy's structure, which has been designed with *browsing tasks* and ease of searching rather than with interest profile generation in mind. For instance, the taxonomy features topic nodes labelled with letters for alphabetical ordering of authors from the same genre, e.g., Books → Fiction → ... → Authors, A-Z → G. Hence, two Sci-Fi books from two different authors with the same initial of their last name would be classified under the same node, while another Sci-Fi book from an author with a *different* last-name initial would *not*. Though the problem's impact is largely marginal, owing to the relatively deep level of nesting where such branchings occur, the procedure appears far from intuitive.

An alternative approach to further investigate the accuracy of taxonomy-driven similarity measurement, and its limitations, would be to have *humans* do the clustering, e.g., by doing card sorts or by estimating the similarity of any two books contained in the book database. The results could then be matched against the topic diversification method's output.

6. RELATED WORK

Few efforts have addressed the problem of making top-N lists more diverse. Only considering literature on collaborative filtering and recommender systems in general, none have been presented before, to the best of our 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" [6]. Most approaches use variations of the "linear combination of score" model (LC), described by Vogt and Cottrell [32]. The LC model effectively resembles our scheme for merging the original, accuracy-based ranking with the current dissimilarity ranking, but is more general and does not address the diversity issue. Fagin *et al.* [7] propose metrics for measuring the distance between top-N lists, i.e., inter-list similarity metrics, in order to evaluate the quality of merged ranks. Oztekin *et al.* [21] extend the linear combination approach by proposing rank combination models that also incorporate content-based features in order to identify the most relevant topics.

More related to our idea of creating lists that represent the whole plethora of the user's topic interests, Kummamuru *et al.* [15] 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. Google (*http://www.google.com*) uses several mechanisms to suppress top-N items too similar in content, showing them only upon the user's explicit request. Unfortunately, no publications on that matter are available.

7. CONCLUSION

We presented topic diversification, an algorithmic framework to increase the diversity of a top-N list of recommended products. In order to show its efficiency in diversifying, we also introduced our new intra-list similarity metric.

Contrasting precision and recall metrics, computed both for user-based and item-based CF and featuring different levels of diversification, 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 were thus able to provide empirical evidence 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 3. 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 [16] and TiVo [1], are item-based, owing to the algorithm's computational efficiency.

8. FUTURE WORK

Possible future directions branching out from our current state of research on topic diversification are rife.

First, we would like to study the impact of topic diversification when dealing with application domains other than books, e.g., movies, CDs, and so forth. Results obtained may differ, owing to distinct characteristics concerning the structure of genre classification inherent to these domains. For instance, Amazon.com's classification taxonomy for books is more deeply nested, though smaller, than its movie counterpart [34]. Bear in mind that the structure of these taxonomies severely affects the taxonomy-based similarity measure c^*, which lies at the very heart of the topic diversification method.

Another interesting path to follow would be to parameterize the diversification framework with several different similarity metrics, either content-based or CF-based, hence superseding the taxonomy-based c^*.

We strongly believe that our topic diversification approach bears particularly high relevance for recommender systems involving *sequential consumption* of list items. For instance, think of personalized Internet radio stations, e.g., Yahoo's Launch (*http://launch.yahoo.com*): community members are provided with playlists, computed according to their own taste, which are sequentially processed and consumed. Controlling the right mix of items within these lists becomes vital and even more important than for mere "random-access" recommendation lists, e.g., book or movie lists. Suppose such an Internet radio station playing five Sisters of Mercy songs in a row. Though the active user may actually like the respective band, he may not want all five songs played in sequence. Lack of diversion might thus result in the user leaving the system.

The problem of finding the right mix for sequential consumption-based recommenders takes us to another future direction worth exploring, namely individually adjusting the right level of diversification versus accuracy tradeoff. One approach could be to have the user himself define the degree of diversification he likes. Another approach might involve learning the right parameter from the user's *behavior*, e.g., by observing which recommended items he inspects and devotes more time to, etc.

Finally, we are also thinking about diversity metrics other than intra-list similarity. For instance, we envision a metric that measures the extent to which the top-N list actually reflects the user's profile.

9. ACKNOWLEDGEMENTS

The authors would like to express their gratitude towards Ron Hornbaker, CTO of Humankind Systems and chief architect of BookCrossing, for his invaluable support. Furthermore, we would like to thank all BookCrossing members participating in our online survey for devoting their time and giving us many invaluable comments.

Moreover, we would like to thank John Riedl, Dan Cosley, Yongli Zhang, Paolo Massa, Zvi Topol, and Lars Schmidt-Thieme for fruitful comments and discussions.

10. 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] BALABANOVIĆ, M., AND SHOHAM, Y. Fab - content-based, collaborative recommendation. *Communications of the ACM 40*, 3 (March 1997), 66–72.

[3] 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.

[4] Cosley, D., Lawrence, S., and Pennock, D. REFEREE: An open framework for practical testing of recommender systems using ResearchIndex. In *28th International Conference on Very Large Databases* (Hong Kong, China, August 2002), Morgan Kaufmann, pp. 35–46.

[5] Deshpande, M., and Karypis, G. Item-based top-n recommendation algorithms. *ACM Transactions on Information Systems 22*, 1 (2004), 143–177.

[6] 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.

[7] Fagin, R., Kumar, R., and Sivakumar, D. Comparing top-k lists. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Baltimore, MD, USA, 2003), SIAM, pp. 28–36.

[8] 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.

[9] 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.

[10] Hayes, C., Massa, P., Avesani, P., and Cunningham, P. An online evaluation framework for recommender systems. In *Workshop on Personalization and Recommendation in E-Commerce* (Malaga, Spain, May 2002), Springer-Verlag.

[11] 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.

[12] Herlocker, J., Konstan, J., Terveen, L., and Riedl, J. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems 22*, 1 (2004), 5–53.

[13] Karypis, G. Evaluation of item-based top-N recommendation algorithms. In *Proceedings of the Tenth ACM CIKM International Conference on Information and Knowledge Management* (Atlanta, GA, USA, 2001), ACM Press, pp. 247–254.

[14] Konstan, J., Miller, B., Maltz, D., Herlocker, J., Gordon, L., and Riedl, J. GroupLens: Applying collaborative filtering to usenet news. *Communications of the ACM 40*, 3 (1997), 77–87.

[15] 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 Thirteenth International Conference on World Wide Web* (New York, NY, USA, 2004), ACM Press, pp. 658–665.

[16] Linden, G., Smith, B., and York, J. Amazon.com recommendations: Item-to-item collaborative filtering. *IEEE Internet Computing 4*, 1 (January 2003).

[17] McLaughlin, M., and Herlocker, J. A collaborative filtering algorithm and evaluation metric that accurately model the user experience. In *Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Sheffield, UK, 2004), ACM Press, pp. 329–336.

[18] Melville, P., Mooney, R., and Nagarajan, R. Content-boosted collaborative filtering for improved recommendations. In *Eighteenth National Conference on Artificial Intelligence* (Edmonton, Canada, 2002), American Association for Artificial Intelligence, pp. 187–192.

[19] Middleton, S., Shadbolt, N., and De Roure, D. Ontological user profiling in recommender systems. *ACM Transactions on Information Systems 22*, 1 (2004), 54–88.

[20] Nichols, D. Implicit rating and filtering. In *Proceedings of the Fifth DELOS Workshop on Filtering and Collaborative Filtering* (Budapest, Hungary, 1998), ERCIM, pp. 31–36.

[21] Oztekin, U., Karypis, G., and Kumar, V. Expert agreement and content-based reranking in a meta search environment using Mearf. In *Proceedings of the Eleventh International Conference on World Wide Web* (Honolulu, HW, USA, 2002), ACM Press, pp. 333–344.

[22] Resnick, P., Iacovou, N., Suchak, M., Bergstorm, P., and Riedl, J. GroupLens: An open architecture for collaborative filtering of netnews. In *Proceedings of the ACM 1994 Conference on Computer Supported Cooperative Work* (Chapel Hill, NC, USA, 1994), ACM, pp. 175–186.

[23] Resnick, P., and Varian, H. Recommender systems. *Communications of the ACM 40*, 3 (1997), 56–58.

[24] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Analysis of recommendation algorithms for e-commerce. In *Proceedings of the 2nd ACM Conference on Electronic Commerce* (Minneapolis, MN, USA, 2000), ACM Press, pp. 158–167.

[25] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. Application of dimensionality reduction in recommender systems. In *ACM WebKDD Workshop* (Boston, MA, USA, August 2000).

[26] 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).

[27] Schafer, B., Konstan, J., and Riedl, J. Meta-recommendation systems: User-controlled integration of diverse recommendations. In *Proceedings of the 2002 International ACM CIKM Conference on Information and Knowledge Management* (2002), ACM Press, pp. 43–51.

[28] Schein, A., Popescul, A., Ungar, L., and Pennock, D. Methods and metrics for cold-start recommendations. In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Tampere, Finland, 2002), ACM Press, pp. 253–260.

[29] 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.

[30] Spillman, W., and Lang, E. *The Law of Diminishing Returns*. World Book Company, Yonkers-on-Hudson, NY, USA, 1924.

[31] Tombs, M. *Osmotic Pressure of Biological Macromolecules*. Oxford University Press, New York, NY, USA, 1997.

[32] Vogt, C., and Cottrell, G. Fusion via a linear combination of scores. *Information Retrieval 1*, 3 (1999), 151–173.

[33] 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.

[34] Ziegler, C.-N., Schmidt-Thieme, L., and Lausen, G. Exploiting semantic product descriptions for recommender systems. In *Proceedings of the 2nd ACM SIGIR Semantic Web and Information Retrieval Workshop 2004* (Sheffield, UK, July 2004).