A Comparison of Calibrated and Intent-Aware Recommendations

Mesut Kaya
mesut.kaya@insight-centre.org
Insight Centre for Data Analytics
University College Cork
Cork, Ireand

Derek Bridge
derek.bridge@insight-centre.org
Insight Centre for Data Analytics
University College Cork
Cork, Ireland

ABSTRACT

Calibrated and intent-aware recommendation are recent approaches to recommendation that have apparent similarities. Both try, to a certain extent, to cover the user's interests, as revealed by her user profile. In this paper, we compare them in detail. On two datasets, we show the extent to which intent-aware recommendations are calibrated and the extent to which calibrated recommendations are diverse. We consider two ways of defining a user's interests, one based on item features, the other based on subprofiles of the user's profile. We find that defining interests in terms of subprofiles results in highest precision and the best relevance/diversity trade-off. Along the way, we define a new version of calibrated recommendation and three new evaluation metrics.

CCS CONCEPTS

• **Information systems → Recommender systems**.

KEYWORDS

Calibration; Intent-aware; Diversity

ACM Reference Format:
Mesut Kaya and Derek Bridge. 2019. A Comparison of Calibrated and Intent-Aware Recommendations. In *Thirteenth ACM Conference on Recommender Systems (RecSys '19), September 16–20, 2019, Copenhagen, Denmark.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3298689.3347045

1 INTRODUCTION

It is well-known that recommender systems that focus only on rating prediction accuracy or even recommendation relevance may bring problems: for example, they can narrow a user's horizons, even creating so-called "filter bubbles" [14], by recommending items that are either too obvious, popular or similar to ones in the user's profile; they can lead to monotony in a user's interactions with the system [4]; and they may under-represent a user's more peripheral interests by concentrating on her dominant interests [19]. Hence, prediction accuracy or recommendation relevance should not be the only measures of recommender system quality [5]. For example, in some domains it may be desirable that recommendations also be novel or serendipitous and that a set of recommendations be diverse [6, 12].

In 2018, Steck proposed *calibrated recommendations* [19]. A set of recommendations is calibrated if it reflects the different interests of the user, as revealed by the user profile, and with the appropriate proportions. In [19], interests are defined in terms of item features (such as movie genres). If 75% of the movies in a user's profile are horror movies and 25% are sci-fi movies, then a set of recommendations is calibrated if it exhibits these genres in roughly these proportions.[1]

Dating from 2011 to the current day, *intent-aware recommender systems* diversify a set of recommendations, inspired by work in Information Retrieval (IR) [1, 16, 21]. Faced with an ambiguous query such as "jaguar", an IR system cannot know whether the user intended to search for cats, cars or operating systems. Informally, intent-aware methods in IR ensure that the search results include at least one document for each possible intention. Analogously, intent-aware recommender systems ensure that the set of recommendations contains items that cover each of the user's interests, as revealed by her profile [8, 22, 24]. The user's interests are defined as a probability distribution over a set of aspects $\mathcal{A}$. In the xQuAD and c-plsa systems, for example, aspects are item features [22, 24]. In the SPAD system, aspects are subsets of the items in the user's profile, referred to as subprofiles [8].

Presented in this way, calibrated and intent-aware recommender systems seem very similar: both try to cover the interests revealed by the user's profile. The main difference lies in their objective. Aside from recommendation relevance, the main goal of a calibrated recommender system is to produce recommendations that reflect the user's interests in the right proportions. A calibrated recommendation set might be diverse, but diversity is not an explicit goal.[2] By contrast, diversity is the main goal of intent-aware systems. This is achieved by something similar to calibration and so a set of recommendations might be calibrated to some extent. But, as we will see (Section 2), intent-aware methods define their equivalent to calibration in a relevance-based way: the proportion of items in a recommendation set that cover an aspect are modulated by how relevant those items are to that aspect. Hence, in the example above, more or fewer than 75% of the recommendations may be horror movies.

RecSys '19, September 16–20, 2019, Copenhagen, Denmark

ACM ISBN 978-1-4503-6243-6/19/09…$15.00
https://doi.org/10.1145/3298689.3347045

[1] This example simplifies by ignoring the possibility that a movie might have more than one genre.

[2] In fact, Steck defines a diversity-enhanced calibrated recommender system, which includes diversity as an explicit objective, alongside calibration [19]. We will discuss it in Section 6.5.

This paper compares these two to determine how well each achieves three objectives: calibration, relevance and diversity. More specifically, the paper's contributions are:

- We define a new variant of Steck's calibrated recommender systems, one which calibrates with respect to subprofiles, rather than item features.
- We define three new evaluation metrics, corresponding to existing metrics that are defined using item features. The new metrics use subprofiles, rather than item features. Using the new metrics alongside the existing ones gives a more balanced view of the performance of the recommender algorithms.
- We present an empirical comparison using all these metrics on two datasets to see the extent to which calibrated and intent-aware recommenders do produce calibrated recommendations, relevant recommendations and diverse sets of recommendations.

The comparison shows that, in these datasets, subprofiles are better than genres; it shows that in some cases intent-aware recommendations are calibrated and that calibrated recommendations are diverse; but it shows that the best relevance/diversity trade-offs are achieved by intent-aware methods that use subprofiles.

The rest of this paper is organized as follows. In Sections 2 and 3, we summarize intent-aware and calibrated recommendation, respectively. Section 4 briefly contrasts intent-aware and calibrated recommendations. In Section 5, we give the details of the datasets, metrics and methodology used in our experiments. Section 6 analyzes the experimental results.

2 INTENT-AWARE RECOMMENDATION

Intent-aware recommender systems aim to determine a recommendation set of size N items, denoted here by RL^*, that optimizes an objective function that balances relevance with diversity. Most commonly, the objective function is a linear combination of the relevance of the items in the recommendation set and the diversity of that set, the trade-off between the two being controlled by a parameter λ ($0 \leq \lambda \leq 1$):

$$RL^* = \underset{RL, |RL|=N}{\arg\max}\ (1-\lambda)s(RL) + \lambda\,\mathrm{div}_{\mathrm{IA}}(RL) \quad (1)$$

Typically, s is a modular function that is the sum of the predicted relevance scores $s(u,i)$ of each item i recommended to user u, i.e. $s(RL) = \sum_{i \in RL} s(u,i)$. $\mathrm{div}_{\mathrm{IA}}(RL)$ measures the diversity of the set RL. But this is not simply a measure of how different the items are from each other, as it would be in more conventional approaches to diversity [2, 18, 25]. Instead, it is defined in terms of coverage of the user's interests, but with coverage modulated by recommendation relevance (below).

In Vargas's adaptation to recommender systems [23] of Santos et al.'s Query Aspect Diversification framework (xQuAD) [16], $\mathrm{div}_{\mathrm{IA}}(RL)$ is defined as follows:

$$\mathrm{div}_{\mathrm{IA}}(RL) = \sum_{a \in \mathcal{A}} p(a|u)\left(1 - \prod_{i \in RL}(1 - p(i|u,a))\right) \quad (2)$$

in which a user u's interests are formulated as a probability distribution $p(a|u)$ over aspects $a \in \mathcal{A}$. $p(i|u,a)$ is the probability of choosing an item i from a set of candidate recommendations RS, produced by a conventional recommender algorithm, given an aspect a and user u.

Finding an optimal solution to Eq. 1 is intractable in general, since it requires considering all possible sets of recommendations RL. However, when the objective function is *monotone* and *submodular*, a $1 - \frac{1}{e}$ approximation to the optimal solution can be computed greedily [13], where e is Euler's number. It is easy to adapt the proof given in [1] to show that, where $\mathrm{div}_{\mathrm{IA}}$ is defined as in Eq. 2, it is indeed monotone and submodular [7]. It follows easily that Eq. 1 is also monotone and submodular.

In the greedy algorithm, a conventional recommender algorithm (which we refer to as the *baseline recommender*), produces a set of recommended items, RS, for user u. For each recommended item i in RS, it also produces a score, $s(u,i)$, that estimates the relevance of recommended item i to user u. Then, the greedy algorithm *re-ranks* RS by iteratively inserting into ordered result list RL the item i from RS that maximizes a function, $f_{obj}(i, RL)$:

$$f_{obj}(i, RL) = (1-\lambda)s(u,i) + \lambda\,\mathrm{div}_{\mathrm{IA}}(i, RL) \quad (3)$$

where $\mathrm{div}_{\mathrm{IA}}(i, RL)$ is the marginal gain in diversity after inserting item i into the set RL. $\mathrm{div}_{\mathrm{IA}}(i, RL)$ is defined as:

$$\mathrm{div}_{\mathrm{IA}}(i, RL) = \mathrm{div}_{\mathrm{IA}}(\{i\} \cup RL) - \mathrm{div}_{\mathrm{IA}}(RL) \quad (4)$$

Using Eq. 2, we can obtain the following [1, 16, 23]:

$$\mathrm{div}_{\mathrm{IA}}(i, RL) = \sum_{a \in A}\left[p(a|u)p(i|u,a)\prod_{j \in RL}(1 - p(j|u,a))\right] \quad (5)$$

We now show ways of instantiating $\mathrm{div}_{\mathrm{IA}}(RL)$ and $\mathrm{div}_{\mathrm{IA}}(i, RL)$, differing in how aspects are defined.

2.1 xQuAD

In xQuAD, the aspects are explicit item features $\mathcal{F}$, such as genres, i.e. $\mathcal{A} = \mathcal{F}$. Hence we will write $p(f|u)$ instead of $p(a|u)$ and $p(i|u,f)$ instead of $p(i|u,a)$. Let $\mathcal{F}_i$ be the subset of $\mathcal{F}$ that describes item i (e.g. the genres of movie i) and let I_u denote the items that are in the user's profile. Then in Eqs. 2 and 5, $p(f|u)$ can be estimated as:

$$p(f|u) = \frac{|\{i \in I_u : f \in \mathcal{F}_i\}|}{\sum_{f' \in \mathcal{F}} |\{i \in I_u : f' \in \mathcal{F}_i\}|} \quad (6)$$

$p(i|u,f)$ can be estimated as:

$$p(i|u,f) = \frac{\mathbb{1}(i,f)s(u,i)}{\sum_{j \in RS} \mathbb{1}(j,f)s(u,j)} \quad (7)$$

where $\mathbb{1}(i,f) = 1$ if $f \in \mathcal{F}_i$ and 0 otherwise.

2.2 SPAD

Recently, we have introduced Subprofile-Aware Diversification (SPAD) [7, 8, 10]. SPAD is a form of intent-aware recommendation, hence it uses the same objective function (Eq. 2) and the same function for greedy re-ranking (Eq. 3). But it differs in the way it models the user's interests. Instead of item features, it uses subprofiles $\mathcal{S}$ as aspects, i.e. $\mathcal{A} = \mathcal{S}$. Hence, we write $p(S|u)$ and $p(i|u,S)$ instead of $p(a|u)$ and $p(i|u,a)$ in the two equations.

A subprofile is a set of items that capture one of the user's interests. In SPAD, subprofiles are mined from the user's liked-item-set,

I_u^+. In the case of systems that use implicit feedback, the user's profile I_u contains items that the user has shown an interest in and therefore, $I_u^+ = I_u$. In the case of systems that use explicit feedback in the form of numeric ratings, for example, then the liked-item-set $I_u^+ \subseteq I_u$ will be just those items to which the user has given a high rating (e.g. 4s or 5s on a 1–5 rating scale).

We have published several ways to mine the subprofiles from the liked-item-set [7, 10]. In this paper, we will use the one that we used in [9] and which most often outpeforms the others. By this algorithm, there is a candidate subprofile for each item $i \in I_u^+$. Informally, the subprofile for item $i \in I_u^+$ should contain other members of I_u^+ that together represent one user interest. Specifically, the candidate subprofile for each $i \in I_u^+$ contains i itself and also $j \in I_u^+$ if j's nearest-neighbours contain i, i.e. the candidate subprofile for i is $\{j \in I_u^+ : i \in \text{KNN}(j), i \neq j\} \cup \{i\}$. KNN($j$) are the most similar items to j in I, for which we use cosine similarity on the rating vectors. We will denote the size of this set by k_{nn}. The candidate subprofiles are pruned to a final set of subprofiles by excluding those that are wholly contained in any of the others. We will denote user u's final set of subprofiles by $\mathcal{S}_u$.

Returning to Eq. 2, in SPAD $p(S|u)$ is estimated as:

$$p(S|u) = \frac{|S|}{\sum_{S' \in \mathcal{S}_u} |S'|} \tag{8}$$

and $p(i|u, S)$, the probability of choosing i from recommendations RS given subprofile S and user u, is estimated as:

$$p(i|u, S) = \frac{\mathbb{1}(i, S)s(u, i)}{\sum_{j \in RS} \mathbb{1}(j, S)s(u, j)} \tag{9}$$

where item i is *related to* subprofile S, as follows:

$$\mathbb{1}(i, S) = \begin{cases} 1 & \text{if } i \in \bigcup_{j \in S} \text{KNN}(j) \\ 0 & \text{otherwise} \end{cases} \tag{10}$$

where KNN(j) is again the set of j's nearest-neighbours in I. In other words, i must be a neighbour of a member of S. We will denote the size of this set by k_{ind}, since it need not be the same as k_{nn}.

3 CALIBRATED RECOMMENDATION

Calibrated recommendation aims to produce a recommendation set that covers the user's different interests in the proportions that they occur in the user's profile [19].

3.1 Calibrated recommendation using features

In [19], the degree of calibration, $\text{C}_{\text{KL}}(p, q)$, is quantified by taking the Kullback-Leibler divergence between two probability distributions: the first, $p(f|u)$, is the distribution of features f across the items in user u's liked-item-set; the second, $q(f|u)$, is the distribution of features f across the items in a recommendation set RL. These two distributions are defined in terms of another distribution, $p(f|i)$, which is the distribution of features for each item i. A simple definition of $p(f|i)$ is that, for each feature f possessed by an item i, $p(f|i)$ will be equal and such that $\sum_f p(f|i) = 1$ [19].

Specifically, Steck defines $p(f|u)$ as:

$$p(f|u) = \frac{\sum_{i \in I_u^+} w_{u,i} p(f|i)}{\sum_{i \in I_u^+} w_{u,i}} \tag{11}$$

In this definition, $w_{u,i}$ is a weight for item i. As an example, Steck suggests that the weight could be based on how recently item i was consumed by user u. However, in the rest of his paper, he takes $w_{u,i} = 1$ for all u and i, and we do the same in this paper.

Steck defines $q(f|u)$ as:

$$q(f|u) = \frac{\sum_{i \in RL} w_{r(i)} p(f|i)}{\sum_{i \in RL} w_{r(i)}} \tag{12}$$

$w_{r(i)}$ denotes the weight of item i due to its rank, $r(i)$ in RL, although again both in his paper, and in ours, $w_{r(i)} = 1$.

The degree of calibration is the Kullback-Leibler divergence of the two distributions, taking p as the target:

$$\text{C}_{\text{KL}}(p, q) = KL(p||\tilde{q}) = \sum_f p(f|u) \log \frac{p(f|u)}{\tilde{q}(f|u)} \tag{13}$$

Because Kullback-Leibler divergence diverges if $q(f|u) = 0$ and $p(f|u) > 0$, Steck uses $\tilde{q}$ instead of q:

$$\tilde{q} = (1 - \alpha)q(f|u) + \alpha p(f|u) \tag{14}$$

with a small positive value of α so that $q \approx \tilde{q}$. Steck uses $\alpha = 0.01$, and we do the same.

Now that we can measure the degree of calibration, we can define an objective function. It is a linear combination of the relevance of the items in the recommendation set and the degree of calibration of that set:

$$RL^* = \underset{RL, |RL|=N}{\arg\max} \; (1 - \lambda)s(RL) - \lambda \, \text{cal}(p, q) \tag{15}$$

Here, while p is different for each user, q is different for each recommendation set.

Steck proves that Eq. 15 satisfies the conditions for a greedy re-ranking approach to find a $1 - \frac{1}{e}$ approximation to the optimal solution. The greedy re-ranking approach uses the following objective function:

$$f_{obj}(i, RL) = (1 - \lambda)s(u, i) + \lambda \, \text{cal}(i, RL) \tag{16}$$

with $\text{cal}(i, RL) = -(\text{C}_{\text{KL}}(p, q(RL \cup \{i\})) - \text{C}_{\text{KL}}(p, q(RL)))$.

We will refer to this instantiation of calibrated recommendation as $\text{CR}_{\mathcal{F}}$.

3.2 Calibrated recommendation using subprofiles

Steck defines user interests in terms of item features. But the subprofile idea that we introduced for SPAD opens an opportunity to define a new variant of calibrated recommendation, one which uses subprofiles instead of features, much as SPAD uses subprofiles where xQuAD uses features. We refer to this variant as $\text{CR}_{\mathcal{S}}$.

In $\text{CR}_{\mathcal{S}}$, the distributions p and q (Eqs. 11 and 12) are defined in the same way, writing S in place of f.

$$p(S|u) = \frac{\sum_{i \in I_u^+} w_{u,i} p(S|i)}{\sum_{i \in I_u^+} w_{u,i}} \tag{17}$$

$$q(S|u) = \frac{\sum_{i \in RL} w_{r(i)} p(S|i)}{\sum_{i \in RL} w_{r(i)}} \tag{18}$$

To define $p(S|i)$, which replaces $p(f|i)$ in these equations, we use $p(S|i) = \frac{1}{|\mathcal{S}_i|}$, where $\mathcal{S}_i$ is the set of user u's final subprofiles that item i is *related to* (using Eq. 10).

4 INTENT-AWARE VS. CALIBRATED

As we have mentioned, there is an apparent similarity between intent-aware and calibrated recommendation. Both try to cover the user's different interests, as revealed by her profile. Indeed, for the latter, covering the tastes and interests in the same proportion as they occur in the user's profile is the main goal. Intent-aware recommendation may result in calibrated recommendations, but it does not directly aim to cover the user's interests in the same proportion as they occur in the user's profile. As we saw, it modulates coverage by recommendation relevance. We illustrate this difference with an informal example.[3]

Consider a user who listens to jazz 70% of the time and to rock music 30% of the time. Suppose that the goal is to recommend a list of top-10 recommendations. Calibrated recommendation tries to generate 10 recommendations such that seven (70%) are jazz and the remaining three (30%) are rock. An intent-aware approach, such as xQuAD, considers how much an item satisfies a given aspect (genre in this case) through $p(i|u, f)$. Suppose the first four songs that the recommender includes in the recommendation list are jazz songs which this user is likely to choose from a recommendation list, i.e. $p(i|u, f = \mathit{jazz})$ is high for each of these songs, i. Now suppose we are trying to decide whether to add a fifth piece of jazz to the recommendation list. It will be penalized because each of the four existing items j in the recommendation list have high $p(j|u, f = \mathit{jazz})$: see the factor $\prod_{j \in RL}(1 - p(j|u, f))$ in Eq. 5. A fifth jazz song might only be included if it can overcome the 'penalty' imposed by the songs that have been added to the recommendation list already. Because of this, the final top-10 might not have seven jazz songs; it may even have more rock than jazz. On the other hand, if the first seven songs to be added to the recommendation list are (informally speaking) not jazzy enough for this user's tastes (more precisely, if they have very low values for $p(i|u, f = \mathit{jazz})$), then it is possible that more jazz songs will be added to the recommendation list. The final top-10 might not have three rock songs; it may even have no rock at all.

5 EXPERIMENTS

In this section, we compare intent-aware and calibrated recommendations. We reveal the extent to which intent-aware approaches produce calibrated recommendations. We evaluate calibrated recommendation more thoroughly than was done in [19]. Steck's goal in [19] was just "to illustrate that the proposed approach [i.e. calibrated recommendation] works as expected". Hence, he used just one dataset. He compared calibrated recommendation with just its baseline. While he did measure recall (as defined in [17]), his focus was on measuring calibration. Here, we will use two datasets; we will compare two forms of calibrated recommendation ($\mathrm{CR}_{\mathcal{F}}$ and $\mathrm{CR}_{\mathcal{S}}$) to two forms of intent-aware recommendation (xQuAD and SPAD); and we will measure calibration but also precision and diversity. In the case of diversity, we will show results for four different metrics, and we will explore the trade-off the recommenders make between precision and diversity.

Table 1: Datasets

	ML20M	TasteProfile
# of users	137,765	375,749
# of items	20,631	190,629
# of interactions	9,990,460	19,328,665

5.1 Datasets

We follow [19] in using datasets that have implicit ratings, since, as Steck says, these are more common in real applications. We use the MovieLens 20 Million dataset (ML20M)[4] (which is the same dataset used in [19]) and the Taste Profile Subset dataset (TasteProfile)[5].

In the original ML20M dataset, all users have at least 20 movies in their profile and movies have one or more of 19 genres. We preprocess the ML20M in the same way as [19]. Specifically, we binarize the numeric ratings by dropping ratings lower than 4 stars and we eliminate movies that have no genre information. The resulting dataset has $\sim$ 10 million implicit ratings made by $\sim$ 140k users for $\sim$ 21k movies.

TasteProfile contains counts of the number of times a user has listened to a song. Each song appears in the Million Song Dataset, from which we can take information about up to 21 genres.[6] We eliminate songs that have no genre information and users who have fewer than 20 songs in their profile. We also binarize the song counts so that we get an implicit dataset; the dataset is binarized in the same way in [11]. The resulting dataset is comprised of $\sim$ 19 million implicit ratings made by $\sim$ 375k users for $\sim$ 190k songs.

The characteristics of the datasets after preprocessing are summarized in Table 1.

5.2 Evaluation metrics

For recommendation relevance, we measure Precision. For diversity, we use Intra-List Diversity (ILD) [25] and α-nDCG [3]. To measure the degree of calibration, we use $\mathrm{C_{KL}}$, the definition of which has been already presented as Eq. 13.

Intra-List Diversity, ILD, measures the average pairwise distance of the items in a recommendation set [25].

$$\mathrm{ILD}(L) = \frac{2}{|L|(|L|-1)} \sum_{i \in L} \sum_{j \in L, j \neq i} \mathrm{dist}(i, j) \tag{19}$$

Most commonly, $\mathrm{dist}(i, j)$ is the Jaccard distance between $\mathcal{F}_i$ and $\mathcal{F}_j$, the set of features of items i and j, respectively:

$$\mathrm{dist}(i, j) = \frac{|\mathcal{F}_i \cap \mathcal{F}_j|}{|\mathcal{F}_i \cup \mathcal{F}_j|} \tag{20}$$

[3] As before, the example simplifies by ignoring the possibility that a movie might have more than one genre.

[4] https://grouplens.org/datasets/movielens/20m/

[5] https://labrosa.ee.columbia.edu/millionsong/tasteprofile

[6] http://www.ifs.tuwien.ac.at/mir/msd/partitions/msd-MAGD-genreAssignment.cls

α-nDCG is based on nDCG but it is aspect and redundancy-aware, which makes it a measure of diversity:

$$\alpha\text{-nDCG}(L) = \frac{1}{\alpha\text{-IDCG}} \sum_{i \in L} \left[\frac{1}{\log_2(\mathrm{r}(i,L)+1)} \sum_{f \in \mathcal{F}} \left(\mathrm{rel}(i|u,f) \prod_{\substack{j \in L, \\ \mathrm{r}(j,L)<\mathrm{r}(i,L)}} (1 - \alpha\, \mathrm{rel}(j|u,f)) \right) \right] \tag{21}$$

where α-IDCG is the highest possible value of α-nDCG where the recommendation set is made of ideally diversified relevant items, L is the set of recommended items (of size N), $\mathrm{r}(i,L)$ is the position of i in L, and $\mathrm{rel}(i|u,f)$ is 1 if item i has feature f and is relevant to user u but 0 otherwise. α is the parameter that controls the penalty for redundancy. We use $\alpha = 0.5$, as in [23].

These two diversity metrics and the calibration metric are defined with respect to the item features $\mathcal{F}$. This gives $\mathrm{CR}_{\mathcal{F}}$ and xQuAD an advantage, since they make use of these features in their re-ranking. This is in line with the observation that algorithms can be expected to perform well regarding the metrics they were developed for [19, 20].

To obtain a more rounded picture of the relative performances of the algorithms, we propose to adapt C_{KL}, ILD and α-nDCG to produce versions of these metrics that use subprofiles instead of item features (below). Of course, these new metrics will favour $\mathrm{CR}_{\mathcal{S}}$ and SPAD, but taken together with the original metrics we will get a balanced view.

The new version of ILD is obtained simply by using a different definition of $\mathrm{dist}(i,j)$, one that uses subprofiles instead of item features:

$$\mathrm{dist}(i,j) = \frac{|\mathcal{S}_i \cap \mathcal{S}_j|}{|\mathcal{S}_i \cup \mathcal{S}_j|} \tag{22}$$

where $\mathcal{S}_i$, $\mathcal{S}_j$ are the set of a user's final subprofiles that item i and j are *related to* (Eq. 10). We will refer to ILD measured by item features as $\mathrm{ILD}^{\mathcal{F}}$ and ILD measured by subprofiles as $\mathrm{ILD}^{\mathcal{S}}$.

We modify α-nDCG by replacing $rel(i|u,f)$ with $rel(i|u,S)$ where S is a subprofile and such that $rel(i|u,S)$ is 1 if item i is *related to* subprofile S and user u but 0 otherwise. We will refer to α-nDCG measured by item features as $\alpha\text{-nDCG}^{\mathcal{F}}$ and α-nDCG measured by subprofiles as $\alpha\text{-nDCG}^{\mathcal{S}}$

Finally, we modify C_{KL} to use subprofiles instead of item features by replacing the distributions p and q over features in Eq. 13 with distributions over subprofiles, much as we did when we defined $\mathrm{CR}_{\mathcal{S}}$ in Section 3.2. We will refer to C_{KL} measured by item features as ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{F}}$ and C_{KL} measured by subprofiles as ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{S}}$.

5.3 Methodology

Our experiments use the methodology adopted in our previous work, e.g. [10]. After preprocessing the datasets, we randomly partition the ratings into training, validation and test sets such that 60% of each user's ratings are in the training set, 20% of them are in the validation set and 20% are in the test set.

We compare $\mathrm{CR}_{\mathcal{F}}$, $\mathrm{CR}_{\mathcal{S}}$, xQuAD and SPAD. All four recommenders use greedy re-ranking, therefore they need a baseline recommender, whose recommendation sets are re-ranked. We use a fast alternative least-squares matrix factorization recommender (MF) that works for implicit and explicit datasets [15]. It has two hyper-parameters: d, the number of latent factors; and α, the confidence level factor. The values that optimize the precision on the validation sets are selected. For ML20M, we find $d = 40$ and $\alpha = 6$. For TasteProfile, $d = 100$ and $\alpha = 20$.

$\mathrm{CR}_{\mathcal{S}}$ and SPAD have hyper-parameters k_{nn} and k_{ind} (see Section 2.2). To find their values, we follow the same methodology as in [10] to select values with best average $\alpha\text{-nDCG}^{\mathcal{F}}$ on the validation set. For both ML20M and TasteProfile, we find that $k_{nn} = 10$ and $k_{ind} = 10$ for $\mathrm{CR}_{\mathcal{S}}$, and it turns out that these are the values for SPAD also.

Once the hyper-parameter values have been found, for each user we generate a recommendation set RS, where $|RS| = 100$ by using MF on the union of the training and validation sets. Then, we re-rank each RS to produce ranked lists RL using each of the re-ranking algorithms. We do this for different values of λ from the set $\{0.1, 0.2, 0.3, \ldots, 1.0\}$. Then, from each RL we select the top-N recommendations, $N = 10$, and compute the metrics we defined in Section 5.2.

6 RESULTS

In this section, we report our empirical comparison of $\mathrm{CR}_{\mathcal{F}}$, $\mathrm{CR}_{\mathcal{S}}$, xQuAD and SPAD on the ML20M and TasteProfile datasets. First, we compare their performances on the two versions of the calibration metric, ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{F}}$ and ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{S}}$. Next, we see how they affect the relevance of the baseline recommendations by measuring Precision. Then, we look at their effect on 'pure' diversity metrics, $\mathrm{ILD}^{\mathcal{F}}$ and $\mathrm{ILD}^{\mathcal{S}}$, and relevance-aware diversity metrics, $\alpha\text{-nDCG}^{\mathcal{F}}$ and $\alpha\text{-nDCG}^{\mathcal{S}}$. Finally, we look at the trade-off between precision and diversity measured by the ILD metrics.

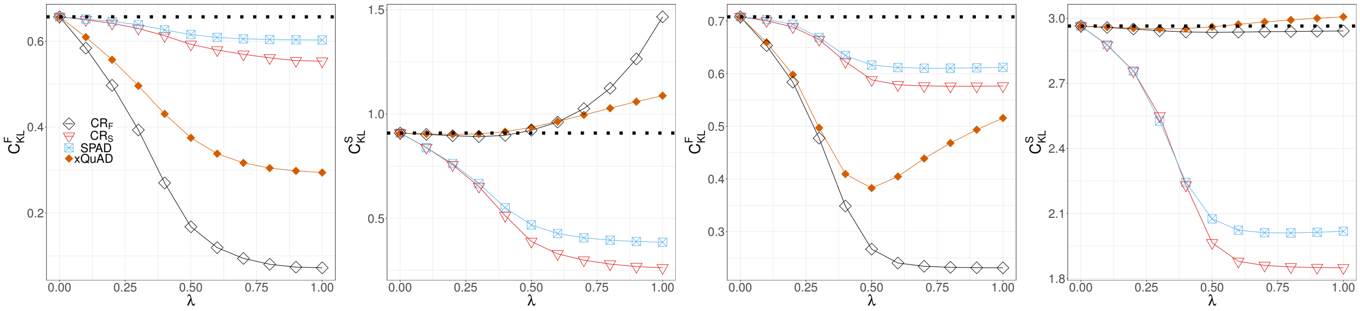
6.1 Calibration results

The calibration results are shown in Figure 1. It is important to keep in mind that for C_{KL} (unlike other results in this paper) smaller values are better: smaller values mean better coverage of the user's interests.

Figure 1a shows results on the ML20M dataset when calibration is measured using item features, ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{F}}$. $\mathrm{CR}_{\mathcal{F}}$ performs best: it has the smallest values of ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{F}}$ for all values of λ. This result is to be expected: $\mathrm{CR}_{\mathcal{F}}$ re-ranks baseline recommendations using the ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{F}}$ metric.

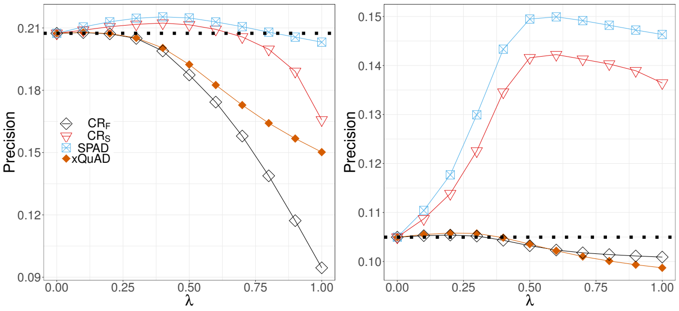
We hypothesized that intent-aware approaches would result in calibrated recommendations to a certain extent. xQuAD tries to cover different user interests defined by item features, and so it should do well for this metric. Indeed, it can be seen in Figure 1a that, for all values of λ, xQuAD results in good calibration: not as good as $\mathrm{CR}_{\mathcal{F}}$, of course, but better than $\mathrm{CR}_{\mathcal{S}}$ and SPAD. However, for all values of λ, even $\mathrm{CR}_{\mathcal{S}}$ and SPAD recommendations are more calibrated than the baseline. This is noteworthy, since $\mathrm{CR}_{\mathcal{S}}$ and SPAD make no use of item features.

Figure 1b shows results on ML20M when calibration is measured using subprofiles, ${\mathrm{C}_{\mathrm{KL}}}^{\mathcal{S}}$. This time, it is xQuAD and $\mathrm{CR}_{\mathcal{F}}$ that are at a disadvantage. We can see that, $\mathrm{CR}_{\mathcal{S}}$ performs best, as expected. But, for all values of λ, SPAD performs better than the baseline too. For smaller values of λ, xQuAD and $\mathrm{CR}_{\mathcal{F}}$ perform close to the



(a) $\mathrm{C_{KL}}^{\mathcal{F}}$ on ML20M **(b) $\mathrm{C_{KL}}^{\mathcal{S}}$ on ML20M** **(c) $\mathrm{C_{KL}}^{\mathcal{F}}$ on TasteProfile** **(d) $\mathrm{C_{KL}}^{\mathcal{S}}$ on TasteProfile**

Figure 1: For ML20M and TasteProfile, $\mathrm{C_{KL}}$ measured using features and subprofiles for different values of λ. Values for MF are shown by dotted lines.



(a) Precision on ML20M **(b) Precision on TasteProfile**

Figure 2: Precision values for different values of λ for ML20M and TasteProfile. Values for MF are shown by dotted lines.

baseline. But for larger values of λ, they perform worse than the baseline.

Figures 1c and 1d show calibration on TasteProfile. The results are quite similar to those for ML20M. The difference is, when calibration is measured by subprofiles (Figure 1d), $\mathrm{CR}_{\mathcal{F}}$ only performs close to the baseline.

6.2 Precision

Figures 2a and 2b plot precision for different values of λ on ML20M and TasteProfile test sets, respectively. Consider ML20M first. $\mathrm{CR}_{\mathcal{S}}$ and SPAD perform well. Precision only falls below the baseline for high values of λ: from 0.7 for $\mathrm{CR}_{\mathcal{S}}$ and from 0.8 for SPAD. $\mathrm{CR}_{\mathcal{F}}$ and xQuAD do not do so well: for many values of λ, their precision is lower than that of the baseline's original recommended set. $\mathrm{CR}_{\mathcal{F}}$ suffers even more than xQuAD: its precision falls even more sharply as λ grows. These $\mathrm{CR}_{\mathcal{F}}$ results confirm those reported by Steck: he found that, for larger values of λ, $\mathrm{CR}_{\mathcal{F}}$'s accuracy drops quickly (although he measured recall rather than precision) [19]. We see similar results when we turn to the TasteProfile dataset. The main differences are that, $\mathrm{CR}_{\mathcal{S}}$ and SPAD now achieve higher precision than the baseline for all values of λ; and $\mathrm{CR}_{\mathcal{F}}$ and xQuAD suffer smaller decreases in precision relative to the baseline than they did on ML20M. These results for precision show a clear preference for approaches that use subprofiles, rather than item features.

6.3 Diversity results

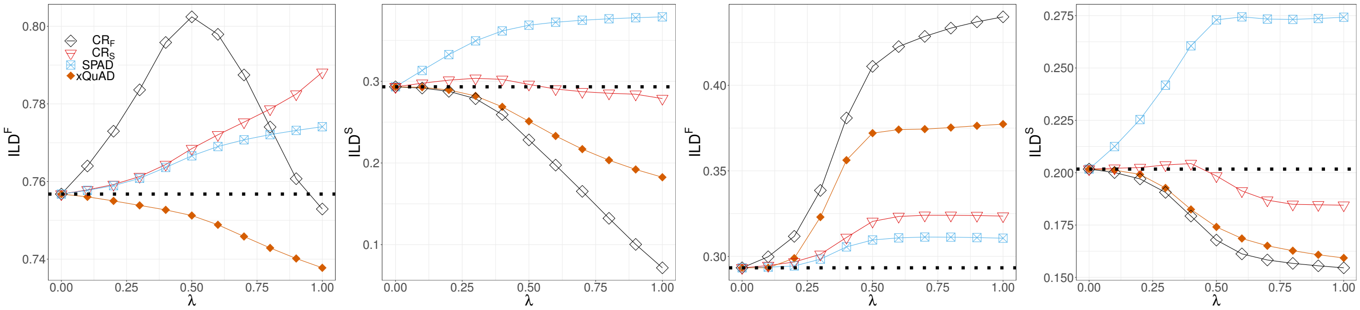
In this subsection, we measure diversity in different ways, enabling us to see the extent to which calibrated recommendations are diverse recommendations.

Figure 3 shows results for ILD. Figure 3a shows results on the ML20M dataset when ILD is measured using item features, $\mathrm{ILD}^{\mathcal{F}}$. Surprisingly, $\mathrm{CR}_{\mathcal{F}}$, which is not an algorithm that explicitly seeks to diversify result sets, achieves the highest values of $\mathrm{ILD}^{\mathcal{F}}$. $\mathrm{CR}_{\mathcal{S}}$ and SPAD are at their usual disadvantage when a metric uses features. But, for all values of λ, their $\mathrm{ILD}^{\mathcal{F}}$ exceeds the baseline. xQuAD is another surprise. For all values of λ, it performs worse than the baseline. This appears to be an idiosyncratic result, specific to this dataset. xQuAD's $\mathrm{ILD}^{\mathcal{F}}$ on a number of datasets that contain explicit ratings, for example, usually exceeds the baseline [10, 23], and it exceeds the baseline by a small amount for most values of λ when we run it on ML20M without binarization (not shown here).

When ILD is measured using subprofiles, $\mathrm{ILD}^{\mathcal{S}}$ (Figure 3b), $\mathrm{CR}_{\mathcal{F}}$ and xQuAD are the ones at a disadvantage. Sure enough, their diversity according to this metric always falls below the baseline, with $\mathrm{CR}_{\mathcal{F}}$ worse than xQuAD. On the other hand, $\mathrm{CR}_{\mathcal{S}}$ and SPAD are at an advantage. SPAD, the algorithm that actually seeks to diversify, produces the most diverse results sets, better than the baseline recommender for all values of λ; $\mathrm{CR}_{\mathcal{S}}$, which only seeks to calibrate its recommendations, produces result sets that are quite similar in diversity to those produced by the baseline.

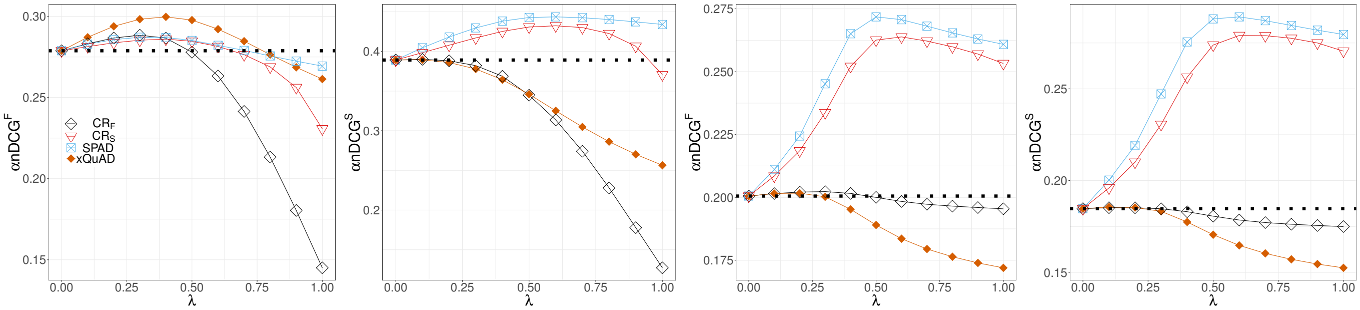
Figures 3c and 3d show ILD results on TasteProfile. Results for $\mathrm{ILD}^{\mathcal{S}}$ (Figure 3d) are similar to those for ML20M. The difference is $\mathrm{ILD}^{\mathcal{F}}$ (Figure 3c). For all values of λ, all of the re-ranking algorithms, including xQuAD, have higher values of $\mathrm{ILD}^{\mathcal{F}}$ than the baseline.

The α-nDCG results are in Figure 4. Figure 4a shows results on the ML20M dataset when α-nDCG is measured using item features, $\alpha\text{-nDCG}^{\mathcal{F}}$. For almost all values of λ, xQuAD performs better than the other re-ranking algorithms. For small values of λ, $\mathrm{CR}_{\mathcal{F}}$ is competitive with $\mathrm{CR}_{\mathcal{S}}$ and SPAD, but soon suffers from the largest decreases in $\alpha\text{-nDCG}^{\mathcal{F}}$. $\mathrm{CR}_{\mathcal{S}}$ and SPAD perform similarly to each



(a) $\text{ILD}^{\mathcal{F}}$ **on ML20M** **(b)** $\text{ILD}^{\mathcal{S}}$ **on ML20M** **(c)** $\text{ILD}^{\mathcal{F}}$ **on TasteProfile** **(d)** $\text{ILD}^{\mathcal{S}}$ **on TasteProfile**

Figure 3: For ML20M and TasteProfile, ILD **measured using features and subprofiles for different values of** λ**. Values for MF are shown by dotted lines.**



(a) α-nDCG$^{\mathcal{F}}$ **on ML20M** **(b)** α-nDCG$^{\mathcal{S}}$ **on ML20M** **(c)** α-nDCG$^{\mathcal{F}}$ **on TasteProfile** **(d)** α-nDCG$^{\mathcal{S}}$ **on TasteProfile**

Figure 4: For ML20M and TasteProfile, α-nDCG **measured using features and subprofiles for different values of** λ**. Values for MF are shown by dotted lines.**

other. Even though they are at a disadvantage, they achieve higher α-nDCG$^{\mathcal{F}}$ than the baseline for all but large values of λ.

Figure 4b shows ML20M results for α-nDCG$^{\mathcal{S}}$, where xQuAD and CR$_{\mathcal{F}}$ are the algorithms that are at a disadvantage. SPAD always has higher α-nDCG$^{\mathcal{S}}$ than all the other algorithms, including the baseline. CR$_{\mathcal{S}}$ has higher α-nDCG$^{\mathcal{S}}$ than the baseline, except when λ is large. xQuAD and CR$_{\mathcal{F}}$ are never better than the baseline and perform particularly poorly as λ grows.

Figures 4c and 4d show α-nDCG results on TasteProfile. Results for α-nDCG$^{\mathcal{S}}$ (Figure 4d) are similar to those for ML20M. The difference is α-nDCG$^{\mathcal{F}}$ (Figure 4c). Now, SPAD and CR$_{\mathcal{S}}$ are always higher than the baseline, while xQuAD and CR$_{\mathcal{F}}$ are similar to the baseline for small values of λ and fall a little below the baseline for large values of λ.

The diversity results show that calibrated recommendations can be diverse as well. Approaches that use subprofiles as aspects, CR$_{\mathcal{S}}$ and SPAD, perform particularly well according to the relevance-aware diversity metric, α-nDCG. In fact, SPAD increases diversity, no matter how it is measured, on both datasets for almost all values of λ. To aid visualization of the relevance/diversity trade-off better, the next subsection plots precision and ILD together.

6.4 The precision/diversity trade-off

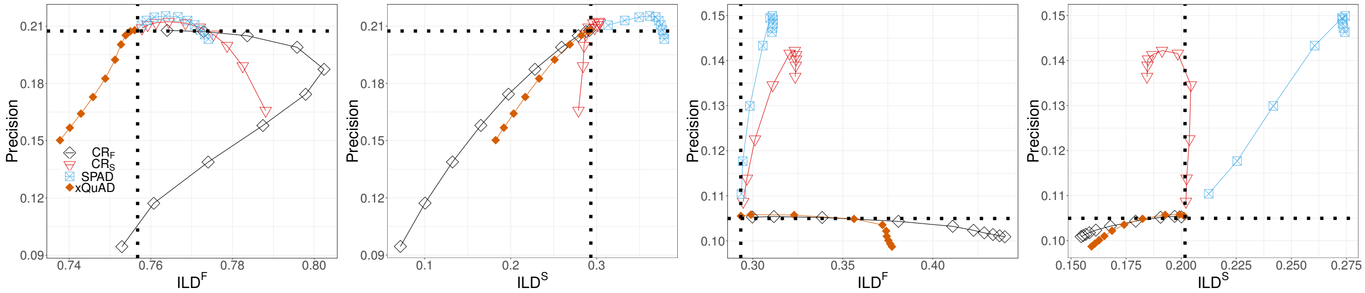
Each subfigure in Figure 5 is divided into four by the dotted lines that plot the precision and ILD values of the MF baseline. When, for a given value of λ, a re-ranking algorithm improves both precision and ILD over the baseline, for example, it appears as a point in the top-right quadrant.

We can see that across all four subfigures (i.e. for both datasets and both version of ILD), assuming that we regard precision and diversity as equally important, SPAD is best by far. It most often increases both precision and diversity. CR$_{\mathcal{S}}$ is second best according to these visualizations. CR$_{\mathcal{F}}$ and xQuAD are not competitive.

6.5 Other results

For completeness, we mention two approaches to diversification whose results we do not show because they clutter the plots but are not competitive.

The first is Maximal Marginal Relevance (MMR) [2]. In MMR, there is no attempt to cover the user's interests. Instead, the item that gets greedily inserted into the re-ranked list is one that is relevant but 'different' from the ones that have already been inserted. 'Difference' is usually measured using Eq. 20. While MMR has some of the highest values for ILD, as one would expect, since this is



(a) Prec./ILD$^{\mathcal{F}}$ on ML20M **(b) Prec./ILD$^{\mathcal{S}}$ on ML20M** **(c) Prec./ILD$^{\mathcal{F}}$ on TasteProfile** **(d) Prec./ILD$^{\mathcal{S}}$ on TasteProfile**

Figure 5: For ML20M and TasteProfile, Precision vs. ILD trade-off plots measured using features and subprofiles for different values of λ. Values for MF are shown by dotted lines.

close to what it optimizes, it performs poorly on all other metrics, almost always having the worst values for precision, for example.

The second is proposed by Steck in [19] with the goal of bringing extra diversity to calibrated recommendations. Steck introduces a diversity-promoting prior, $p_0(f)$, to recommend from features that are not in the user's profile. Specifically, probability distribution $p(f|u)$ is replaced by $\tilde{p}(f|u) = \beta p_0(f) + (1-\beta)p(f|u)$. There are no experimental results in [19]. But we have implemented it. We take $p_0(f)$ to be the average $p(f|u)$ over all the users [19]. The most useful results were obtained when we optimized β for α-nDCG rather than for precision or ILD. In this case, we found that, although the approach does improves diversity a little (measured by ILD), it harms precision and the $\mathrm{C_{KL}}$ metrics. Overall, it was not competitive with the simpler forms of calibrated recommendation.

7 CONCLUSIONS

In this paper, we compared calibrated and intent-aware recommendation. They have apparent similarities in that both try to cover the user's interests to a certain extent. We showed that intent-aware approaches can define interests in terms of item features (xQuAD) or subprofiles (SPAD). In a similar vein, we defined a new instantiation of calibrated recommendation that uses subprofiles ($\mathrm{CR}_{\mathcal{S}}$) in place of features ($\mathrm{CR}_{\mathcal{F}}$). All four can be implemented by greedy re-ranking algorithms that offer a $1 - \frac{1}{e}$ optimality guarantee.

On two publicly available datasets, we compared xQuAD and SPAD with $\mathrm{CR}_{\mathcal{F}}$ and $\mathrm{CR}_{\mathcal{S}}$, measuring calibration, precision and diversity. Since existing calibration and diversity metrics use features in their formulation, re-ranking approaches using subprofiles are at a disadvantage in the experiments. We adapted the existing metrics to produce new ones that use subprofiles instead of features. By using all these metrics in the experiment, we get a more balanced view of the performance of the algorithms.

We find that intent-aware recommendation results in calibrated recommendations to a certain extent, and calibrated recommendation results in diverse recommendations to a certain extent. We also see that re-ranking approaches using features, xQuAD and $\mathrm{CR}_{\mathcal{F}}$, harm precision a lot. The re-ranking approaches that use subprofiles, SPAD and $\mathrm{CR}_{\mathcal{S}}$, achieve the highest precision, achieve good calibration according to both calibration metrics, and achieve good diversity according to both α-nDCG metrics. SPAD also achieves good diversity according to both ILD metrics and suffers least from the relevance/diversity trade-off.

Future work should include user trials and on-line evaluation to determine how users *perceive* calibrated and diverse recommendations both where calibration and diversification are done with respect to features and to subprofiles.

ACKNOWLEDGMENTS

This paper emanates from research supported by a grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 which is co-funded under the European Regional Development Fund.

REFERENCES

[1] Rakesh Agrawal, Sreenivas Gollapudi, Alan Halverson, and Samuel Ieong. 2009. Diversifying search results. In *Procs. of the 2nd ACM International Conference on Web Search and Data Mining*. 5–14.

[2] Jaime Carbonell and Jade Goldstein. 1998. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Procs. of the 21st ACM SIGIR International Conference on Research and Development in Information Retrieval*. ACM, 335–336.

[3] Charles L A Clarke, Maheedhar Kolla, Gordon V Cormack, Olga Vechtomova, Azin Ashkan, Stefan Büttcher, and Ian MacKinnon. 2008. Novelty and diversity in information retrieval evaluation. In *Procs. of the 31st ACM SIGIR International Conference on Research and Development in Information Retrieval*. 659–666.

[4] Farzad Eskandanian, Bamshad Mobasher, and Robin Burke. 2017. A Clustering Approach for Personalizing Diversity in Collaborative Recommender Systems. In *Procs. of the 25th Conference on User Modeling, Adaptation and Personalization*. ACM, 280–284.

[5] Jonathan L Herlocker, Joseph A Konstan, Loren G Terveen, and John T Riedl. 2004. Evaluating collaborative filtering recommender systems. *ACM Transactions on Information Systems (TOIS)* 22, 1 (2004), 5–53.

[6] Marius Kaminskas and Derek Bridge. 2016. Diversity, Serendipity, Novelty, and Coverage: A Survey and Empirical Analysis of Beyond-Accuracy Objectives in Recommender Systems. *ACM Trans. Interact. Intell. Syst.* 7, 1 (2016), 2:1–2:42.

[7] Mesut Kaya. 2019. *Subprofile Aware Diversification of Recommendations.* Ph.D. Dissertation. School of Computer Science & Information Technology, University College Cork.

[8] Mesut Kaya and Derek Bridge. 2018. Accurate and Diverse Recommendations Using Item-Based SubProfiles. In *Procs. of the Thirty-First International Florida Artificial Intelligence Research Society Conference*. AAAI, 462–467.

[9] Mesut Kaya and Derek Bridge. 2018. Automatic Playlist Continuation using Subprofile-Aware Diversification. In *Proceedings of the ACM Recommender Systems Challenge 2018.*

[10] Mesut Kaya and Derek Bridge. 2019. Subprofile-aware diversification of recommendations. *User Modeling and User-Adapted Interaction* (2019).

[11] Dawen Liang, Jaan Altosaar, Laurent Charlin, and David M Blei. 2016. Factorization meets the item embedding: Regularizing matrix factorization with item co-occurrence. In *Procs. of the 10th ACM Conference on Recommender Systems*. 59–66.

[12] Sean M McNee, John Riedl, and Joseph A Konstan. 2006. Being Accurate is Not Enough: How Accuracy Metrics Have Hurt Recommender Systems. In *Procs. of the CHI'06 Extended Abstracts on Human Factors in Computing Systems.* 1097–1101.

[13] George L Nemhauser, Laurence A Wolsey, and Marshall L Fisher. 1978. An analysis of approximations for maximizing submodular set functionsâĂŤI. *Mathematical programming* 14, 1 (1978), 265–294.

[14] Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you.* Penguin UK.

[15] István Pilászy, Dávid Zibriczky, and Domonkos Tikk. 2010. Fast ALS-based Matrix Factorization for Explicit and Implicit Feedback Datasets. In *Procs. of the Fourth ACM Conference on Recommender Systems.* ACM, 71–78.

[16] Rodrygo L T Santos, Craig Macdonald, and Iadh Ounis. 2010. Exploiting Query Reformulations for Web Search Result Diversification. In *Procs. of the 19th International Conference on World Wide Web*. 881–890.

[17] Guy Shani and Asela Gunawardana. 2011. Evaluating Recommendation Systems. In *Recommender Systems Handbook.* Springer, 257–297.

[18] Barry Smyth and Paul McClave. 2001. Similarity vs. diversity. In *Procs. of the International Conference on Case-Based Reasoning*. Springer, 347–361.

[19] Harald Steck. 2018. Calibrated Recommendations. In *Procs. of the 12th ACM Conference on Recommender Systems.* 154–162.

[20] Saúl Vargas, Linas Baltrunas, Alexandros Karatzoglou, and Pablo Castells. 2014. Coverage, redundancy and size-awareness in genre diversity for recommender systems. In *Procs. of the 8th ACM Conference on Recommender systems.* 209–216.

[21] Saúl Vargas, Pablo Castells, and David Vallet. 2011. Intent-oriented Diversity in Recommender Systems. In *Procs. of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval.* ACM, 1211–1212.

[22] Saúl Vargas, Pablo Castells, and David Vallet. 2012. Explicit Relevance Models in Intent-oriented Information Retrieval Diversification. In *Procs. of the 35th ACM SIGIR International Conference on Research and Development in Information Retrieval.* ACM, 75–84.

[23] Saúl Vargas Sandoval. 2015. *Novelty and Diversity Evaluation and Enhancement in Recommender Systems.* Ph.D. Dissertation. Universidad Autónoma de Madrid, Spain.

[24] Jacek Wasilewski and Neil Hurley. 2016. Intent-Aware Diversification Using a Constrained PLSA. In *Procs. of the 10th ACM Conference on Recommender Systems.* 39–42.

[25] Cai-Nicolas Ziegler, Sean M McNee, Joseph A Konstan, and Georg Lausen. 2005. Improving recommendation lists through topic diversification. In *Procs. of the 14th International Conference on World Wide Web.* 22–32.