Proceedings of the Twenty-Ninth International
Florida Artificial Intelligence Research Society Conference

Incorporating Diversity in a Learning to Rank Recommender System

Jacek Wasilewski and **Neil Hurley**
Insight Centre for Data Analytics
University College Dublin
Dublin, Ireland
{jacek.wasilewski, neil.hurley}@insight-centre.org

Abstract

Regularisation is typically applied to the optimisation objective of matrix factorisation methods in order to avoid overfitting. In this paper, we explore the use of regularisation to enhance the diversity of the recommendations produced by these methods. Given a matrix of pairwise item distances, we add regularisation terms dependent on the item distances to the accuracy objective of a learning to rank matrix factorisation formulation. We examine the impact of these regularisers on the latent factors produced by the algorithm and show that such regularisation does indeed promote diversity. The regularisation comes at a cost of performance in terms of accuracy and ultimately the approach cannot greatly enhance diversity without a consequent fall-off in accuracy.

Introduction

Recommender systems have become ubiquitous in online systems and services. Their goal is to help address the choice overload problem, by filtering a large set of possible selections into a much smaller set of recommended items that a user is likely to be interested in. Recommendations are generally based on a set of implicit or explicit item ratings gathered from users in past interactions. Matrix factorisation has proven an effective means of producing accurate recommendations. In such methods, the rating matrix is factored into two low-rank matrices, representing user- and item-latent factors and predicted ratings are obtained by multiplying the corresponding user and item factors. Interest in promoting the diversity of recommendations has increased in recent years. In general, the promotion of diversity is in opposition to the requirement of high accuracy. Many offline studies have found that the more diverse a recommendation is, the less likely it is to match the user's preference and conversely, a highly accurate set is likely to consist of many similar recommendations. Some work, e.g. (Ekstrand et al. 2014), has found positive correlations between diversity and accuracy as subjectively perceived by users in user trials. To date, many of the approaches to diversity enhancement have been developed in the context of memory-based algorithms, or alternatively, diversity enhancement has been considered as a separate post-processing step carried out after the initial rating predictions have been obtained. In this paper, we explore whether it is possible to tackle the accuracy and diversity problems together in a single training phase.

Regularisation has been used in matrix factorisation algorithms, principally to control for overfitting, by constraining the size of the latent factors. However, in the literature on recommender systems, different types of regularisers have been proposed in order to incorporate other side information into the objective function optimisation to support the recommendation. For example, in (Jamali and Ester 2010) a *social regularisation* is proposed to incorporate social network information into the optimisation, by encouraging users who are close in the social network to have similar user latent factors. We are motivated by such work to consider whether appropriate regularisation can be used to enhance diversity. Given an item distance matrix, in this paper, we propose a number of regularisations that use the distance matrix to encourage the optimisation to produce factors that result in a diverse set of items in the recommendation list. We evaluate the regularisation methods on two datasets. Ultimately, we observe that, while some of the proposed regularisers are effective in promoting diversity, the diversity cannot be largely improved without a consequent fall-off in accuracy.

The paper is organised as follows: after reviewing the state-of-the-art and summarising the learning-to-rank method upon which our regularisers are applied, we propose a number of different regularisation terms and discuss their likely effectiveness. We then describe how to incorporate such regularisation in an alternating least squares optimisation framework. In the evaluation section, we test the regularisers and compare their performance.

Related Work

The generation of personalised rankings from implicit feedback data has received some attention in recent work in recommender systems (Hu, Koren, and Volinsky 2008; Pilászy, Zibriczky, and Tikk 2010; Jahrer and Töscher 2012; Takács and Tikk 2012). In this paper, we focus on incorporating diversity into the learning to rank algorithm for implicit feedback, proposed in (Takács and Tikk 2012), although our method can be applied to any matrix factorisation formulation. Work on diversity has largely focused on variants of the Maximum Marginal Relevance (MMR) re-ranking principle introduced originally in (Carbonell and Goldstein 1998) and used the diversification of recommendations in work such as (Ziegler et al. 2005; Zhang and Hurley 2008). In this approach, the final recommendation is produced in two steps: first a list of recommendation candidates is produced for each user and then the top-N items are selected one by one in a way that an item and a list of already selected items has the highest diversity value. Re-ranking based on the intent-aware framework has also been proposed (Vargas, Castells, and Vallet 2011). In contrast to this work, we focus on tackling accuracy and diversity jointly during model training. A comprehensive framework for evaluating novelty and diversity is given in (Vargas and Castells 2011). We use this framework in the evaluation of our proposed method.

Measuring Diversity

Given a set, U, of users with $n = |U|$ and a set, I, of items with $m = |I|$ and an $n \times m$ matrix R containing ratings given by the each user for some of the m items, the top-N recommendation problem is, for a given user u, to recommend a list L_u of N items that the user is likely to enjoy. The accuracy of the recommendation algorithm can be measured using various different metrics, by comparing L_u with hold-out data. We assume that there also exists an $m \times m$ matrix D, with elements $d(i, j)$ giving a distance between items i and j and that, as well as being accurate, we would like the recommendation to be novel or diverse. Within the framework for evaluating novelty and diversity in recommender systems proposed in (Vargas and Castells 2011), the novelty of items is measured with the respect to a particular context. We concentrate on the *expected intra-list diversity* (EILD) which measures the novelty of recommended items with respect to the other items in the recommended list. A recommended list with a high EILD value contains items that are very different to each other, according to the distance measure, $d(.,.)$. The full expression incorporates rank discount and relevance-awareness, such that, given a recommended list $L_u = \{i_1, \ldots, i_N\}$ of size $N = |L_u|$ for a user u,

$$\mathrm{EILD}(L_u) = \sum_{k=1,l=1;l\neq k}^{N} C_k \mathrm{disc}(k)\mathrm{disc}(l|k)p(rel|i_k,u)p(rel|i_l,u)d(i_k,i_l)\,,$$

where $\mathrm{disc}(l|k) = \mathrm{disc}(\max(1, l-k))$ reflects a relative rank discount for an item at position l knowing that position k has been reached, rel is the relevance of an item to a user, and C_k is a normalising constant. Ignoring rank and relevance, the metric reduces to the intra-list distance (ILD) (Zhang and Hurley 2008; Ziegler et al. 2005):

$$\mathrm{ILD}(L_u) = \frac{1}{N(N-1)} \sum_{i,j \in L_u} d(i,j)$$

Learning to Rank for Recommendation

We focus on matrix factorisation approaches to recommendation in which the training phase involves learning a low-rank $n \times k$ latent user matrix P and a low-rank $m \times k$ latent item matrix Q, such that the estimated rating $\hat{r}_{ui}$ can be expressed as:

$$\hat{r}_{ui} = \mathbf{p}_u^T \mathbf{q}_i\,,$$

where $\mathbf{p}_u^T$ is the u^{th} row of P, $\mathbf{q}_i^T$ is the i^{th} row of Q and k is the chosen dimension of the latent space. P and Q are learned through the minimisation of an accuracy-based objective. A number of such objectives have been proposed in the literature and the regularisation methods we propose here could be incorporated with any such objective. Since we are interested in ranking rather than rating prediction, we focus on the learning to rank objective function proposed originally in (Jahrer and Töscher 2012) and further developed in (Takács and Tikk 2012) i.e. we take $\mathrm{acc}(\mathrm{P}, \mathrm{Q})$ to be:

$$\sum_{u\in U}\sum_{i\in I} c_{ui} \sum_{j\in I} s_j[(\hat{r}_{ui} - \hat{r}_{uj}) - (r_{ui} - r_{uj})]^2 + \beta(\|\mathrm{Q}\|^2 + \|\mathrm{P}\|^2) \tag{1}$$

where c_{ui} and s_j are parameters of the objective function and β is the standard regularisation parameter of norm-based regularisation to avoid over-fitting. We consider the implicit feedback case in which $c_{ui} = 0$ if $r_{ui} = 0$, and 1 otherwise. The role of c_{ui} is to select user-item pairs corresponding to positive feedbacks from all possible pairs. s_j is an importance weighting for item j.

Regularisation to Enhance Diversity

We explore the use of regularisation to enhance the diversity of the recommendation, by choosing an optimisation objective of the form

$$\min_{\mathrm{P},\mathrm{Q}} \mathrm{acc}(\mathrm{P}, \mathrm{Q}) + \lambda \mathrm{reg}(\mathrm{P}, \mathrm{Q})$$

where $\mathrm{acc}(.)$ is the accuracy objective and $\mathrm{reg}(.)$ is a regularisation term. To choose an appropriate regularisation, it is useful to initially consider the diversity objective alone. Representing a recommendation as an m-dimensional vector $\mathbf{x}$ such that $x(i) = 1$ when item $i \in L_u$ and $x(i) = 0$ otherwise, the ILD of the recommendation may be written as the quadratic form

$$\frac{1}{N(N-1)} \mathbf{x}^T \mathrm{D} \mathbf{x} = \frac{1}{N(N-1)} \sum_{i}^{m} \sum_{j}^{m} d(i,j) x(i) x(j)\,. \tag{2}$$

Expanding (2) using the real eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_m$ of the symmetric matrix D, with corresponding eigenvalues $\alpha_1 \geq \cdots \geq \alpha_m$ we have that

$$\mathbf{x}^T \mathrm{D} \mathbf{x} = \sum_{j=1}^{m} \alpha_j (\mathbf{v}_j^T \mathbf{x})^2\,. \tag{3}$$

This suggests that a high diversity set can be selected by choosing $\mathbf{x}$ to maximise $(\mathbf{v}_1^T \mathbf{x})^2$, that is, to choose $x(i) = 1$ when i is among the N largest components $(v_1(1)^2, \ldots, v_1(m)^2)$. Applying this rationale to the selection of a regulariser, we note that fixing P and solving

$$\max_{\mathrm{Q}} \sum_{l=1}^{k} \mathrm{Q}(\cdot,l)^T \mathrm{D} \mathrm{Q}(\cdot,l) = \mathrm{tr}(\mathrm{Q}^T \mathrm{D} \mathrm{Q})\,, \tag{4}$$

for $\|\mathrm{Q}\|$ fixed, results in $Q(.,l) = \mathbf{v}_1$ for $l = 1, \ldots k$. The resulting ratings, $\hat{r}_{ui}$, are then proportional to $v_1(i)$, since $\mathbf{p}_u^T \mathbf{q}_i = (\sum_l p_{ul}) v_1(i)$. A potential drawback of this approach is that, as we cannot control for the sign of the eigenvector, the largest magnitude components of $\mathbf{v}_1$ may correspond to the smallest ratings rather than the largest.

An alternative regulariser is given by:

$$\sum_{ij} d(i,j) \|\mathbf{q}_i - \mathbf{q}_j\|^2 = \mathrm{tr}(\mathrm{Q}^T \mathrm{L}_D \mathrm{Q}) \tag{5}$$

where $\mathrm{L}_D = \mathrm{E} - \mathrm{D}$ is the Laplacian matrix of D and E is the diagonal matrix with i^{th} diagonal entry equal to $\sum_j d(i,j)$. Minimising a quadratic form of the Laplacian is a well-known strategy for minimising the *edge-cut* of a bipartitioning of the rows of the matrix D which in our context amounts to minimising

$$\sum_{i \in R} \sum_{j \notin R} d(i,j) \,.$$

The critical points again occur at the eigenvectors, $\mathbf{v}$, of L_D. The Laplacian has some nice properties with respect to optimisation. It is positive semi-definite with a minimum eigenvalue of 0 obtained for the eigenvector $\mathbf{v} = (1, \ldots, 1)^T$. It follows that for all other eigenvectors, $\sum_j v(j) = 0$. Moreover, a vector in the span of the eigenvectors corresponding to the largest few eigenvalues tends to have high magnitude values in components corresponding to high diversity sets. Since $\sum_j v(j) = 0$, these components occur on either side of the sorted vector i.e. they are either large positive or negative values, so that the sign problem of the previous regulariser is no longer a problem. The above regularisers can be naturally extended to P using the following expressions:

$$\sum_u \sum_{ij} d(i,j) (\mathbf{p_u}^T (\mathbf{q}_i - \mathbf{q}_j))^2 = \mathrm{tr}(\mathrm{P} \mathrm{Q}^T \mathrm{L}_D \mathrm{Q} \mathrm{P}^T) \,,$$

and

$$\sum_u \sum_{ij} d(i,j) (\mathbf{p_u}^T \mathbf{q}_i)(\mathbf{p_u}^T \mathbf{q}_j) = \mathrm{tr}(\mathrm{P} \mathrm{Q}^T \mathrm{D} \mathrm{Q} \mathrm{P}^T) \,.$$

We explore these regularisers in the case of the Netflix genre distance (see Evaluation section for a description of the dataset). In this case, the eigenvalues of D lie in the range $[-1412.4, 7437.9]$ and the eigenvalues of L_D lie in the range $[0, 9319.5]$. We perform a gradient descent update

$$\mathrm{Q}^{(l+1)} = \mathrm{Q}^{(l)} - \alpha(\lambda \mathrm{Q}^{(l)} \mathrm{H} \pm \nabla \mathrm{reg}(\mathrm{Q}^{(l)})$$

to optimise the above regularisers using the the Netflix genre distance (see later section) and a randomly chosen P matrix. Here $\mathrm{H} = \mathrm{P}^T \mathrm{P}$ when using one of the P-dependent regularisers and $\mathrm{H} = \mathrm{I}$ otherwise. λ is chosen to ensure a positive definite Hessian and the sign in front of the gradient term is chosen as negative when maximisation of the regularisation term is required and positive when minimisation is required. To ensure global convergence, $\alpha = 1.9/(\lambda + \alpha_1)$ for maximisation and $\alpha = 1.9/(\lambda - \alpha_m)$ for minimisation. Ten update iterations are carried out. The resulting Q is used to generate ratings and the top $N = 50$ ratings are selected to form the recommender set. The average diversity of the resulting set over 100 randomly selected users is shown in Table 2, which can be compared in Table 1 with the baseline mean ILD values obtained for random sets, sets produced by the non-diversified learning-to-rank algorithm (`RankALS`) and maximum diversity sets obtained through a greedy maximisation from a random initial item. From this analysis, maximisation of the Laplacian regularisers would appear to be the best strategy. We will evaluate if this holds true when the regulariser is combined with the accuracy objective.

Table 1: Netflix dataset, $N = 50$ items, Genre Distance

Baseline	**ILD**
`RankALS`	0.66
Random Set	0.77
Max Diversity Set	0.95

Table 2: Netflix dataset, Genre Distance, $N = 50$, ILD achieved by different regularisers

Regulariser		**ILD**	λ
$\max \mathrm{tr}(\mathrm{Q}^T \mathrm{L}_D \mathrm{Q})$	LapDQ-max	0.83	9,320
$\min \mathrm{tr}(\mathrm{Q}^T \mathrm{L}_D \mathrm{Q})$	LapDQ-min	0.77	0
$\max \mathrm{tr}(\mathrm{P}\mathrm{Q}^T \mathrm{L}_D \mathrm{Q}\mathrm{P}^T)$	PLapDQ-max	0.86	9,320
$\max \mathrm{tr}(\mathrm{P}\mathrm{Q}^T \mathrm{L}_D \mathrm{Q}\mathrm{P}^T)$	PLapDQ-min	0.15	0
$\max \mathrm{tr}(\mathrm{Q}^T \mathrm{D} \mathrm{Q})$	DQ-max	0.28	7,500
$\min \mathrm{tr}(\mathrm{Q}^T \mathrm{D} \mathrm{Q})$	DQ-min	0.00	1,500

ALS Algorithm

Naive minimization of (1) is expensive as the number of terms is $T \cdot I$, where T is the number of transactions in the rating matrix R. The original algorithm employed in (Jahrer and Töscher 2012) used the stochastic gradient descent (SGD) algorithm and this was improved to an alternating least squares (ALS) approach in (Takács and Tikk 2012). The ALS consists of two steps – the *P-step* and the *Q-step*, in which the objective function is initially minimised w.r.t. P, keeping Q fixed and then w.r.t. Q keeping P fixed. This requires a calculation of the gradients with respect to P and Q. The P-step may be rearranged into the following linear system to solve for each row of P at step l:

$$\left[\beta \mathrm{I} + \sum_{i \in I} c_{ui} \sum_{j \in I} s_j (\mathbf{q}_i^{(l-1)} - \mathbf{q}_j^{(l-1)})(\mathbf{q}_i^{(l-1)} - \mathbf{q}_j^{(l-1)})^T\right] \mathbf{p}_u^{(l)} =$$
$$\sum_{i \in I} c_{ui} \sum_{j \in I} s_j (r_{ui} - r_{uj})(\mathbf{q}_i^{(l-1)} - \mathbf{q}_j^{(l-1)})$$

Similarly, for each row of Q, at step l, the Q-step may be rearranged as the linear system:

$$\left[\beta \mathrm{I} + \sum_{j \in I} s_j \sum_u c_{ui} \mathbf{p}_u^{(l)} {\mathbf{p}_u^{(l)}}^T\right] \mathbf{q}_i^{(l)} =$$
$$(\sum_u c_{ui} \mathbf{p}_u^{(l)} {\mathbf{p}_u^{(l)}}^T)(\sum_{j \in I} s_j \mathbf{q}_j^{(l-1)}) + \sum_{j \in I} s_j \sum_u c_{ui} (r_{ui} - r_{uj}) \mathbf{p}_u^{(l)} \,.$$

Table 3: Gradients of the Regularisers for ALS algorithm

Regulariser	∇_p	∇_q
PLapDQ	$\mathrm{Q}^T\mathrm{L}_D\mathrm{Q}\mathrm{P}^T$	$\mathrm{P}^T\mathrm{P}\mathrm{Q}^T\mathrm{L}_D$
LapDQ	0	$\mathrm{Q}^T\mathrm{L}_D$
DQ	0	$\mathrm{Q}^T\mathrm{D}$

Note that the Q-step requires old values of Q from the previous iteration in the RHS vector. We refer to this algorithm as RankALS.

We add $\lambda\mathrm{reg}(\mathrm{P},\mathrm{Q})$ to the objective, where we choose $\lambda < 0$ to promote solutions that maximise the regulariser and $\lambda > 0$ to promote solutions that minimise the regulariser. The gradients of the regularisers are summarised in Table 3. For example, in the case of the PLapDQ regulariser, the derivative w.r.t. P is given by $\mathrm{Q}^T\mathrm{L}_D\mathrm{Q}\mathrm{P}^T = \sum_{i,j} l_{ij}\mathbf{q}_i\mathbf{q}_j^T\mathbf{p}_u$. Hence, one possibility to incorporate this regularisation into the ALS update equations is to add $\lambda\sum_{i,j} l_{ij}\mathbf{q}_i^{(l-1)}{\mathbf{q}_j^{(l-1)}}^T$ to the LHS matrix. However, we find that more stable solutions are obtained if instead the RHS is updated with $-\lambda\sum_{i,j} l_{ij}\mathbf{q}_i^{(l-1)}{\mathbf{q}_j^{(l-1)}}^T\mathbf{p}_u^{(l-1)}$. In general, we incorporate the diversity regularisation by modifying the RHS of the update equations using values of P and Q from the previous step.

Diversity Distribution

Testing Diversity

We have discussed some approaches to enhancing the diversity as measured by the ILD of a recommendation list. To test our models, we generate recommendations for a set of $U_{\text{test}} \subseteq U$ of randomly selected test users and estimate the expected ILD of recommendation lists generated by the model using the sample mean ILD observed over these users:

$$\hat{\mu} = \frac{1}{|U_{\text{test}}|N(N-1)} \sum_{u\in U_{\text{test}}} \sum_{i\neq j\in R_u} d(i,j) \qquad (6)$$

To test the significance of any observed differences in ILD, we need the standard error of this sample estimator. However, the pairwise differences are not independent, as each item index i appears $N-1$ times in the set of distances that are averaged to obtain the ILD of each user. In (Giorgi and Bhattacharya 2012), assuming only independence across individuals, an unbiased estimator of the sampling variance is obtained in the context of comparing intra-individual genetic diversity between populations. Adopting this to our context, we find an unbiased estimator for $\mathrm{Var}\{\hat{\mu}\}$ as a weighted sum of the sample covariances:

$$\hat{\sigma}_0 = \sum_u \sum_{i<j<k<l\in R_u} (d(i,j)-\hat{\mu})(d(k,l)-\hat{\mu})$$

$$\hat{\sigma}_1 = \sum_u \sum_{i<j<k\in R_u} (d(i,j)-\hat{\mu})(d(i,k)-\hat{\mu})$$

$$\hat{\sigma}_2 = \sum_u \sum_{i<j<\in R_u} (d(i,j)-\hat{\mu})(d(i,j)-\hat{\mu})\,.$$

Choosing the weights to ensure that the expected value of the estimator is $\mathrm{Var}\{\hat{\mu}\}$, it is possible to show that the resulting estimator is:

$$\hat{\sigma}^2 = \frac{8\hat{\sigma}_0 + 8\hat{\sigma}_1 + 4\hat{\sigma}_2}{|U_{\text{test}}|(|U_{\text{test}}|-1)(N(N-1))^2}$$

which, if we assume that the covariance is zero when all indices differ, reduces to

$$\begin{aligned}\hat{\sigma}^2 &= (8\hat{\sigma}_1 + 4\hat{\sigma}_2)\,/\\ &\quad [|U_{\text{test}}|N(N-1)\times\\ &\quad (|U_{\text{test}}|(N-2)(N-3) + (|U_{\text{test}}|-1)(4N-6))]\end{aligned}$$

a formula which agrees with that obtained in (Giorgi and Bhattacharya 2012) when $|U_{\text{test}}| = 1$. Note that $|U_{\text{test}}|\hat{\sigma}^2$ is an unbiased estimator of the standard deviation of the ILD. We can use this estimator in a paired test for difference in ILD between two models, using the t-statistic,

$$t_{\text{paired}} = \frac{\hat{\mu}_{m_1} - \hat{\mu}_{m_2}}{\sqrt{\hat{\sigma}^2_{m_1} + \hat{\sigma}^2_{m_2}}}$$

which is approximately normally distributed, when $|U_{\text{test}}|$ is large. The mean and spread of ILD values that are possible among a set of items, depends on the method used to calculated the item distance. It is therefore useful to use standardised measures to compare the impact of our diversification methods. We use such standard measures in the following evaluation section.

Evaluation

In this section we present the results of our experiments. First we briefly describe the data sets we have used, followed by the evaluation methodology and finally the results.

Datasets

Two datasets are evaluated as follows:

- *Netflix*: The full Netflix data set (Bennett and Lanning 2007) consists of 100,480,507 ratings from 1 to 5 from 480,189 users on 17,770 items. Using IMDb, 28 movie genres have been identified and associated with the movies in the dataset, such that 9,320 movies have at least one associated genre. Ratings for movies without genres have been removed. Following (Takács and Tikk 2012), ratings are implicitized by assigning 1 if the rating value is 5, and 0 otherwise, leaving 17,678,861 positive implicit ratings for 9,315 items and 457,107 users. This final set has been split an 80/20 ratio into train and test sets, containing, respectively,14,143,088 and 3,535,773 ratings.
- *MovieLens 20m*: The biggest MovieLens data set[1] released in 2015 consists of 20,000,263 ratings from 0.5 to 5 with a step-size of 0.5, from 138,493 users on 27,278 items, enriched by 18 genres. Items without genre information have been removed, implicit ratings have been created from ratings equal to 5, giving a data set consisting of 2,898,660 ratings from 131,839 users and 14,474 items. This has been split into a training set containing 2,318,928 items and test set with 579,732 items.

[1] http://grouplens.org/datasets/movielens/

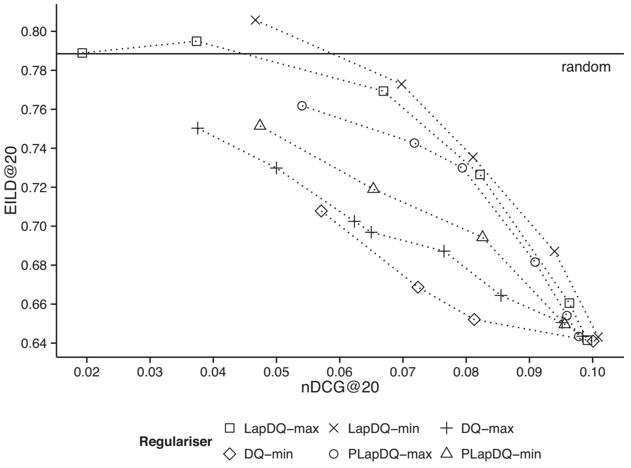

Figure 1: Diversity - accuracy trade-off on the Netflix data set.

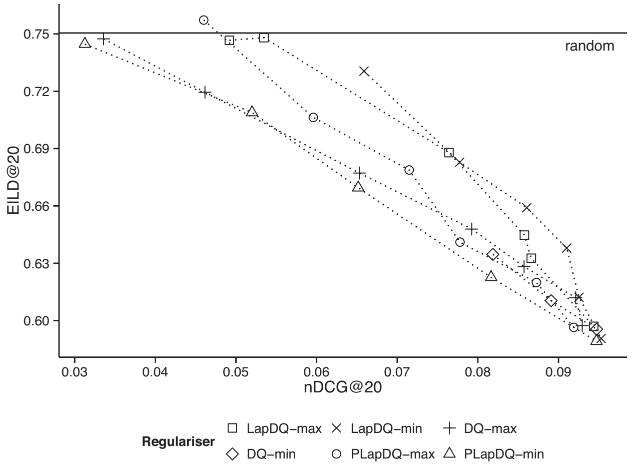

Figure 2: Diversity - accuracy trade-off on the MovieLens 20m data set.

Evaluation Protocol

In all experiments with `RankALS`, we set the number of factors $k = 20$ and we run the training phase for 10 iterations. We set $\beta = 0$, following (Takács and Tikk 2012) in which it is reported that no accuracy improvement is obtained when standard regularisation is used. We set the item importance weighing $s_j = |U_j|$, the number of users who rated j. Accuracy and diversity has been checked for different λ values which control the level of diversity. A set of different metrics has been used to measure accuracy and diversity. For accuracy we report results of Precision, Recall and nDCG, the diversity is measured by EILD; all metrics are evaluated at $N = 20$; As the ratings are binary, we do not employ a relevance model in the calculation of EILD, but we do use the logarithmic rank discount, which is the same as that employed in the nDCG accuracy metric: $\mathrm{disc}(k) = \frac{1}{\log_2(k+1)}$. We also report the expected profile distance (EPD) metric (Vargas and Castells 2011). High value indicates high diversity.

As a baseline, we have used a diversity-enhancing MMR re-ranker. The re-ranker has a λ parameter that controls the accuracy-diversity trade-off. In our experiment this parameter has been set to $\lambda = 0.5$ which means that we equally weight diversity and accuracy. In order to benefit from the re-ranker, a larger candidate set of items has to be picked before generating the final recommendations. We set the size to be twice bigger than the N, which in our case is 40.

The *RankSys*[2] framework has been used to run and evaluate the experiments using built-in metrics.

Results

Figures 1 and 2 show the diversity/accuracy trade-off plots of different regularisation methods, for different λ values on, respectively, the Netflix and MovieLens 20m datasets. For both data sets, the LapDQ regularisers produce the best results. Varying λ allows the level of diversity to be controlled, by deciding how much accuracy to be sacrificed in order to gain diversity. LapDQ-max shows higher tuning possibility. For example, on the Netflix data set, we can increase the diversity from 0.6466 up to 0.7693 with a drop in nDCG from 0.1001 to 0.0669. Using the LapDQ-min regulariser, diversity of 0.6870 can be achieved with drop only to 0.0939. Similar behaviour can be observed on the MovieLens data set: the LapDQ-min regulariser increases diversity from 0.5934 to 0.6468 with a small decrease in the accuracy, from 0.0951 to 0.0887. The DQ regulariser does not perform well on either data sets, confirming the observations of our regulariser analysis. With this regulariser, on the Netflix data set some increase in diversity can be observed, although the decrease in the accuracy is significant, while on MovieLens, it was hard to find any setting that would allow for a useful accuracy/diversity trade-off. The PLapDQ-max regulariser performs reasonable well, while not reaching the performance of the LapDQ regulariser.

Table 4 shows the overall results of our experiments of regulariser approaches. We may observe that all accuracy measures decrease with the increase of diversity. Results are similar whether or not the discount model is used, except for the LapDQ regularisers where the logarithmic discount results in slightly better diversity performance, at least for the Netflix data set. Increase in EILD leads to an increase in the EPD metric as well, with approximately the same trade-off.

Comparison of the re-ranker approach and regularisers approach shows that even though regularisers can beat the re-ranker in terms of higher diversity in some settings, they suffer more in terms of accuracy. Right now, the re-rankers offer better accuracy-diversity trade-off in the off-line evaluation.

We have standardised the EILD results of the best performing algorithms using the mean and standard deviation of a random recommendation. The mean EILD score of `RankALS + LapDQ-min` algorithm on Netflix data set is

[2] http://ranksys.org/

		Prec	Recall	nDCG	EILD		EPD	
					no disc	log disc	no disc	log disc
Netflix	Random	0.0009	0.0022	0.0015	0.7886	0.7885	0.7643	0.7643
	RankALS	0.0479	0.1302	0.1002	0.6431	0.6367	0.6737	0.6721
	+ MMR	**0.0466**	**0.1255**	**0.0959**	**0.7513**	0.7662	0.7167	0.7164
	+ LapDQ-min	0.0423	0.1145	0.0698	0.7476	**0.7729**	0.7350	0.7708
	+ LapDQ-max	0.0414	0.1087	0.0669	0.7466	0.7693	0.7405	0.7757
	+ DQ-min	0.0294	0.0782	0.0571	0.7058	0.7078	0.7356	0.7409
	+ DQ-max	0.0228	0.0513	0.0375	0.7477	0.7503	**0.7827**	0.7911
	+ PLapDQ-min	0.0388	0.0896	0.0653	0.7098	0.7190	0.7465	0.7611
	+ PLapDQ-max	0.0316	0.0925	0.0540	0.7491	0.7618	0.7802	**0.8065**
MovieLens 20m	Random	0.0004	0.0015	0.0008	0.7506	0.7505	0.7429	0.7430
	RankALS	0.0312	0.1467	0.0951	0.6001	0.5935	0.6239	0.6207
	+ MMR	0.0298	0.1401	**0.0897**	0.7170	0.7336	0.6731	0.6717
	+ LapDQ-min	**0.0305**	**0.1418**	0.0887	0.6451	0.6469	0.6522	0.6593
	+ LapDQ-max	0.0212	0.0887	0.0535	**0.7450**	**0.7481**	**0.7751**	**0.7880**
	+ DQ-min	0.0297	0.1378	0.0891	0.6163	0.6104	0.6348	0.6323
	+ DQ-max	0.0301	0.1391	0.0903	0.6132	0.6070	0.6337	0.6308
	+ PLapDQ-min	0.0218	0.0836	0.0520	0.6960	0.7089	0.7363	0.7484
	+ PLapDQ-max	0.0235	0.0953	0.0596	0.6998	0.7063	0.7157	0.7244

Table 4: Results on Precision@20 (Prec), Recall@20, nDCG@20, EILD@20 and EPD@20 on different data sets and different regularisers. For EILD and EPD metrics results without discount are present and with logarithmic discount model. ILD values have been calculated over all users and all differences are significant according to the paired t test. The best results for each metric across all of tested diversification methods are highlighted in bold.

0.8885 standard deviations smaller than the random algorithm. The `RankALS` is 3.1544 standard deviations smaller than then the random algorithm.

The diversity regulariser is a global regulariser, in so far as it seeks to maximise the average performance over the entire population. This means it does not necessarily improve the diversity of each individual user and it is possible that some users experience a decrease in diversity, in comparison to the non-diversified algorithm. It is therefore interesting to look at the impact of the method across users. For the `RankALS + LapDQ-min` on the Netflix data set we observe that 87% of the users have increased diversity over `RankALS` and 12% have decreased diversity. For the remaining 1%, there is no change. For the same algorithm but a smaller λ value – i.e. less diversification – we have 46% of the users experiencing an increase in diversity but 52% with decreased diversity compared to `RankALS`. This illustrates an issue with the approach of global diversification – that improved performance for some users can come at a cost of a reduction in performance for others.

Conclusions and Further Work

The research presented here aimed to continue and explore the work published in (Hurley 2013). A number of diversity regularisers have been proposed and evaluated, showing that it is possible to incorporate diversity into the training phase of a learning to rank algorithm for recommender systems. Of the proposed regularisers, the LapDQ regulariser showed the best performance among those compared. A number of short-comings of this approach can be identified. In particular, optimising for an global average improvement in diversity means that boosting the diversity for some users can mean a reduction in diversity for others. Ultimately, when the diversity term is strong enough, all users experience a diversity boost, but the accuracy generally deteriorates. Moreover, it may be better if diversification focused on the top candidate items, rather than across all items. Ways to address these short-comings may be directions for future work.

Acknowledgments

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

References

Bennett, J., and Lanning, S. 2007. The Netflix Prize. *KDD Cup and Workshop* 3–6.

Carbonell, J., and Goldstein, J. 1998. The Use of MMR, Diversity-Based Reranking for Reordering Documents and Producing Summaries. *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval* 335–336.

Ekstrand, M. D.; Harper, F. M.; Willemsen, M. C.; and Konstan, J. a. 2014. User perception of differences in recommender algorithms. *Proceedings of the 8th ACM Conference on Recommender systems - RecSys '14* 161–168.

Giorgi, E. E., and Bhattacharya, T. 2012. A Note on Two-Sample Tests for Comparing Intra-Individual Genetic Sequence Diversity between Populations. *Biometrics* 68(4):1323–1326.

Hu, Y.; Koren, Y.; and Volinsky, C. 2008. Collaborative filtering for implicit feedback datasets. In *Proceedings of the 2008 Eighth IEEE International Conference on Data Mining*, ICDM '08, 263–272. Washington, DC, USA: IEEE Computer Society.

Hurley, N. J. 2013. Personalised Ranking with Diversity. *Proceedings of the 7th ACM conference on Recommender systems - RecSys '13* 2(1):379–382.

Jahrer, M., and Töscher, A. 2012. Collaborative filtering ensemble for ranking. *Journal of Machine Learning Research - Proceedings Track* 18:153–167.

Jamali, M., and Ester, M. 2010. A Matrix Factorization Technique with Trust Propagation for Recommendation in Social Networks. *Proceedings of the fourth ACM conference on Recommender systems - RecSys '10* 135–142.

Pilászy, I.; Zibriczky, D.; and Tikk, D. 2010. Fast als-based matrix factorization for explicit and implicit feedback datasets. In *Proceedings of the fourth ACM conference on Recommender systems*, RecSys '10, 71–78. New York, NY, USA: ACM.

Takács, G., and Tikk, D. 2012. Alternating Least Squares for Personalized Ranking. *Proceedings of the 6th ACM conference on Recommender systems - RecSys '12* 83.

Vargas, S., and Castells, P. 2011. Rank and Relevance in Novelty and Diversity Metrics for Recommender Systems. *Proceedings of the 5th ACM conference on Recommender systems - RecSys '11* 109.

Vargas, S.; Castells, P.; and Vallet, D. 2011. Intent-oriented Diversity in Recommender Systems. *Proceedings of the 34th International ACM SIGIR Conference on Research and Development in Information Retrieval* 1211–1212.

Zhang, M., and Hurley, N. 2008. Avoiding monotony: Improving the diversity of recommendation lists. In *Proceedings of 2nd ACM International Conference on Recommender Systems*.

Ziegler, C.; McNee, S. M.; Konstan, J. A.; and Lausen, G. 2005. Improving recommendation lists through topic diversification. In *Proceedings of the 14th International Conference on World Wide Web*, 22–32.