Promoting Diversity in Recommendation by Entropy Regularizer

Lijing Qin, Xiaoyan Zhu
State Key Laboratory of Intelligent Technology and Systems
Tsinghua National Laboratory for Information Science and Technology
Dept. of Computer Science and Technology, Tsinghua University, Beijing, China
qinlijing@gmail.com; zxy-dcs@tsinghua.edu.cn

Abstract

We study the problem of diverse promoting recommendation task: selecting a subset of diverse items that can better predict a given user's preference. Recommendation techniques primarily based on user or item similarity can suffer from the risk that users cannot get expected information from the over-specified recommendation lists. In this paper, we propose an entropy regularizer to capture the notion of diversity. The entropy regularizer has good properties in that it satisfies monotonicity and submodularity, such that when we combine it with a modular rating set function, we get submodular objective function, which can be maximized approximately by efficient greedy algorithm, with provable constant factor guarantee of optimality. We apply our approach on the top-K prediction problem and evaluate its performance on MovieLens data set, which is a standard database containing movie rating data collected from a popular online movie recommender system. We compare our model with the state-of-the-art recommendation algorithms. Our experiments show that entropy regularizer effectively captures diversity and hence improves the performance of recommendation task.

1 Introduction

Recommender system has become an important information filtering tool to provide people with items (e.g. movie, online app) they appreciate [1, 4, 2, 18]. Different from another famous information filtering task search engine, one of the challenges of recommender system is that we do not exactly know what users needs. What recommender systems do is to anticipate users' interests according to existing data. On this point, many recommendation techniques [9, 10, 15, 23] rank the objects based on overlap between users's past activities, i.e., items are recommended to a user based on the similar items this user has selected and other users with similar patterns of selected items. This strategy however will lead to a large potential risk as the recommended lists become more and more narrowing popular or over-specified. Users may become bored with the low amount of information of recommendation lists. Taking movie recommendation for example, a user that has rated the movie Godfather five stars may obtain a recommendation list containing all the other Godfather sequels. Though the user probably appreciates the Godfather sequels, usually this recommended set is not satisfying, since the user can hardly get expected information.

To increase information of recommendation results, most previous proposals consider defining diversity by semantic information (including the genre or directors of a movie, the authors of a book, etc.). An intuitive idea is formulating distance between a pair of items by their attributes and maximizing the sum of distance of a set. However, these semantic diversification has two main drawbacks. First, there may be lack of such semantic information. Second, it maybe unreliable to define diversity based on given semantic information. For example, movies of the same directors/actors are not necessarily similar. In this paper we construct a novel *entropy regularizer* to capture the notion of diversity for a recommended subset. Entropy regularizer is different from existing diversity-promoting techniques in that it is not based on traditional pairwise distance between elements in subset, but defined on the feature space of the overall subset.

The entropy regularizer has several good properties. First, the entropy regularizer is maximized when the set of item-feature vectors are orthogonal, and is minimized when the vectors are linearly dependent, which intuitively captures diversity of a set of feature vectors. Second, the entropy regularizer satisfies non-decreasing monotonicity and submodularity, such that when combined with a modular rating function, we get a monotone submodular objective function. There exists efficient greedy approximation algorithm for monotone submodular function maximization, such that the recommendation set is theoretically guaranteed to be a constant-factor approximation of the optimal solution.

In this paper, we make the following contributions: (i) we formulate the diversity promoting recommendation task as a linear combination of the rating function and the entropy regularizer; (ii) we discuss the properties of the entropy regularizer, and provide theoretical proof; (iii) we construct greedy algorithm to maximize the objective set function subject to a cardinality constraint, and discuss the approximate rate; (iv) we conduct experiments on real dataset, and show that compared to state-of-the-art baselines, our approach significantly improves the performance of top-K prediction [6]. Our approach reaches a tradeoff between enhancing the the personalization of individual user intent and increasing the information of recommendation set.

2 Related Work

The importance of result diversification in recommendation systems has been recognized recently [8, 26]. There exist two main challenges to address the problem of diversity promoting recommendation. The first one is how to define diversity in objective function, and the other one is how to design effective and efficient algorithm to find the optimal solution.

The earlier work [27] proposed to define a similarity metric based on a taxonomy-based classification. The similarity is used to compute an intra-list similarity metric to determine the overall diversity of the recommended set. The authors provide a heuristic post-processing algorithm to increase the diversity of top-N recommendation list. As we discussed in the introduction, taxonomy information is not always available, even worse it is not always reliable to determine the similarity of two items. Several proposals [25, 24, 3] define diversity based on feature space of items. Typically, the diversity of a given set is defined as the sum of distance between items in the set. These proposals differ on the algorithms, which generally fall into two classes: greedy heuristics, where the recommendation list is constructed one-by-one by maximizing a given distance function at each step; refinement heuristics, where an initial relevant item set is first provided and then refined by a series of actions that forms the final recommendation set. While the objective functions of these proposals are set function, in that the objective seeks for a set which maximize the value of the objective set function, almost all heuristic algorithms can not guarantee a theoretical bound on the solution.

We argue that the separate consideration of definition of diversity and the design of algorithm is suboptimal, which is an obstacle to design a principled algorithm. In this paper, we construct diversity promoting regularizer using a carefully chosen entropy function based on feature space. The entropy regularizer, on one hand quantifies the expected value of information contained in the given set of items, which describes the notion of diversity in principle manner, on the other hand satisfies submodularity and monotonicity such that there exists a provable approximation algorithm with a bound that guarantees the found recommendation set is almost as good as the optimal solution.

Submodular fucntions have roots in economics, game theory, and computational optimization. Recently, submodular functions have started attracting attentions in many practical computer science fields, such as machine learning [20], data mining [12, 20] and natural language processing communities [14]. The most recent work is document summarization, where the authors studied a class of submoduar function to select sentences which meet the requirements of both representativeness and diversity. Different from previous work, we introduce entropy as the measurement of diversity. Our definition of diversity and our algorithm to find the best subset over all possible subsets can also be applied in other work such as topic diversification tasks in information retrieval.

3 Submodular Functions

In this section, we give some background on submodularity [7]. Submodularity can be viewed as a discrete analog to convexity [16] in continuous optimization. Like convexity functions, submodularity functions also arise in many applications, and lead to nice theory results and efficient optimization algorithms.

Suppose we are given a finite set of items $V = \{v_1, \ldots, v_n\}$, the ground set 2^V denotes the set of all subsets of V. A function $f : 2^V \to \mathbb{R}$ returns a real value for any subset $S \subseteq V$.

Definition 1. *A set function $f : 2^V \to \mathbb{R}$ is submodular if, for any subset $S, T \subseteq V$,*

$$f(S) + f(T) \geq f(S \cap T) + f(S \cup T).$$

Meanwhile, for any subset $S \subseteq T \subseteq V \backslash v$, the submodularity of a set function $f : 2^V \to \mathbb{R}$ is also equivalent to: $f(S + v) - f(S) \geq f(T + v) - f(T)$. There are mainly two types of Submodular functions: monotone and non-monotone. A submodular function is *monotone*, if for any subset $S \subseteq T \subseteq V$, $f(S) \leq f(T)$. In this paper, we will focus on the monotone submodular functions.

Definition 2. *A set function $f : 2^V \to \mathbb{R}$ is called modular if for any sub set $S, T \subseteq V$,*

$$f(S) + f(T) = f(S \cap T) + f(S \cup T).$$

Modular set functions, which also satisfy submodularity according to definition 1, are a class of simple set functions. By induction, we can get $f(S) = f(\emptyset) + \sum_{v \in S} f(\{v\})$. If we identify every subset of S with its incidence vector, modular functions will correspond to linear functions. In practice, modular set functions are less useful than the more general submodular ones, but in this paper, our rating function derived by probability matrix factorization is a modular set function.

4 Promoting Diversity in Recommendation

In this work, we consider the task of recommendation as choosing the best set of items for a user, coupled with the desire to choose as "diverse" items as possible. We formulate this task as a combinatorial optimization problem in the following.

Problem 1. *Suppose there are M distinct items, for $K > 0$, find a set $S \subseteq [M]$[1] such that*

$$\arg\max_{S:|S| \leq K} f(S) \triangleq R(S) + \lambda g(S), \tag{1}$$

where $R(S)$ measures the quality of recommendation set, $g(S)$ is the diversity promoting regularizer and $\lambda > 0$ is the regularization constant.

In the rest of this section, we construct $R(S)$ and $g(S)$ in principled way. We will also show that both $R(S)$ and $g(S)$ are submodular and non-decreasing. These properties enable us to design an efficient approximation algorithm for the optimization problem Eq. 1.

[1] $[M]$ is shorthand for $\{1, \ldots, M\}$.

4.1 Probablistic matrix factorization

Our construction for $R(S)$ and $g(S)$ will use deep properties of probabilistic matrix factorization (PMF) model [22]. Therefore, in this part, we briefly review the assumptions and learning algorithms of PMF. Let r_{ij} denote the rating of user i for item j, and $\mathbf{R} = \{r_{ij}\} \in \mathbb{R}^{N \times M}$ denote the rating matrix, where N is the number of users. PMF is a Bayesian method for modeling these ratings by deriving a low-rank approximation of the rating matrix:

$$\mathbf{R} = \mathbf{U}^T\mathbf{V} + \mathbf{E}, \tag{2}$$

where $\mathbf{U} \in \mathbb{R}^{D \times N}$ is the user preference matrix and $\mathbf{V} \in \mathbb{R}^{D \times M}$ denotes the item feature matrix. We can further write $\mathbf{U} = [\mathbf{u}_1, \ldots, \mathbf{u}_N]$ and $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_M]$. Now, for each $i \in [N]$, vector $\mathbf{u}_i \in \mathbb{R}^D$ represents the D-dimensional latent preference vector of user i, and for each $j \in [M]$, $\mathbf{v}_j \in \mathbb{R}^D$ denotes the D-dimensional latent feature vector of item j.

More specifically, PMF places Gaussian priors on $\mathbf{U}, \mathbf{V}$ and $\mathbf{E}$:

$$u_{ik} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_u^2), \quad v_{jk} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_v^2), \quad e_{ij} \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_e^2).$$

The rating r_{ij} of user i on item j is then modeled as:

$$r_{ij} = \mathbf{u}_i^T\mathbf{v}_j + e_{ij}, \quad \text{for all } (i,j) \in [N] \times [M]. \tag{3}$$

Given a collection of observations $\Phi \subseteq [N] \times [M]$, we can estimate $\mathbf{U}$ and $\mathbf{V}$ by maximizing the following joint likelihood using gradient descent algorithm:

$$\begin{aligned} \log \Pr(\mathbf{R}, \mathbf{U}, \mathbf{V}) = &-\frac{1}{2\sigma_e^2} \sum_{i,j \in \Phi} (r_{ij} - \mathbf{u}_i^T\mathbf{v}_j)^2 \\ &- \frac{1}{2\sigma_u^2} \sum_i \|\mathbf{u}_i\|_2^2 - \frac{1}{2\sigma_v^2} \sum_j \|\mathbf{v}_j\|_2^2 + C, \end{aligned}$$

where C is a constant that does not dependent on the parameters or hyper-parameters.

Predictive distribution of ratings

In real-world recommendation scenario, we often assume that the item feature matrix $\mathbf{V}$ is reliably estimated by learning a PMF model. But for each specific user, the point estimation of preference vector $\mathbf{u}$ is often not reliable, since there is no sufficient rating information for one user. Bayesian approach is a promising solution for this issue: we integrate out the user preference vector $\mathbf{u}$ to take into account the uncertainty of $\mathbf{u}$. In other words, without estimating $\mathbf{u}$, we can compute the conditional distribution of the ratings of a user, which is often termed as *predictive distribution* in Bayesian literature, based on the estimated item feature matrix $\mathbf{V}$ and the known ratings of the user.

Formally, for a given user, we denote the set of rated items of the user as $\Omega \subseteq [M]$ and the set of unrated items as $\bar{\Omega}$. Without loss of generality, consider a set of unrated items $S \subseteq \bar{\Omega}$, $\mathbf{r}_S = \{r_i : i \in S\}$ denotes the rating of items in S, we are interested in the predictive distribution $\Pr(\mathbf{r}_S|\mathbf{r}_\Omega, \mathbf{V})$, where $\mathbf{r}_\Omega = \{r_i : i \in \Omega\}$ denotes the known ratings.

We start from deriving the posterior distribution of ratings $\mathbf{r} \in \mathbb{R}^M$ for a particular user. Given the item feature matrix $\mathbf{V}$, we integrate out the user's preference vector $\mathbf{u}$, and then we have

$$\begin{aligned} \Pr(\mathbf{r}|\,\mathbf{V}) &= \int_{\mathbf{u}} \Pr(\mathbf{r}|\mathbf{u}, \mathbf{V}) \Pr(\mathbf{u}) d\mathbf{u} \\ &= \mathcal{N}(\mathbf{r}|0, \Sigma + \sigma_e^2\mathbf{I}), \end{aligned} \tag{4}$$

where the second line follows the self conjugacy property of Gaussian distribution, and $\Sigma \triangleq \sigma_u^2\mathbf{V}^T\mathbf{V}$ is the covariance matrix of items. Then, by Eq. 4, the joint distribution of $\mathbf{r}_S$ and $\mathbf{r}_\Omega$ is given by:

$$\begin{bmatrix} \mathbf{r}_S \\ \mathbf{r}_\Omega \end{bmatrix} |\,\mathbf{V} \sim \mathcal{N}\left(\begin{bmatrix} \mathbf{0} \\ \mathbf{0} \end{bmatrix}, \begin{bmatrix} \Sigma_{SS} + \sigma_e^2\mathbf{I} & \Sigma_{S\Omega} \\ \Sigma_{\Omega S} & \Sigma_{\Omega\Omega} + \sigma_e^2\mathbf{I} \end{bmatrix} \right),$$

where $\Sigma_{AB} \in \mathbb{R}^{|A| \times |B|}$ denotes sub-matrix of Σ that consists of rows and columns indexed by A and B respectively. The desired predictive distribution of $\mathbf{r}_S$ is given by:

$$\begin{aligned} &\mathbf{r}_S|\,\mathbf{r}_\Omega, \mathbf{V} \sim \mathcal{N}(\mu_S, \Sigma_S), \\ &\quad \mu_S = \Sigma_{S\Omega}(\Sigma_{\Omega\Omega} + \sigma_e^2\mathbf{I})^{-1}\mathbf{r}_\Omega, \\ &\quad \Sigma_S = \Sigma_{SS} + \sigma_e^2\mathbf{I} - \Sigma_{S\Omega}(\Sigma_{\Omega\Omega} + \sigma_e^2\mathbf{I})^{-1}\Sigma_{\Omega S}. \end{aligned} \tag{5}$$

4.2 Rating function

We define rating function $R(S)$ as the sum of offset expected rating of items in set S. Intuitively, $R(S)$ encourages to select items with high expected rating. Formally,

$$\begin{aligned} R(S) &\triangleq \sum_{\omega \in S} (\mathbb{E}[r_\omega|\mathbf{r}_\Omega, \mathbf{V}] - c) \\ &= \mathbf{1}^T\mu_S - |S|c \\ &= \mathbf{1}^T\Sigma_{S\Omega}(\Sigma_{\Omega\Omega} + \sigma_e^2\mathbf{I})^{-1}\mathbf{r}_\Omega - |S|c, \end{aligned} \tag{6}$$

where $c \leq \min_{\omega \in \bar{\Omega}} \mathbb{E}[r_\omega|\mathbf{r}_\Omega, \mathbf{V}]$, which is an offset constant that does not depend on S, can be computed in the preprocessing step. Here, the role of c is to ensure $R(\{\omega\}) \geq 0$ for all $\omega \in \bar{\Omega}$, and therefore make R itself a monotone nondecreasing set function. In addition, note that $R(S)$ exactly matches the deduction of definition 2, i.e. a modular set function can be rewritten as $f(S) = f(\emptyset) + \sum_{v \in S} f(\{v\})$, when every subset of S is identified by its incidence vector. Hence, the rating function $R(S)$ satisfies modularity. In this way, we have successfully constructed a rating function $R(S)$ that is both *modular* and *non-decreasing*.

4.3 Entropy regularizer

There is no unique way to define diversity of an item set S. Intuitively, the diversity promoting function should be high if the items in S are very dissimilar with each other and vice versa. Since each item j is represented by its feature vector $\mathbf{v}_j$, it is natural to quantify such similarity based on these feature vectors. Principally, for a set of vectors, they are most dissimilar if these vectors are orthogonal, and are most similar when they are linear dependent. In addition, we seek to find the most informative unrated items, which implies the criterion measuring the posterior uncertainty of the ratings. Following this insight, we tap into the notion of entropy of a set of random rating variables.

As we have shown before, the posterior of $\mathbf{r}_S$ is multivariate Gaussian distribution. The differential entropy of multivariate Gaussian random variables is a function of the determinant of the covariance matrix:

$$h(\mathbf{r}_S|\mathbf{r}_\Omega, \mathbf{V}) \triangleq \int_{\mathbf{r}_S} \log(p(\mathbf{r}_S|\mathbf{r}_\Omega, \mathbf{V}))p(\mathbf{r}_S|\mathbf{r}_\Omega, \mathbf{V})d\mathbf{r}_S \quad (7)$$
$$= \frac{1}{2}|S|\log(2\pi e) + \frac{1}{2}\log\det(\Sigma_S),$$

where Σ_S is given by Eq. 5.

We call the set function

$$g(S) \triangleq h(\mathbf{r}_S|\mathbf{r}_\Omega, \mathbf{V})$$

as *entropy regularizer*. From information theoretical viewpoint, our definition of the entropy regularizer $g(S)$ quantifies the uncertainty of the set of ratings $\mathbf{r}_S$ given observations of rated items $\mathbf{r}_\Omega$. If the items in S are dissimilar from each other, and meanwhile these items are dissimilar from rated items Ω, the uncertainty about the ratings of $\mathbf{r}_S$ should be high. From this perspective, our choice of $g(S)$ matches the intuition of a diversity promoting regularizer.

We can establish similar intuitions of $g(S)$ using a geometric interpretation. Recall that each item ω can be characterized by its item feature vector $\mathbf{v}_\omega$ in PMF model. Naturally, two arbitrary items ω_1, ω_2 are most dissimilar if their feature vectors are orthogonal $\mathbf{v}_{\omega_1}^T\mathbf{v}_{\omega_2} = 0$ and therefor implies a high diversity. In fact, we can show that $g(S)$ is maximized if the item feature vectors of items in S are orthogonal to each other and the feature vectors of rated items. Formally, we have the following result.

Lemma 1. *Without loss of generality, assume that spectral norm (maximum singular value) of $\mathbf{V}_S$ is bounded by $\|\mathbf{V}_S\| \le \theta$. Then, we have*

$$g(S) \le \frac{|S|}{2}\log(2\pi e) + \frac{|S|}{2}\log(\theta^2\sigma_u^2 + \sigma_e^2),$$

where the equality holds when $\mathbf{V}_S^T\mathbf{V}_S = \theta^2\mathbf{I}$ and $\mathbf{V}_S^T\mathbf{V}_\Omega = \mathbf{0}$.

We have shown that $g(S)$ is an appropriate notion of diversity. Next, we will show that $g(S)$ is both submodular and non-decreasing based on mild assumptions.

Proofs of all results are given in the appendix.

Submodularity. We can show that the entropy regularizer $g(S)$ is a submodular set function. By definition, it suffices to show that $g(\mathcal{A} \cup \{\omega\}) - g(\mathcal{A}) \ge g(\mathcal{B} \cup \{\omega\}) - g(\mathcal{B})$ for all $\mathcal{A} \subseteq \mathcal{B} \subseteq \Omega$. Using the definition of $g(S)$, this is equivalent to $h(r_\omega|\mathbf{r}_\mathcal{A}, \mathbf{r}_\Omega, \mathbf{V}) \ge h(r_\omega|\mathbf{r}_\mathcal{B}, \mathbf{r}_\Omega, \mathbf{V})$. The last inequality is the well-known "information never hurts" bound [17]. Hence, we have the following result.

Lemma 2. *The set function $g(S)$ is submodular.*

Monotonicity. In general, the differential entropy of set of random variables is not necessarily monotone. Fortunately, however, we can show $g(S)$ is monotone for most parameter settings of PMF used in practical scenario. Formally, we have the following result.

Lemma 3. *If $\sigma_e^2 \ge (2\pi e)^{-1}$, the set function $g(S)$ is monotone.*

Recall that σ_e^2 is the prior belief of variance of noise on the ratings r_{ij} as defined in Eq. 3. The typical parameter setting of σ_e is $\sigma_e^2 \leftarrow 1.0$ and is much larger than the required threshold $(2\pi e)^{-1} \approx 0.0586$ [22]. Therefore, $g(S)$ is a monotone set function under the usual settings of σ_e.

5 Algorithm

The optimization problem of Eq. 1 is NP-hard [13]. In this section, we describe an efficient approximation algorithm with a provable guarantee by utilizing the submodular and non-decreasing properties of $f(S)$ constructed in the previous section.

Our algorithm is greedy, in that it chooses items in sequence and selects the next item which provides the maximum increase in $f(S)$. Formally, suppose we have already chosen a set of items S where $|S| < K$, our goal is to greedily select next item ω which maximizes:

$$\begin{aligned} &f(S \cup \{\omega\}) - f(S) \\ =&R(S \cup \{\omega\}) - R(S) + \lambda(g(S \cup \{\omega\}) - g(S)) \\ =&\mathbb{E}[r_\omega|\mathbf{r}_\Omega, \mathbf{V}] + \lambda h(r_\omega|\mathbf{r}_S, \mathbf{r}_\Omega, \mathbf{V}). \end{aligned}$$

Using the results of predictive distribution of r_ω in Eq. 5, Algorithm 1 shows our approximation algorithm for solving optimization problem Eq. 1.

```
Input: item feature matrix $\mathbf{V}$; existing ratings $\mathbf{r}_\Omega$.
Output: item selection $S$.
$S \leftarrow \emptyset$, $\mathcal{A} \leftarrow \Omega$;
$\mathbf{D} \leftarrow (\Sigma_{\Omega\Omega} + \sigma_e^2\mathbf{I})^{-1}$;
for $j = 1$ to $K$ do
    $\mathbf{C} \leftarrow (\Sigma_{\mathcal{A}\mathcal{A}} + \sigma_e^2\mathbf{I})^{-1}$;
    for $\omega \in [M]\backslash(S \cup \Omega)$ do
        $\delta_\omega^{(g)} \leftarrow \frac{1}{2}\log(2\pi e(\sigma_e^2 + \Sigma_{\omega\mathcal{A}}\mathbf{C}\Sigma_{\mathcal{A}\omega}))$;
        $\delta_\omega^{(R)} \leftarrow \Sigma_{\omega\Omega}\mathbf{D}\mathbf{r}_\Omega$;
    end
    $\omega^* \leftarrow \underset{\omega \in [M]\backslash(S \cup \Omega)}{\arg\max} \ \delta_\omega^{(R)} + \lambda\delta_\omega^{(g)}$;
    $S \leftarrow S \cup \{\omega^*\}$, $\mathcal{A} \leftarrow A \cup \{\omega^*\}$;
end
```

Algorithm 1: Approximation algorithm for Eq. 1

Complexity analysis. The time complexity of Algorithm 1 is $O(K(|\Omega| + K)^3)$, which is acceptable in real-world settings, since the number of rated items of a single user $|\Omega|$ is often very limited.

Approximation ratio. Combining the fact that $g(S)$ is submodular and $R(S)$ is modular, we have the objective function $f(S)$ is also submodular by the closure property of submodularity. Nemhauser et al.[21] proved that for maximizing a nondecreasing submodular set function, the performance guarantee of greedy algorithm is $(1-1/e)OPT$, where OPT is the value of optimal solution of Eq. 1. Nemhauseretal's theorem can be stated formally as following.

Theorem 1. *Under the assumption of Lemma 3, Algorithm 1 is guaranteed to find a set S of k items satisfying*

$$f(S) \ge (1 - 1/e)OPT,$$

where $OPT = \max_{|S^*| \leq K} f(S^*)$ is the value of optimal solution.

6 Evaluation

In this section, we conduct several experiments to evaluate the quality of our diversity promoting recommendation.

6.1 Dataset

The MovieLens dataset consists of 1,000,209 ratings for 3900 movies by 6040 users of homonym online movie recommender service [19]. All users are selected randomly, and each of them has rated at least 20 movies. The element of the dataset is represented by a tuple: $t_{i,j} = (u_i, v_j, r_{i,j})$, where u_i denotes userID, v_j denotes movieID, and $r_{i,j}$, which is an integer score between 1 and 5, denotes the rating of user i for movie j (higher score indicates higher preference).

6.2 Performance of top-K prediction

Following the work [25], we split the dataset into a training dataset and a test dataset. The training dataset was used to train the PMF model, such that we can get the item feature matrix. For each user profile P_u in test dataset, we further select $0.5|P_u|$ movies at random as *profile test set* T_u, where $|P_u|$ denotes the size of the profile. Then the profile test set T_u, i.e. 50% of the user profile, is removed. The reminder of profile P_u is used to make a top-K prediction for user u.

Metrics

We use precision to measure the quality of our proposed approach. Suppose we recommend K sized set S_u to user u, the precision is defined as:

$$precision \triangleq \sum_{u \in U} \frac{|T_u \cap S_u|}{N|P_u|}, \tag{8}$$

where U denotes the set of all tested user.

Note that we do not aim to predict the ratings of movie, but to make a more satisfying top-K recommendation [11]. Hence, precision is more appropriate metric rather than the the Mean Absolute Error (MAE) metrics.

Baselines

In order to show the performance improvement of our approach, we compare our approach with the following baselines:

1. PMF: Standard probabilistic matrix factorization without diversity promoting entropy regularizer. PMF is considered as the state-of-the-art recommendation technique. Specifically, we seek to find a set satisfying:

$$S \in \operatorname*{arg\,max}_{S \subset [M], |S| \leq K} \left[\sum_{\mathbf{v} \in V_S} \mathbf{u}_i^T \mathbf{v} \right],$$

which is equal to maximizing the rating function defined in Eq. 6.

2. PMF+MMR: Maximal marginal relevance [5] is proposed to reduce redundancy while maintaining relevance in the field of document summarization. More specifically, MMR sequentially chooses item using the following criteria

$$\omega_i^* = \operatorname*{arg\,max}_{\omega_i \subseteq [M] \setminus \hat{S}} \left[r_{\omega_i} - \lambda_{\text{MMR}} \max_{\omega_j \in \hat{S}} sim(\omega_i, \omega_j) \right],$$

where r_{ω_i} is the estimated rating of item ${\omega_i}^2$, $\hat{S}$ is the subset of items already selected, λ is a parameter specifying the trade-off between rating score and diversity, and $sim(\cdot, \cdot)$ is similarity measure between two items. Without loss of generality, based on the result of PMF model, we specify $sim(\cdot, \cdot)$ as the cosine similarity between the item feature vectors. The MMR algorithm intuitively selects an item that maximizes the MMR objective function until a given cardinality constraint is met, i.e. $|\hat{S}| \leq K$. Note that though MMR is widely used in information retrieval and document summarization, this heuristic method is not theoretically justified.

Results

We consider predicting different amount of movies, i.e. we recommend top-$5, 10, 15, 20, 25, 30$ movies to each user. For the parameter setting, when training the PMF model, we set the size of the latent space $D = 80/100/120$. Both of the baselines and our approach take the same setting. In addition, $\lambda = 0.15$ and $\lambda_{\text{MMR}} = 0.15$. In the next section, we will discuss the impact of parameter λ. The experimental results are shown in Table 1, where our approach is denoted as "PMF+MMR". As is obvious from the table, "PMF" provides the worst performance, "PMF+MMR" performances better, and our approach significantly outperforms both of them. This results indicate that diversity promoting methods (MMR and ER) help predict user preference, since users may like dissimilar movies that standard PMF cannot cover. Furthermore, compared with popular diversity promoting algorithm MMR, our entropy regularizer is more effective for capturing the notion of diversity. Moreover, for individual case of top-K prediction, where $K = 5/10/15/20/25/30$, "PMF+ER" on average improves the precision by $1.77\%/2.60\%/3.47\%/4.04\%/4.97\%/5.51\%$ relative to "PMF". This results indicates that as K increases, the diversity promoting regularizer plays more important role in predicting user preference.

6.3 Impact of parameter λ

In our model, the parameter λ balances rating and diversity. If $\lambda = 0$, the model seeks to find the movies with the highest rating, and if $\lambda = +\infty$, the model seeks to find the most diverse movies. Figure 1 shows the impacts of λ on precision. We also plots the precision by baseline PMF as a control. Clearly, in every case of top-K prediction (i.e. $K \in \{5, 10, 20, 30\}$), PMF+ER outperforms PMF significantly. More remarkable, as λ increases, while the performance of PMF stay constant, the prediction precision by PMF+ER increases at first, but

[2] Standard MMR use a similarity measure between a query and a candidate item to capture the notion of relevance. There is no explicit query for recommendation task, such that we consider replacing it with the estimated rating of an item.

Table 1: Precision comparison with baselines

K	$D = 80$			$D = 100$			$D = 120$		
	PMF	PMF+MMR	PMF+ER	PMF	PMF+MMR	PMF+ER	PMF	PMF+MMR	PMF+ER
5	0.436755	0.439073	**0.441060**	0.446358	0.449007	**0.460596**	0.427152	0.423510	**0.447351**
10	0.384437	0.382616	**0.392219**	0.390894	0.389735	**0.401821**	0.387252	0.388079	**0.397517**
15	0.348013	0.347461	**0.361700**	0.356402	0.357064	**0.368874**	0.354084	0.355629	**0.366556**
20	0.324421	0.323675	**0.336424**	0.332699	0.334851	**0.345116**	0.328477	0.328974	**0.341391**
25	0.304768	0.306490	**0.321523**	0.315298	0.317219	**0.330530**	0.308808	0.310265	**0.323113**
30	0.290728	0.291115	**0.306015**	0.300993	0.302539	**0.315894**	0.293322	0.295419	**0.308223**

when the value of λ surpass a certain threshold, the precision begins to decrease. This phenomenon demonstrates that there is a tradeoff between enhancing the the personalization of users and increasing the information of recommendation set. We also see that in most cases, when $\lambda \approx 0.15$, PMF+ER can achieve the best performance. Except the cases $K = 5$ and $K = 10$, where smaller value of λ is better. This is reasonable, since when we recommend less movies, users usually prefer movies more relevant, but when the amount of recommended items increase, the requirement of diverse begin to emerge.

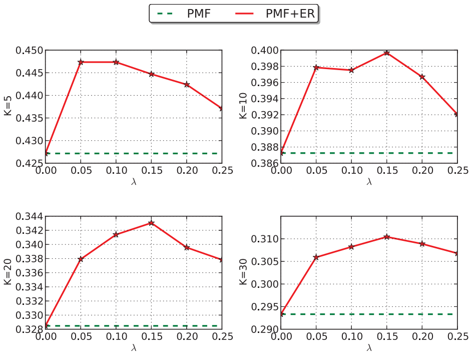

Figure 1: Impact of Parameter λ

7 Conclusion

In this paper, we (i) propose an entropy regularizer to capture the notion of diversity in recommendation, (ii)show the objective set function combined by the rating set function and the entropy regularizer satisfies submodularity, (iii) discuss the constraint by which the entropy regularizer will meet monotonicity, (iv) provide an approximation algorithm with $(1-1/e)$ theoretical bound. Our empirical results indicate entropy regularizer performs better than popular diversity promoting algorithm MMR, and improves the precision and recall of top-N prediction recommendation task.

Acknowledgments

This work is supported by China 973 Project No.2012CB316301.

Appendix

Proof of Lemma 1. By the defintion of $g(S)$, it suffices to show

$$\det(\Sigma_{SS} + \sigma_e^2 \mathbf{I} - \Sigma_{S\Omega}(\Sigma_{\Omega\Omega} + \sigma_e^2 \mathbf{I})^{-1}\Sigma_{\Omega S}) \leq (\sigma_u^2 + \sigma_e^2)^{|S|}.$$

At this point, it is easy to check the equality holds if $\mathbf{V}_S^T \mathbf{V}_S = \mathbf{I}$ and $\mathbf{V}_S^T \mathbf{V}_\Omega = \mathbf{0}$ and therefore $\Sigma_{SS} = \sigma_u^2 \mathbf{I}$ and $\Sigma_{S\Omega} = \mathbf{0}$ in this case. Now we upper bound the determinant by its spectral norm. Denote $\mathbf{E} = \Sigma_{SS} + \sigma_e^2 \mathbf{I} - \Sigma_{S\Omega}(\Sigma_{\Omega\Omega} + \sigma_e^2 \mathbf{I})^{-1}\Sigma_{\Omega S}$. We have $\det(\mathbf{E}) \leq \lambda_{\text{max}}(\mathbf{E})^{|S|}$, where $\lambda_1(\mathbf{E})$ is the maximum eigenvalue of $\mathbf{E}$. We can bound $\lambda_1(\mathbf{E})$ by following

$$\begin{aligned} \lambda_1(\mathbf{E}) &= \max_{\|\mathbf{x}\|_2 = 1} \mathbf{x}^T \mathbf{E} x \\ &= \mathbf{x}^T \left[\Sigma_{SS} + \sigma_e^2 \mathbf{I} - \Sigma_{S\Omega}(\Sigma_{\Omega\Omega} + \sigma_e^2 \mathbf{I})^{-1}\Sigma_{\Omega S}\right] \mathbf{x} \\ &\leq \mathbf{x}^T \Sigma_{SS} \mathbf{x} + \sigma_e^2 \mathbf{x}^T \mathbf{x} \leq \sigma_u^2 \|\mathbf{V}_S \mathbf{x}\|_2^2 + \sigma_e^2 \|\mathbf{x}\|_2^2 \\ &\leq \theta^2 \sigma_u^2 + \sigma_e^2, \end{aligned}$$

where the last inequality holds by the assumption on the spectral norm of $\|\mathbf{V}_S\| \leq \theta$. Combining the facts, it follows immediately that

$$\det(\mathbf{E}) \leq (\theta^2 \sigma_u^2 + \sigma_e^2)^{|S|}.$$

□

Proof of Lemma 3. By definition of g and the property of Gaussian distribution, we have

$$g(S \cup \{\omega\}) - g(S) = \frac{1}{2}\log(2\pi e \sigma^2_{r_\omega|\mathcal{A}}),$$

where we have denoted $\mathcal{A} = S \cup \Omega$. Since r_ω is conditionally Gaussian distributed, we have

$$\sigma^2_{r_\omega|\mathcal{A}} = \sigma_e^2 + \Sigma_{\omega\omega} - \Sigma_{\omega\mathcal{A}}(\Sigma_{\mathcal{A}\mathcal{A}} + \sigma_e^2 \mathbf{I})^{-1}\Sigma_{\mathcal{A}\omega} \geq \sigma_e^2,$$

where the inequality holds since $\Sigma_{\omega\omega} - \Sigma_{\omega\mathcal{A}}(\Sigma_{\mathcal{A}\mathcal{A}} + \sigma_e^2 \mathbf{I})^{-1}\Sigma_{\mathcal{A}\omega}$ is the Schur complement of positivie definitive matrix

$$\begin{bmatrix} \Sigma_{\omega\omega} & \Sigma_{\omega\mathcal{A}} \\ \Sigma_{\mathcal{A}\omega} & \Sigma_{\mathcal{A}\mathcal{A}} + \sigma_e^2 I \end{bmatrix}$$

and is therefore nonnegative.

Hence $g(S \cup \{\omega\}) - g(S) > 0$ since $\sigma_e^2 > (2\pi e)^{-1}$ by assumption.

□

References

[1] G. Adomavicius and A. Tuzhilin. Toward the next generation of recommender systems: A survey of the state-of-the-art and possible extensions. *Knowledge and Data Engineering, IEEE Transactions on*, 17(6):734–749, 2005.

[2] P. Bedi, H. Kaur, and S. Marwaha. Trust based recommender system for semantic web. In *proceedings of the 2007 International Joint Conferences on Artificial Intelligence*, pages 2677–2682, 2007.

[3] R. Boim, T. Milo, and S. Novgorodov. Diversification and refinement in collaborative filtering recommender. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 739–744. ACM, 2011.

[4] R. Burke. Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370, 2002.

[5] 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*, pages 335–336. ACM, 1998.

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

[7] S. Fujishige. *Submodular functions and optimization*, volume 58. Elsevier Science Limited, 2005.

[8] M. Ge, C. Delgado-Battenfeld, and D. Jannach. Beyond accuracy: evaluating recommender systems by coverage and serendipity. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 257–260. ACM, 2010.

[9] T. Hofmann. Latent semantic models for collaborative filtering. *ACM Transactions on Information Systems (TOIS)*, 22(1):89–115, 2004.

[10] T. Hofmann. Collaborative ltering via gaussian probabilistic latent semantic analysis. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 39–46. ACM, 2007.

[11] G. Karypis. Evaluation of item-based top-n recommendation algorithms. Technical report, DTIC Document, 2000.

[12] D. Kempe, J. Kleinberg, and É. Tardos. Maximizing the spread of influence through a social network. In *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 137–146. ACM, 2003.

[13] C.W. Ko, J. Lee, and M. Queyranne. An exact algorithm for maximum entropy sampling. *Operations Research*, 43(4):684–691, 1995.

[14] H. Lin and J. Bilmes. A class of submodular functions for document summarization. In *The 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL-HLT), Portland, OR, June*, 2011.

[15] G. Linden, B. Smith, and J. York. Amazon. com recommendations: Item-to-item collaborative filtering. *Internet Computing, IEEE*, 7(1):76–80, 2003.

[16] L. Lovász et al. Submodular functions and convexity. *Mathematical programming: the state of the art*, pages 235–257, 1983.

[17] L. Lovász et al. Submodular functions and convexity. *Mathematical programming: the state of the art*, pages 235–257, 1983.

[18] P. Massa and P. Avesani. Trust-aware collaborative filtering for recommender systems. *On the Move to Meaningful Internet Systems 2004: CoopIS, DOA, and ODBASE*, pages 492–508, 2004.

[19] MovieLens dataset. In *http://movielens.org*.

[20] M. Narasimhan and J. Bilmes. Local search for balanced submodular clusterings. In *Twentieth International Joint Conference on Artificial Intelligence (IJCAI07), Hyderabad, India, January*, 2007.

[21] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functionsi. *Mathematical Programming*, 14(1):265–294, 1978.

[22] R. Salakhutdinov and A. Mnih. Probabilistic matrix factorization. *Advances in neural information processing systems*, 20:1257–1264, 2008.

[23] J. Wang, A.P. De Vries, and M.J.T. Reinders. Unifying user-based and item-based collaborative filtering approaches by similarity fusion. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 501–508. ACM, 2006.

[24] C. Yu, L. Lakshmanan, and S. Amer-Yahia. It takes variety to make a world: diversification in recommender systems. In *Proceedings of the 12th International Conference on Extending Database Technology: Advances in Database Technology*, pages 368–378. ACM, 2009.

[25] M. Zhang and N. Hurley. Avoiding monotony: improving the diversity of recommendation lists. In *Proceedings of the 2008 ACM conference on Recommender systems*, pages 123–130. ACM, 2008.

[26] T. Zhou, Z. Kuscsik, J.G. Liu, M. Medo, J.R. Wakeling, and Y.C. Zhang. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515, 2010.

[27] C.N. 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*, pages 22–32. ACM, 2005.