Personalised Diversification Using Intent-Aware Portfolio

Jacek Wasilewski
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
jacek.wasilewski@insight-centre.org

Neil Hurley
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
neil.hurley@insight-centre.org

ABSTRACT

The intent-aware diversification framework considers a set of aspects associated with items to be recommended. A baseline recommendation is greedily re-ranked using an objective that promotes diversity across the aspects. In this paper the framework is analysed and a new intent-aware objective is derived that considers the minimum variance criterion, connecting the framework directly to portfolio diversification from finance. We derive an aspect model that supports the goal of minimum variance and that is faithful to the underlying baseline algorithm. We evaluate diversification capabilities of the proposed method on the MovieLens dataset.

KEYWORDS

recommender systems; diversity; intent-aware recommendations; portfolio theory

1 INTRODUCTION

Recommender systems have become ubiquitous in the interfaces to product catalogues provided by on-line retailers. Recommender algorithms are used to filter a large set of possible selections into a much smaller set of items that the user is likely to be interested in. Engaging and holding a user's interest is a complex matter. Simply identifying relevant items, without taking into account the user's context, can lead to accurate recommendations when measured directly against user's past preferences, but may not be very engaging. Recommendation diversification addresses this issue, by widening the range of possible item types recommended to the user, to match true user needs in an uncertain environment.

Different frameworks have been developed for diversification of recommendations. Initial work [17] was focused on the intra-list diversity (ILD), that is, the average distance between a set of items. Meanwhile, in Information Retrieval (IR), a similar problem of diversifying search query results was explored [1, 9], and the challenge was to address the ambiguity in search queries. The *intent-aware* framework models the problem as one of identifying intents and attempting to ensure all possible intentions are addressed. If a "jaguar" is queried, the framework will return documents that are relevant to the animal and the car manufacturer, to make sure the user receives at least one satisfying result. The work of Vargas et al. [15] unifies the IR and recommender system perspectives.

In this paper, we carry out a detailed analysis of the intent-aware diversification framework. We identify two components of the framework as being, (a) an objective which drives the diversification and (b) an aspect model that holds user intent and item aspects used by the objective. We show that the current state-of-the-art does not achieve its objective and we propose and discuss a different intent-aware personalised diversification objective, based on the minimum variance criterion that drives Markowitz portfolio diversification in modern portfolio theory. We argue that to achieve good diversification, an accurate aspect model that is faithful to the underlying baseline algorithm is required. This deviates from current practise in which the state-of-the-art aspect models contain certain biases which make them unsuitable for accurate recommendation and prevent them from achieving their diversification objective, which should be paramount, since it defines the type of diversification that the framework is trying to inject.

The main contributions of this paper are: (a) an analysis of the intent-aware framework and its components, (b) a normalised relevance-based aspect model that exactly matches the baseline scores over which the diversifier is built, (c) a personalised intent-aware covariance and an intent-aware portfolio ranking objective.

The paper is organised as follows: we discuss the state-of-the-art in Section 2, in Section 3 we introduce the intent-aware diversification framework. In Section 4 we propose an intent-aware portfolio and in Section 5 we derive a faithful aspect model. Evaluation is reported in Section 6 and we conclude in Section 7.

Notation. We consider the task of recommending top-N items $i \in \mathcal{I}$ to a user $u \in \mathcal{U}$. We write $R_u \subset \mathcal{I}$ for the recommended set, or simply R if the user context is clear. Write $P_u \subseteq \mathcal{I}$ as the *user profile* of items rated by user u. A set $C \subset \mathcal{I}$ of top-rated *candidate* items is provided by a *baseline* recommender for *re-ranking*, along with the scores $s(i|u) \in \mathbb{R}^+$ used to sort the items $i \in C$. We select $R_u \subset C$, such that $|R_u| = N$. There exists a set of explicit, known *aspects* $\mathcal{A}$, and we write $\mathcal{A}_i \subseteq \mathcal{A}$ for the aspect associated with item i. Generally single users are denoted by u, items by i and j, and we write i_k to represent the k^{th} item in a ranked list.

2 MODELLING AND EVALUATING DIVERSITY OF RECOMMENDATIONS

Different models and evaluation frameworks have been proposed for diversity enhancement in information retrieval and recommender systems. As there are many frameworks to choose from, to analyse the success of the diversification, it is necessary to have some sense of what an ideal result list should consist of. Do we simply want to maximise the total number of aspects covered by

This project has been funded by Science Foundation Ireland under Grant No. SFI/12/RC/2289.

UMAP'17 Adjunct, July 9–12, 2017, Bratislava, Slovakia

DOI: http://dx.doi.org/10.1145/3099023.3099067

the recommendation, or do we want to ensure that the aspects are not covered by just a few items in the result list, but rather spread across the recommendation. These and other possibilities are reflected in a number of frameworks proposed in the literature.

Intra-List Diversity (ILD), the most commonly considered type of diversity, emphasises mutual dissimilarity of items. Dissimilarity is based on a distance metric between items, and many possibilities exist, e.g. content-based distance. Framework favours recommendations where items are as different as possible according to this distance. Typically greedy optimisation techniques [2, 17] are used (such as `MMR`) where final recommendations are obtained by mixing relevance scores and dissimilarities. Work of [13] extends the framework to include rank- and relevance-awareness.

Another view on diversification has been proposed by Wang and Zhu in [16], inspired by the work of Markowitz [8] in the modern portfolio theory of finance—the **Mean-Variance diversification framework**. In investment applications, it states that buying two correlated stocks might result in high profits or high losses. It is argued that accepting lower variance in profits/loses leads to better and safer overall outcomes. A re-ranking method has been proposed that transforms possibly high-risk recommendations into lower-risk by accepting lower mean relevance. In recommendations, high risk arises due to uncertainty of user needs and this should be reduced if possible. Wang and Zhu used a covariance matrix based on historical ratings to model item-item relationships. A variation using a covariance matrix created on explicit item-aspect assignments was proposed and evaluated in [9]. Another adaptation of the Mean-Variance framework has been proposed in [10] for latent factor models by employing variance of learned user latent factors to address users' individual situations.

More recently, the **proportionality framework** [4], focuses on the coverage of aspects, claiming that these should be represented by items in the final recommendation proportionally to a user's interest in them. A greedy re-ranking method has been proposed called the *proportionality method* (PM), that is inspired on the seat assignment system for elections in some countries. This framework emphasises coverage of aspects and proportionality of user's interests, however redundant items are not penalised and framework does not consider item's rank or model item relevance.

In [12], Vargas et al. focus on genre-based diversity and propose the **Binomial framework** as a way of measuring and enhancing recommendations in terms of genre coverage, redundancy and list size-awareness. A recommendation list can be seen as a sequence of Bernoulli trials, where a trial models the selection of an item covering each genre. This framework emphasises coverage by measuring the loss of not covering a particular genre, and redundancy through tolerance an user has for getting more items of the same aspect. Rank and relevance however is not taken into account.

3 INTENT-AWARE DIVERSIFICATION

In this paper we focus on the intent-aware diversification framework of Agrawal et al. [1], later extended in [9, 14, 15]. The framework aims to ensure a good spread of explicit aspects among the items in the result list to mitigate the uncertainty. In IR, if multiple aspects are related to a search query, without explicit information about the query's intent, items representing many different relevant aspects are returned, hoping that at least one will satisfy user's need. In the recommender systems (RS) setting, Vargas et al. [15] introduced the notion of user intent as an analogue of query intent. User intents are described in terms of a probability distribution over a set of aspects based on interest previously expressed. A similar ambiguity as in IR arrises—for any given interaction, what aspect is the user currently interested in? If an incorrect assumption is made, an RS may generate recommendations that user finds irrelevant.

Explicit Query Aspect Diversification framework (`xQuAD`) is the state-of-the-art optimisation method for selecting the recommendation set so that redundancy over aspects is reduced. It is a generalisation of the `IA-Select` method [1] and it assumes the existence of a probabilistic *aspect model* [6], defined as probabilities of a user's interests in aspects, $p(a|u)$ such that $\sum_{a\in\mathcal{A}} p(a|u) = 1$, and probabilities of selecting an item, $p(i|a,u)$.

Final recommendations are obtained through a multi-objective greedy optimisation of the relevance (baseline score $s(i|u)$) and a diversity set objective (`IA-Select`). Starting with $S = \emptyset$, recommendations are constructed in an iterative way, by greedily selecting at each iteration the item i that satisfies:

$$i^* = \arg\max_{i\notin S}(1-\lambda)s(i|u) + \lambda \sum_{a\in\mathcal{A}} p(a|u)p(i|a,u)\prod_{j\in S}(1-p(j|a,u)) \quad (1)$$

and updating $S \leftarrow S \cup \{i^*\}$. Parameter $\lambda \in [0,1]$ controls the trade-off between relevance and diversity. `IA-Select` can be considered as a sub-case of `xQuAD` when $\lambda = 1$.

The above shows that the intent-aware framework is comprised of: (a) an aspect model consisting of the probabilities $p(a|u)$ and $p(i|a,u)$; and (b) a redundancy model, or set-oriented objective, such as `IA-Select`. In [14], relevance-aware model is proposed where instead of $p(i|a,u)$, probabilities of item relevance, $p(\text{rel}_i\,|a,u)$, are used, where rel_i is a binary relevance value for item i. This model is assumed in the rest of the paper. Moreover, we make the simplifying assumption of conditional independence of relevance.

Diversity is a feature of the whole recommendation, rather than a feature of a single item—a *set-oriented* measure—in that it evaluates the entire set of recommended items—or a *rank-oriented* measure, if the rank-order of items is taken into account. Rather than thinking about diversity as a desirable attribute of recommendations that is to one side of, or separate to, performance per se, it is possible to arrive at diversity criteria by considering performance goals and this is indeed how the `IA-Select` expression is formulated.

As we discuss below, different set objectives lead to greedy re-ranking expressions of the form

$$i^* = \arg\max_{i\notin S} \sum_{a\in\mathcal{A}} p(a|u)p(\text{rel}_i\,|a,u)\beta(S,a,u) \quad (2)$$

where S is the set of previously selected items and $\beta(.)$ is a *redundancy* term, which generally penalises items when they contain aspects that are already in the set S.

The `IA-Select` objective is derived from the principle that *at least one* item in the recommended set should be relevant. Given the aspect a, this can be written as $1 - \prod_{i\in R}(1 - p(\text{rel}_i\,|a,u))$. Hence, using conditional independence, the overall objective that at least one item in R is relevant is

$$o_{\texttt{IA-Select}}(R) \triangleq 1 - \sum_{a\in\mathcal{A}} p(a|u)\prod_{i\in R}(1-p(\text{rel}_i\,|a,u))). \quad (3)$$

While maximisation of $o_{\texttt{IA-Select}}(R)$ is NP-hard, the `IA-Select` re-ranker, represents a greedy approach to optimising Eqn.(3) and

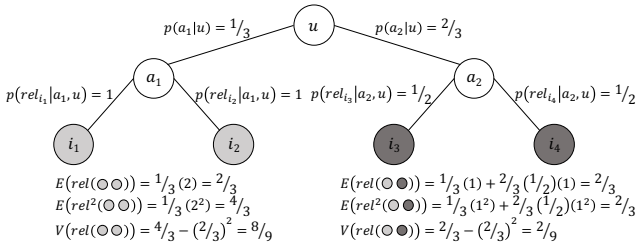

Figure 1: Example of variance minimisation diversification.

performs well in practice due to the submodularity of $o_{\texttt{IA-Select}}(R)$. Hence, greedy optimisation of $o_{\texttt{IA-Select}}(R)$ takes the form of Eqn.(2), where $\beta_{\texttt{IA-Select}}(S,a,u) \triangleq \prod_{j\in S}(1-p(\text{rel}_j\,|a,u))$.

Using *hit* to mean that a set contains at least one relevant item, the $o_{\texttt{IA-Select}}$ is the expected number of hits. Empirically this is the percentage of test users $u \in U$ for whom the model recommends at least one relevant item in the test set T_u:

$$relhits = \frac{1}{|U|}\sum_{u\in U}\mathbb{1}_{(R_u\cap T_u\neq\emptyset)}$$

As we will see later, it is difficult in practice for a greedy re-ranker to have a positive impact on relevance hits. Uncertainty in the aspect model means that the predicted relevance hits may not be sufficiently accurate to allow the best sets to be distinguished from the rest. Instead, we argue that a better context for diversification can be obtained through risk minimisation, as discussed next.

4 INTENT-AWARE PORTFOLIO DIVERSIFICATION

The issue of diversification has long been studied in the context of portfolio optimisation in financial domain. In the financial context, there are two primary objectives in portfolio selection. One is to maximise the expected return on investment, while the other is to minimise the risk, as measured by the variance of the return. Diversification across assets with negatively correlated returns allows the risk on one asset to be hedged by another, so that a good expected return can be maintained while reducing the risk.

The classical Mean-Variance (MV) portfolio selection criterion of Markowitz [8], selects a portfolio of $|C|$ assets by minimising

$$\mathbf{w}^* = \arg\min \mathbf{w}^T\Sigma\mathbf{w} \tag{4}$$

such that $\mathbf{w}^T\boldsymbol{\mu} = q$ and $\mathbf{w}^T\mathbf{1} = 1$, where Σ is the covariance of asset returns, $\boldsymbol{\mu}$ is the vector of mean asset returns and q is the expected return of the portfolio. The weights $\mathbf{w} \in \mathbb{R}^{|C|}$ represent the fraction of each asset to include in the portfolio.

Diversification across the items in recommendation lists may be understood analogously: the items being recommended correspond to assets, and their relevance to the assets' returns. Recommendation algorithms typically seek to maximise a recommendation's return. Similar to portfolio selection, items might be selected so that the variance of the relevance is minimised as well.

In [16], Wang proposed a ranking objective following portfolio optimisation, where the effectiveness is measured as:

$$\text{E}[\text{rel}(R)] = \sum_{k=1}^{N} w_{i_k}\,\text{E}[\text{rel}_{i_k}] \tag{5}$$

where w_{i_k} is a weight assigned to the k^{th} ranked item in the recommender list, such that the weight corresponds to a rank-discount, depending on the item's position in the list and $\sum_{k=1}^{N} w_{i_k} = 1$.

The variance of the expected overall relevance is defined as:

$$\text{Var}[\text{rel}(R)] = \text{E}[\text{rel}(R)^2] - \text{E}[\text{rel}(R)]^2 = \sum_{i,j\in R} w_i w_j c_{ij} \tag{6}$$

where c_{ij} is the covariance of the relevance between items i and j. By combining Eqn.(5) and Eqn.(6), we end up with the following:

$$o_{\texttt{MV}} = \text{E}[\text{rel}(R)] - \alpha\,\text{Var}[\text{rel}(R)] \tag{7}$$

where α, similarly to λ in xQuAD, is a parameter adjusting the risk level, thus diversification. Although greedy re-ranking is sub-optimal for the minimum variance problem, it performs well in practise and hence, we can represent a greedy strategy for variance minimisation as one of selecting a new item according to:

$$i^* = \arg\max_{i\notin S} w_i\left(\text{E}[\text{rel}_i] - \alpha w_i c_{ii} - 2\alpha\sum_{j\in S} w_j c_{ij}\right).$$

The key issue in the mean-variance framework is to obtain accurate covariance and relevance estimates. One approach is to use baseline scores $s(i|u)$ to estimate the relevance and historical rating data to estimate covariance, as suggested in [16]. Another possibility is explored in [9] where explicit item-aspect associations are used to estimate the covariance. Both of these options are global estimates, i.e. they are independent of the user's preferences. A user-dependent covariance has been proposed in [10], however this is in the context of latent factorisation models.

It is possible to derive an intent-aware diversification criterion in terms of variance minimisation that is personalised and uses explicit aspects by deriving an intent-aware covariance. To further illustrate, consider the example, shown in Figure 1. With binary relevance $\text{rel} \in \{0,1\}$, the total relevance of a set of two items can be $\{0,1,2\}$. We assume that the user is interested in one particular aspect a, with probability $p(a|u)$, and, given an aspect, the relevance of items is independent. In our scenario, items i_1 and i_2 are both relevant, once the user is interested in aspect a_1, and i_3 and i_4 have a 50% chance of being relevant when interested in a_2. The expected relevance of any two-item set is $2/3$. However, the variance of relevance depends on the choice of items. In fact, if items i_1 and i_2 are chosen, the variance is $8/9$. If either i_1 or i_2 is paired with i_3 or i_4, the variance reduces to $2/9$. Rather than placing all bets on the user being interested in a_1, by spreading the recommendation across aspects, the variance has been minimised which is desired.

The aspect model provides tractable expressions for the mean and variance that appear in MV. In particular, we can write:

$$E[\text{rel}(R)] = \sum_{i\in R} w_i \sum_a p(a|u)p(\text{rel}_i\,|a,u)$$

Moreover, using conditional independence, the user-dependent covariance is:

$$c_{ij} = \begin{cases} \sum_a p(a|u)p(\text{rel}_i\,|a,u)p(\text{rel}_j\,|a,u) - \\ \sum_a p(a|u)p(\text{rel}_i\,|a,u)\sum_a p(a|u)p(\text{rel}_j\,|a,u) & i\neq j \\ \sum_a p(a|u)p(\text{rel}_i\,|a,u)(1-\sum_a p(a|u)p(\text{rel}_i\,|a,u)) & i=j \end{cases} \tag{8}$$

Both the xQuAD and MV methods are optimisation methods that combine expected relevance and some diversification criterion for selecting items. In fact, we note that the greedy approach to minimising variance, has the same form as the maximisation problem,

Eqn.(2), where the redundancy term may now be written as

$$\beta_{\texttt{IA-MV}}(S, a, u, i) \triangleq w_i \sum_{j \in S} w_j \left(\sum_{a'} p(a'|u) p(\text{rel}_j | a', u) - p(\text{rel}_j | a, u) \right) + w_i^2 (1 - \sum_{a'} p(a'|u) p(\text{rel}_i | a', u)) .$$

to select the item with the largest increase in the negative variance.

Placing our proposed method in the context of the state-of-the-art, IA-MV may be understood as an instance of the xQuAD with a new redundancy term, $\beta_{\texttt{IA-MV}}(S, a, u, i)$. It may also be seen as an instantiation of MV, where we estimate a user-dependent covariance c_{ij} that is obtained directly from the aspect model. A slight difference between the formulations is that, while in xQuAD the redundancy term is traded against the baseline score $s(i|u)$, in MV, we trade variance against the expected relevance, $E[\text{rel}(R)]$. It is noteworthy that the aspect model provides a tractable expression directly for the variance, $\text{Var}[\text{rel}(R)]$, of the performance objective $\text{rel}(R)$. This contrasts with other work on minimum variance, where the covariance is often not of the performance measure per se but derived from side-information, such as item-aspect profiles [9]. Thus our formulation offers the possibility of accepting a drop in the overall mean of our performance metric, for the sake of minimising its variance—a worthy objective, in and of itself, even without considering its impact on diversity.

The success of our method, however, depends crucially on the accuracy of the aspect model. Not only it must accurately predict mean performance, but also accurately estimate the covariance.

5 ASPECT MODELS

The starting point for re-ranking is the output of a baseline recommendation algorithm, consisting of the set C of top-ranked candidate items, along with the scores $s(i|u)$ that the baseline used to rank these items. Given such a baseline, we can consider the problem of computing an aspect model, as one of *factoring* the baseline model into the form:

$$p(\text{rel}_i | u) = \sum_{a \in \mathcal{A}} p(a|u, s) p(\text{rel}_i | a, u, s), \tag{9}$$

where we have added $s = s(i|u)$ to the dependencies, to emphasise that this model is derived from baseline scores. We assume that $s \geq 0$ and that s is correlated with relevance, so that $s_i > s_j$ implies that the baseline predicts item i to be more relevant than item j. This model enables the recommendation process to be understood in terms of item aspects. We emphasise that the aspects are *explicit* attributes associated with items. The associations are *fixed* and *known a-priori* allowing users to directly perceive and understand the impact of any diversification process that is applied.

In previous work [14, 15], Vargas used RxQuAD to denote his version of xQuAD that uses a relevance-based aspect model. Of the two, RxQuAD is developed to model binary relevance of items, rather than to model item choice. As our formulations depend on binary relevance, we compare our method with RxQuAD.

Vargas' Relevance Model (vrm). In the RxQuAD model, Vargas computes $p(a|u)$ and $p(\text{rel}_i | a, u)$ based on user profile and baseline recommendations as:

$$p(a|u) = \frac{|\{i \in P_u : a \in \mathcal{A}_i\}|}{\sum_{a' \in \mathcal{A}} |\{i \in P_u : a' \in \mathcal{A}_i\}|}, \quad p(\text{rel}_i | a, u) = \frac{2^{\mathbb{1}(a \in \mathcal{A}_i) \frac{s(i|u)}{s^*(a|u)}} - 1}{2}$$

where $s^*(a|u)$ is the highest score given by the baseline recommendation engine for an item containing aspect a. The probability $p(a|u)$ is based only on the user profile, while $p(\text{rel}_i | a, u)$ is entirely based on the recommendation scores $s(i|u)$ and item-aspect associations. The formulae are entirely heuristic. If a top-N recommendation is formed using item relevance as computed using this aspect model and Eqn.(9), then the recommendation performance is very poor in comparison to the original baseline. As we will see in our evaluation, this model cannot be used as a basis from which to compute a meaningful estimate of $E[\text{rel}(R)]$ and $\text{Var}[\text{rel}(R)]$.

Normalised Relevance Model (nrm). We seek an aspect model that provides good estimates of $E[\text{rel}(R)]$ and $\text{Var}[\text{rel}(R)]$. Assuming that the baseline is a high-accuracy recommendation engine, we develop an aspect model that faithfully represents the baseline i.e. one for which $E[\text{rel}_i | u] \geq E[\text{rel}_j | u]$, whenever $s(i|u) > s(j|u)$. Our starting point is a mapping of the baseline score $s(i|u)$ to the probability of relevance, which we write as $g(s(i|u)) = p(\text{rel}_i | u, s)$. Given $g(.)$, we write the joint distribution of relevance and aspect as: $p(\text{rel}_i, a|u) = g(s(i|u)) p(a | \text{rel}_i, u)$, which implies

$$p(\text{rel}_i | a, u) = \frac{g(s(i|u)) p(a | \text{rel}_i, u)}{p(a|u)}.$$

It follows that $p(a|u) \geq g(s(i|u)) p(a | \text{rel}_i, u)\ \forall i$. There is clearly some degree of freedom in choosing $p(a|u)$. We choose $p(a|u) \propto \max_{j \in C} g(s(j|u)) p(a|rel_j, u) \triangleq m_a$. Then, since $\sum_{a \in \mathcal{A}} p(a|u) = 1$, we have $p(a|u) = m_a / m$, where $m = \sum_{a \in \mathcal{A}} m_a$. From this, we obtain a model that satisfies $p(rel_i|u) = mg(s(i|u))$, so that the ranking according to the model is identical to the baseline ranking. We call this model the normalised relevance model, nrm.

The conditional probabilities, $p(a | \text{rel}_i, u)$, can be taken as uniform, i.e. $p(a | \text{rel}_i, u) = 1/|\mathcal{A}_i|$. For all user-item pairs (u, i) in the training set, scores $s(i|u)$ are generated and a threshold th is chosen such that we come up with class labels $\text{rel}(i|u) = 1$ if $r_{ui} \geq th$ and $\text{rel}(i|u) = 0$ otherwise. We generate a sample of scores $s(i|u)$ for a set of randomly chosen user item pairs and choose a class label $\text{rated}(i|u) = 1$ if (u, i) is in our training set and $\text{rated}(i|u) = 0$ otherwise. The relevance function $g(.)$ is inferred using logistic regression as a combination of the probability of relevance of a (u, i) pair given that it *has been rated* by the user ($l(s)$), and the probability that a user-item pair is rated ($r(s)$), given score s:

$$\begin{aligned} g(s) &= p(\text{rel} | \text{rated}, s) p(\text{rated}|s) + p(\text{rel} | \neg\text{rated}, s) p(\neg\text{rated}|s) \\ &= l(s) r(s) + \text{rel}_b (1 - r(s)) \end{aligned}$$

where rel_b is a prior relevance score for unrated user-item pairs. Since precision is calculated over held-out rated items, to match our relevance model against precision, we take $\text{rel}_b = 0$.

In summary, with nrm, we have obtained aspect probabilities, $p(a|u)$ and $p(\text{rel}_i | a, u)$, such that the use of Eqn.(9), to obtain relevance scores to rank the items , yields exactly the same ranking as the original baseline. Moreover, we have chosen the aspect probabilities so that they provide a good estimate of the covariance.

6 EVALUATION

6.1 Experiment Setup

We performed our evaluation on the MovieLens 1M dataset [5] containing 1M ratings from 6K users on 3.7K movies. Movies are described using at least one genre out of 18 available, with an average of 1.98 genres per item. Ratings are made on a 5-star scale.

Several algorithms available in the RankSys framework have been used to produce baseline recommendations, on which we evaluate different re-ranking methods. Due to space restrictions, we report results only on the matrix factorisation (MF) baseline [7]—diversification over other baselines follows similar patterns. Candidate sets of $|C| = 100$ top-ranked items are generated. The re-ranking strategies are then applied to produce a top-N set of items, where $N = 20$. Trade-offs between the relevance and diversity are explored for different values of λ and α by performing a grid search.

Although our formulation of the `MV` and `IA-MV` objectives includes a rank-discount weight, w_i, since `IA-Select` is a set-oriented objective, we take $w_i = 1/N$ for fair comparison.

6.2 Evaluation Metrics

A large number of metrics have been proposed to measure diversity. A good review can be found in [11]. Agrawal et al. [1] proposed an intent-aware generalisation of some standard metrics to account for aspects, for instance ERR-IA [3], the intent-aware expected reciprocal rank. ERR-IA can be seen as a special case of another commonly used diversity metric, α-nDCG, which we focus on here. In contexts such as movie recommendation, where a single movie can simultaneously belong to multiple genres, the relevance of multi-genre items gets multiply counted by α-nDCG, when the above expression is used. The impact is that items with e.g. 3 aspects is more important than relevant item with just one aspect. An alternative is to weight an item's relevance by the number of aspects it contains: $rel(i|u,a) = rel(i|u)\mathbb{1}(a \in \mathcal{A}_i) \,/\, |\mathcal{A}_i|$.

As well as α-nDCG, we focus on some more intuitive metrics which we briefly describe below:

(1) S-recall (*subtopic recall*): a measure of how well the recommendation covers the aspect space; S-recall $= \frac{|\cup_{i \in R_u | \text{rel}(i|u)=1} \mathcal{A}_i|}{|\mathcal{A}|}$

(2) DNG: a measure of how early new aspects are introduced to a ranked list; DNG $= \sum_{k=1}^{N} 2^{-(k-1)} \text{rel}(i_k|u) G(k)$, where $G(k)$ is the number of new aspects at rank k at which a relevant item appears.

(3) SDI (*subtopic dispersion index*): a measure of the distribution of aspects across items; SDI $= \sigma^2_{c_a} \,/\, \mu_{c_a}$ where $\forall_{a \in \mathcal{A}} : c_a = \sum_{i \in R_u | rel(i|u)=1} \mathbb{1}(a \in \mathcal{A}_i)$. This is the ratio of the variance to the mean of the occurrence count of subtopics in the recommendation. Unlike the S-recall, it measures whether subtopics are over-represented or not. A lower value indicates a well-balanced recommendation set, with subtopics evenly distributed across the items.

Accuracy is evaluated through precision. We use binary relevance with threshold 4. Also, relevance-aware versions of the diversity metrics are used (following [13]).

6.3 Results and Analysis

Diversification Objective. As discussed, the `IA-Select` objective is to maximise the probability that at least one relevant item is recommended. We empirically checked if the `xQuAD` optimisation

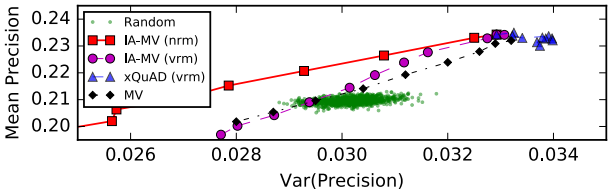

Figure 2: Variance vs precision (expected return).

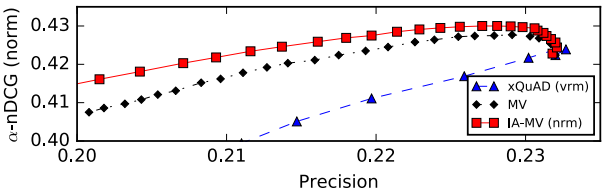

Figure 3: Precision/Diversity trade-offs.

method meets this objective. For a various values of λ we can see *relhits* is decreasing as the redundancy term contributes more and more (λ increases), counter to the objective of the optimiser. Also, using the `IA-Select` re-ranker, more users lose relevant items in the recommendations than gained. Although `IA-MV` does not optimise for this objective, we can check its impact on it. We find that slightly fewer users benefit by gaining a relevant item when they had none in the baseline, however fewer users see disimprovements in their overall recommendations. Neither method is able to actually improve on this performance metric.

A similar analysis of the target objective can be carried out for `IA-MV`. Figure 2 shows the trade-off between mean precision, and its variance, evaluated against the test set. First, `xQuAD` fails to explore the trade-off–it focuses on a very narrow region of precision/variance and is actually increasing the variance while keeping the same or similar mean precision. The `IA-MV` with different risk preferences α, allows a set of recommendations with varying levels of mean precision to be explored. For comparison, we plot the performance of random re-rankings. We see that when our method produces a higher mean precision to these random recommendations for a similar variance of precision.

Importance of the aspect model. In Figure 2 the performance of `IA-MV` is shown when either the `vrm` and the `nrm` aspect models are used to compute the covariance. It can be seen that `IA-MV` with `nrm` outperforms `IA-MV` with `vrm`. What is more, `IA-MV` with `vrm` is not much better than random re-ranking of the baseline recommendations. This illustrates the importance of selecting an accurate aspect model.

Precision/Diversity Trade-off. Figure 3 presents the trade-off between precision and diversity as measured by α-nDCG. `IA-MV` is offering the best precision/diversity trade-off, improving α-nDCG, when the normalised relevance measure is used. With the non-normalised relevance measure, `xQuAD` out-performs `IA-MV`, giving a larger α-nDCG value for a given precision operation point. It may be observed that `xQuAD` with the `vrm` aspect model tends to promote items with multiple aspects, which, as discussed earlier, results in a large contribution to the non-normalised α-nDCG, whenever the item is relevant. It can be debated as to whether or not this is a desirable feature of a re-ranker. In our opinion, the normalised

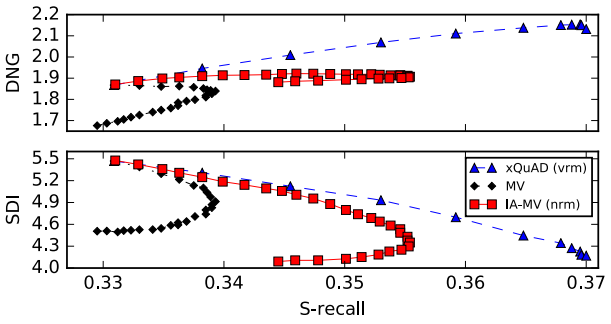

Figure 4: Diversity trade-offs: S-recall against DNG and SDI.

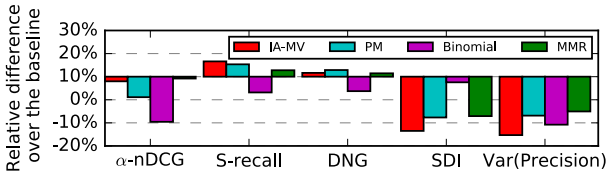

Figure 5: Comparison against other diversification methods.

weighting is a fairer measure of an item's relevance. Users are not doubly or triply more satisfied with an item just because it covers two or three aspects—they are simply satisfied when it covers the aspect of current interest to them.

Aspect-based diversity. As pointed out in [12], certain diversity metrics are correlated and it can be useful to analyse them against each other. In Figure 4 we plot DNG and SDI, against S-recall. This allows to assess how aspects are spread across the list when a certain level of diversity, as measures using S-recall, is achieved. It can be seen that MV offers lowest SDI but also lowest DNG for a given S-recall value, indicating little tendency for aspects to be introduced early in the list. On the other hand, xQuAD offers high DNG and SDI for a given S-recall value and IA-MV falls between xQuAD and MV. We can conclude for IA-MV that there is a tendency for aspects to appear early in recommender lists, but without the strong bias towards multi-aspect items of xQuAD, we see that there is a greater spread (i.e. low SDI) of aspects across the ranks of the list. In IA-MV, as the number of relative aspects appearing in recommender lists increases, they tend to appear around the same rank positions, while in MV, as we achieve more relevant aspects on average, they are more likely to appear lower in the list, as indicated by the decrease in DNG. This indicates that, the diversification of IA-MV is more likely to be perceived by users than that of MV. Moreover this perceptible diversification is being achieved along with the low variance advantage of a minimum variance diversifier.

Overall, IA-MV has some clear advantages over its comparators. As a true minimum variance optimiser, it allows for the trade-off between risk and return to be addressed. In effect, it allows us to ensure that any particular user's experience is not far from the mean experience. MV enables the same trade-off to be explored but not to the extent of IA-MV , but xQuAD does not. Evaluation on the diversity metrics indicates that IA-MV is achieving this result by more perceptible manipulation of explicit aspects than MV—new aspects appear higher in the recommendation list, without resort to strong bias towards multiple-aspect items.

Other diversification methods. Finally, we compare our method against other diversification frameworks. In Figure 5, we show relative differences over the baseline recommendation of these, for a fixed precision of approx. 0.20. IA-MV obtains a small drop on α-nDCG similarly to MMR. IA-MV does not improve DNG more then other methods (as seen before). The most notable feature is that IA-MV is generally beating other methods on S-recall, SDI and variance of precision which shows that, at this level of precision: a) a low variance result is achieved and, b) relevant items contain more aspects and their distribution is more even across the list.

7 CONCLUSIONS

In this paper we tackled the problem of intent-aware diversification as a way of addressing ambiguity of user's needs. We proposed a new optimisation objective bridging the intent-aware framework and the mean-variance portfolio theory through a personalised covariance, and a normalised relevance aspect model to drive the objective. We showed that our proposed objective compromises between xQuAD and MV, and if compared with other diversification frameworks, it shows a better or comparable performance.

REFERENCES

[1] R. Agrawal, S. Gollapudi, A. Halverson, and S. Ieong. Diversifying Search Results. In *Proceedings of the 2nd ACM Conference on Web Search and Data Mining.*

[2] J. Carbonell and J. Goldstein. The Use of MMR, Diversity-based Reranking for Reordering Documents and Producing Summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[3] O. Chapelle, S. Ji, C. Liao, E. Velipasaoglu, L. Lai, and S. Wu. 2011. Intent-based Diversification of Web Search Results: Metrics and Algorithms. *Inf. Retr.* 14, 6 (Dec. 2011).

[4] V. Dang and W.B. Croft. Diversity by Proportionality: An Election-based Approach to Search Result Diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[5] F.M. Harper and J.A. Konstan. 2015. The MovieLens Datasets: History and Context. *ACM Trans. Interact. Intell. Syst.* 5, 4 (Dec. 2015).

[6] T. Hofmann. 2004. Latent Semantic Models for Collaborative Filtering. *ACM Trans. Inf. Syst.* 22, 1 (Jan. 2004).

[7] Y. Hu, Y. Koren, and C. Volinsky. 2008. Collaborative Filtering for Implicit Feedback Datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining (ICDM '08).*

[8] H. Markowitz. 1952. Portfolio Selection. *The Journal of Finance* 7, 1 (1952), 77–91.

[9] R.L.T. Santos, C. Macdonald, and I. Ounis. 2012. On the Role of Novelty for Search Result Diversification. *Inf. Retr.* 15, 5 (Oct. 2012).

[10] Y. Shi, X. Zhao, J. Wang, M. Larson, and A. Hanjalic. Adaptive Diversification of Recommendation Results via Latent Factor Portfolio. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[11] S. Vargas. 2015. *Novelty and Diversity Evaluation and Enhancement in Recommender Systems.* Ph.D. Dissertation. Universidad Autnoma de Madrid.

[12] S. Vargas, L. Baltrunas, A. Karatzoglou, and P. Castells. Coverage, Redundancy and Size-awareness in Genre Diversity for Recommender Systems. In *Proceedings of the 8th ACM Conference on Recommender Systems.*

[13] S. Vargas and P. Castells. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. In *Proceedings of the Fifth ACM Conference on Recommender Systems.*

[14] S. Vargas, P. Castells, and D. Vallet. Explicit Relevance Models in Intent-oriented Information Retrieval Diversification. In *Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[15] S. Vargas, P. Castells, and D. Vallet. Intent-oriented Diversity in Recommender Systems. In *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval.*

[16] J. Wang. Mean-Variance Analysis: A New Document Ranking Theory in Information Retrieval. In *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval.*

[17] C. Ziegler, S.M. McNee, J.A. Konstan, and G. Lausen. Improving Recommendation Lists Through Topic Diversification. In *Proceedings of the 14th International Conference on World Wide Web.*