Mach Learn (2009) 74: 75–109
DOI 10.1007/s10994-008-5081-7

Latent grouping models for user preference prediction

Eerika Savia · Kai Puolamäki · Samuel Kaski

Received: 22 February 2007 / Revised: 12 December 2007 / Accepted: 5 August 2008 /
Published online: 3 September 2008

Abstract We tackle the problem of new users or documents in collaborative filtering. Generalization over users by grouping them into user groups is beneficial when a rating is to be predicted for a relatively new document having only few observed ratings. Analogously, generalization over documents improves predictions in the case of new users. We show that if either users and documents or both are new, two-way generalization becomes necessary. We demonstrate the benefits of grouping of users, grouping of documents, and two-way grouping, with artificial data and in two case studies with real data. We have introduced a probabilistic latent grouping model for predicting the relevance of a document to a user. The model assumes a latent group structure for both users and items. We compare the model against a state-of-the-art method, the User Rating Profile model, where only the users have a latent group structure. We compute the posterior of both models by Gibbs sampling. The Two-Way Model predicts relevance more accurately when the target consists of both new documents and new users. The reason is that generalization over documents becomes beneficial for new documents and at the same time generalization over users is needed for new users.

Keywords Collaborative filtering · Gibbs sampling · Graphical model · Latent topic model

Editor: Dan Roth.

E. Savia (✉) · S. Kaski
Adaptive Informatics Research Centre, Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, 02015 TKK, Finland
e-mail: eerika.savia@tkk.fi

S. Kaski
e-mail: samuel.kaski@tkk.fi

E. Savia · K. Puolamäki · S. Kaski
Helsinki Institute for Information Technology HIIT, Department of Information and Computer Science, Helsinki University of Technology, P.O. Box 5400, 02015 TKK, Finland

K. Puolamäki
e-mail: kai.puolamaki@tkk.fi

1 Introduction

1.1 Background

This paper addresses the task of predicting relevance values for user–item pairs based on a set of observed relevance judgments of users for the items. Especially, we study how to predict relevance when very few relevance judgments, or ratings, are known for each user or item. The models we discuss are generally applicable, but since our prototype application area has been information retrieval, we will refer to the items as documents.

Traditionally, user preferences have been predicted using so-called collaborative filtering methods, where the predictions are based on the opinions of similar-minded users. Collaborative filtering is needed when the task is to make personalized predictions, but there is not yet sufficient amount of data about the user's personal interests available. Then the only possibility is to generalize over users, for instance by grouping them into like-minded user groups.

The early collaborative filtering methods were memory-based; predictions were made by identifying a set of similar users, and using their preferences fetched from memory. See, for instance (Konstan et al. 1997; Shardanand and Maes 1995). Model-based approaches are justified by the poor scaling of the memory-based techniques. Combining a memory-based technique with a model-based part has been suggested (Yu et al. 2004) to avoid the problem of new users not having enough ratings for reliable predictions. Recent work includes probabilistic and information-theoretic models, for instance (Hofmann 2004; Jin and Si 2004; Wettig et al. 2003; Zitnick and Kanade 2004). An interesting family of models are the latent component models, which have been successfully used in document modeling but also in collaborative filtering (Blei et al. 2003; Blei and Jordan 2003; Erosheva et al. 2004; Hofmann 2004; Keller and Bengio 2004; Marlin 2004; Marlin and Zemel 2004; McCallum et al. 2004; Popescul et al. 2001; Pritchard et al. 2000; Rosen-Zvi et al. 2004; Si and Jin 2003; Yu et al. 2005a, 2005b). When applying these models to collaborative filtering, each user is assumed to belong to one or more latent user groups that explain her preferences.

1.2 Tackling the problem of new users and documents

As a collaborative filtering system has to rely on the past experiences of the users, it will have problems when assessing new documents not yet seen by most of the users. Making the collaborative filtering scheme item-based, that is, grouping items or documents instead of users, would in turn introduce the problem of new users that do not have ratings for many documents. To tackle this problem of either new users or documents we have proposed a model that generalizes both ways (Savia et al. 2005). We go one step further from the common probabilistic models which have a latent structure for the users, and introduce a similar latent structure for the documents as well. A similar two-way structure has been suggested by (Si and Jin 2003) with some technical differences that will be discussed in Sect. 2.3.

This paper is structured as follows. We first present the User Rating Profile Model (URP; Marlin 2004) from which we go on to introduce our Two-Way Model that generalizes URP by grouping both users and documents (Sect. 2). After that we discuss the differences and similarities of our model with other related models.

In our model and URP two major choices in model structure have been made differently. The first choice is whether to cluster only users or to cluster both users and documents. The

Table 1 Notation

Symbol	Description
u	User index
d	Document index
r	Binary relevance (relevant = 1, irrelevant = 0)
u^*	User group index (attitude in URP)
d^*	Document cluster index
N_U	Number of users
N_D	Number of documents
N	Number of triplets (u, d, r)
K_U	Number of user groups
K_D	Number of document clusters
$\mathcal{D}$	Observed data

second choice is whether to generate the users and documents or to treat them as covariates of the model. In this paper, we study the effects of these two choices on the prediction performance. In Sect. 3 we introduce one variant of each model to compare whether it is useful to design the model to be fully generative, or to see users and documents as given covariates of the model.

We describe the experimental setups and baseline models in Sect. 4. In Sect. 5 we demonstrate with clearly clustered toy data how the two structural choices make a difference in actual predictions. Finally, in Sect. 6 we show with two case studies with real-world data that the proposed method works as expected also in practice, in addition to the toy demonstrations. Since in the work of Marlin the variational URP model outperformed the other latent topic models, we only had to compare with it. We compared our model to both a URP model that groups the users and a document-based URP that groups the documents. We computed the posteriors of all three models by Gibbs sampling.

2 Models

We first introduce the User Rating Profile Model (URP; Marlin 2004) from which we extend to our Two-Way Model that groups both users and documents. Our main notations are summarized in Table 1.

2.1 User rating profile model

URP is a generative model which generates a binary rating r for a given (user, document) pair.[1] It was originally optimized with variational Bayesian methods (*variational URP*; Marlin 2004). We solved the model with Markov chain Monte Carlo sampling from the full posterior distribution (*User URP*). This is expected to improve estimates especially for the small numbers of known ratings in our applications. We estimate the posterior predictive

[1]Note that the model also allows multiple-valued ratings if the Bernoulli distribution is replaced with a multinomial.

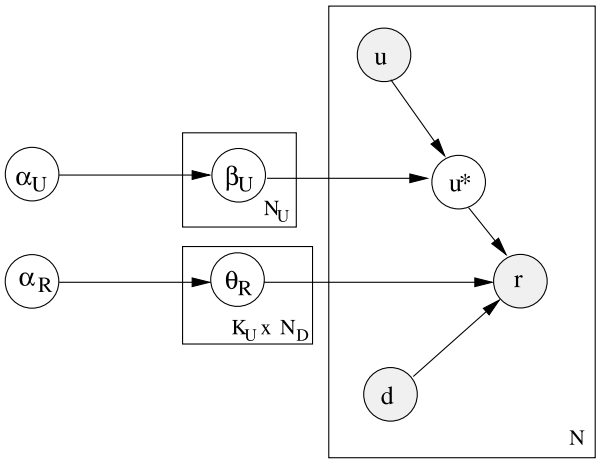

Fig. 1 Graphical model representation of the User Rating Profile model (URP). The *grey circles* indicate observed values. The boxes are "plates" representing replicates and the value at a corner of each plate indicates the number of replicates. The *rightmost* plate is repeated for each given (u, d) pair (altogether N pairs). The *upper left* plate represents the multinomial models of different users. The *lower left* plate represents the relevance probabilities of the different (user group, document) pairs

distribution $P(r|u, d, \mathcal{D})$ by Gibbs sampling; here $\mathcal{D}$ denotes the training data consisting of observations (u, d, r). The model assumes that there are a number of latent user groups whose preferences on the documents are different. The users belong to these groups probabilistically, into different groups in different instances. Alternatively, the groups can be interpreted as different "attitudes" of the user, and the attitude may be different for different documents. The generative process and the sampling formulas are presented in detail in the Appendix A, Sect. A.1. See Fig. 1 for a graphical model representation.

2.2 Two-way latent grouping model

We introduce a model that clusters users into user groups and documents into document clusters, in order to generalize relevance over both groupings. Each user may have several "attitudes," that is, belong to different groups during different relevance evaluations, and likewise each document may have several "aspects." These are modeled as probabilistic soft assignments.

The model generates rating triplets (user, document, rating), or (u, d, r), with binary relevances r. See Fig. 2 for a graphical model representation; the model is presented in detail in the Appendix A, Sect. A.4.

The obvious difference to URP is that the Two-Way Model extends URP by introducing the latent grouping also on the document side of the model. Another difference is the generation of users and documents. In URP the user group is generated for given (user, document) pairs whereas our model first generates the group and from the group a user/document is generated. In effect, this also changes the direction in which the users vs. user groups matrix $[\boldsymbol{\beta}_U]$ is normalized.

The Two-Way Model could easily be extended to handle multiple-valued ratings, by replacing the binary output with a multinomial. Binary responses, however, have the clear advantage that the ratings have a natural ordering: from the posterior we obtain simple probabilities varying continuously between the extremes, 0 and 1. In comparison, the multinomial distribution does not take the ordering of the ratings into account.

2.3 Other related models

The models most closely related to our Two-Way Model are the so-called latent topic models, especially the Flexible Mixture Model (FMM; Si and Jin 2003) but also Hofmann's

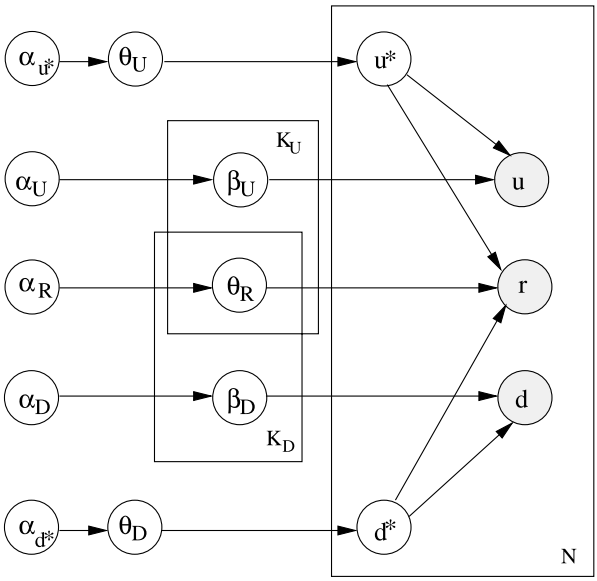

Fig. 2 Graphical model representation of our Two-Way Model. The *grey circles* indicate observed values. The *boxes* are "plates" representing replicates; the value in a corner of each plate is the number of replicates. The *rightmost* plate represents the repeated choice of N (user, document, rating) triplets. The plate labeled with K_U represents the different user groups, and $\boldsymbol{\beta}_U$ denotes the vector of multinomial parameters for each user group. The plate labeled with K_D represents the different document clusters, and $\boldsymbol{\beta}_D$ denotes the vector of multinomial parameters for each document cluster. In the intersection of these plates there is a Bernoulli-model for each of the $K_U \times K_D$ combinations of user group and document cluster

probabilistic latent semantic analysis (pLSA; Hofmann 2004), Latent Dirichlet Allocation (LDA; Blei et al. 2003), also known as multinomial PCA or mPCA (Buntine 2002), and the already introduced URP (Marlin 2004). Most of these models fall into the unifying model framework called Discrete PCA (Buntine and Jakulin 2006).

The main differences of the proposed Two-Way Model from the one-way grouping models pLSA, LDA, and URP, and from the two-way grouping FMM are discussed in this section.

The three one-way grouping models have subtle differences, but in all of them each user is assigned an individual multinomial distribution with parameters $\boldsymbol{\theta}$, and the latent user group u^* is sampled from this multinomial; the sampling is repeated for each document.[2] Each user can therefore belong to many groups with varying degrees. In our model, as well as in the FMM model, both *users* and *documents* can belong to many latent groups, much in the same way as users do in the three one-way models.

In pLSA and URP, each user group has a set of multinomials Mult($\boldsymbol{\beta}$), one for each document—these multinomials immediately determine the probabilities of ratings once the user group has been generated. Each mPCA user group, on the contrary, has a multinomial Mult($\boldsymbol{\beta}$) over the *documents* (Blei et al. 2003). Hence, mPCA could be seen as a model where the occurrence probabilities of documents are interpreted as probabilities of relevance. Thus, multinomial PCA can be interpreted as a binary relevance model, which cannot, as a side note, explicitly represent multiple-valued ratings. URP can be seen as an extension to mPCA with one extra dimension in the matrix of parameter vectors $\boldsymbol{\beta}$ to represent the different rating values. In our model and in the FMM model, the relevance is assumed to depend only on the latent groups, that is, there is a probability distribution of different ratings, Mult($\boldsymbol{\theta}_R$), for each (user group, document cluster) pair.

[2]Note that in text modeling this corresponds to sampling a "topic" Z for each word or token. In such a framework each document has a multinomial distribution Mult($\boldsymbol{\theta}$) over topics.

In addition to being two-way, our model and FMM differ from URP in that the users u and documents d are explicitly generated. In other words, the marginals $P(u)$ and $P(d)$ are estimated from data. In contrast, URP contains no generative process for the users or documents.

The difference between our two-way model and Flexible Mixture Model is that our model defines Dirichlet priors for all the multinomial model parameters, and is computed by sampling from the posterior distribution. FMM simply finds the easily overfitting maximum likelihood solution with the EM algorithm.

Finally, although our model is not far from the diverse set of so-called biclustering models (Madeira and Oliveira 2004; Tanay et al. 2006; Wettig et al. 2003), we aim at prediction instead of clustering, and therefore it is enough that the latent structure makes the predictions accurate. Because the model is computed by sampling, we do not obtain a single explicit set of clusters.

3 Choices in the model structure

In this section, we analyze the differences in model structure between the User Rating Profile Model (URP) and the Two-Way Latent Grouping Model (Two-Way Model). We introduce two variants of the models in order to study in Sect. 5 how the choices in model structure are reflected in the performance.

3.1 Main choices

There are two main differences between the model structures of URP and the Two-Way Model:

(i) One-Way or Two-Way Grouping
In URP, only users are clustered and there is no generalization over documents. An alternative is to use URP to cluster documents but then there is no generalization over users. In the Two-Way Model both users and documents are grouped for generalization.

(ii) Generation of Users and Documents from Marginals
In URP, the users and documents are assumed to be given and the user groups are generated from multinomials given the users. In the Two-Way Model, users and documents are generated from the user groups and document clusters, which in turn are generated from their marginal distributions. This difference also affects the normalization of the users vs. user groups matrix $[\boldsymbol{\beta}_U]$ containing the parameter vectors $\boldsymbol{\beta}_U$.

3.2 Model variants

First, we introduce a slightly different variant of URP, by introducing a mechanism for generating users and documents. This means we are making the model choice number (ii) in the same way it is done in the Two-Way Model. This implies changing the direction of normalization of the matrix $[\boldsymbol{\beta}_U]$ (user groups vs. users). We call this model *URP with Generation of Users and Documents* (*URP-GEN*). A graphical model representation is given in Fig. 3. A detailed description of the model including the generative process and the sampling formulas can be found in the Appendix A, Sect. A.2.

We expect this model to work better than the original URP because it takes the probabilities of different users appearing in the data into account. However, if the frequencies of users

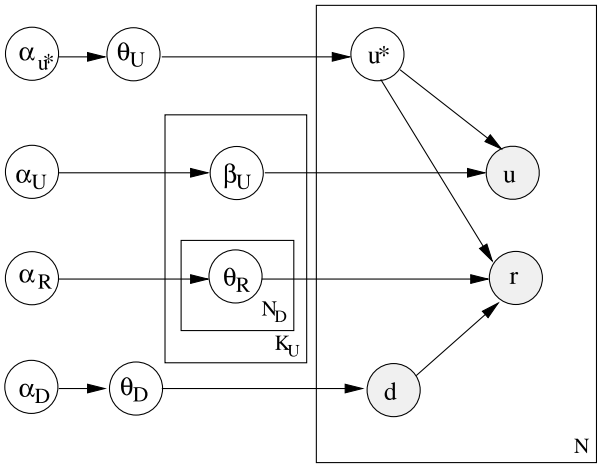

Fig. 3 A graphical model representation of User Rating Profile Model with generation of users and documents (URP-GEN). The *rightmost* plate represents the repeated choice of N (user, document, rating) triplets. The plate labeled with K_U represents the different user groups, and $\boldsymbol{\beta}_U$ denotes the vector of multinomial parameters for each user group. The plate labeled with N_D represents the documents. In the intersection of these plates there is a Bernoulli-model for each of the $K_U \times N_D$ combinations of user group and document. Since α_D and θ_D are conditionally independent of all other parameters given document d, they have no effect on the predictions of relevance $P(r \mid u, d)$ in this model. They only describe how documents d are assumed to be generated

as such are disinformative of the relevance, the original URP should be better. We study the validity of this assumption with toy data in Sect. 5.2.

Second, we introduce a slightly different variant of the Two-Way Model, by assuming the users and documents to be given as in URP. This means we are making the model choice number (ii) in the same way it is done in URP. Both the user groups u^* and the document clusters d^* are now generated from the multinomials of matrices $[\boldsymbol{\beta}_U]$ (users vs. user groups) and $[\boldsymbol{\beta}_D]$ (documents vs. document clusters), which implies changing the direction of normalization of the matrices. This means that the multinomials for user groups $\boldsymbol{\theta}_U$ and document clusters $\boldsymbol{\theta}_D$ become obsolete. We call this model Two-Way Grouping Model Without Generation of Users and Documents (Two-Way NO-GEN). A graphical model representation is given in Fig. 4. A detailed description of the model including the generative process and the sampling formulas can be found in the Appendix A, Sect. A.3.

We expect this model to generally perform worse than the original Two-Way Model because it does not take into account the probabilities of different users and documents appearing in the data.

3.3 Computational complexity

The number of observed (u, d, r) triplets is denoted by N and it dominates the complexities of all the presented models. The time complexity of computing one iteration of Gibbs sampling with the Two-Way Model is $\mathcal{O}(N(K_U^2 + K_D^2) + N_U K_U + N_D K_D + K_U K_D)$, whereas for User URP the complexity is $\mathcal{O}(N K_U^2 + N_U K_U + N_D K_U)$ per iteration, where $N_U =$ number of users, $N_D =$ number of documents, $K_U =$ number of user groups, and $K_D =$

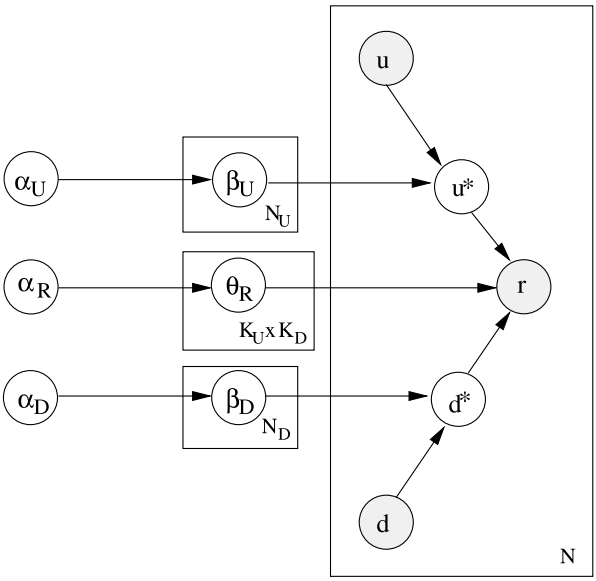

Fig. 4 A graphical model representation of the Two-Way Grouping Model without generation of users and documents (Two-Way NO-GEN). The *rightmost* plate represents the repeated choice of N (user, document, rating) triplets. The plate labeled with N_U represents the different users, and $\boldsymbol{\beta}_U$ denotes the vector of multinomial parameters for each user. The plate labeled with N_D represents the different documents, and $\boldsymbol{\beta}_D$ denotes the vector of multinomial parameters for each document. In the plate labeled with $K_U \times K_D$ there is a Bernoulli-model for each of the combinations of user group and document cluster

number of document clusters.[3] The complexity remains the same, regardless of whether one chooses to generate the users and documents.

In principle the complexity is too high for online computation. However, an approximate prediction for a new user or document can be made efficiently as follows: one could keep a set of randomly selected posterior samples and use it to represent all other users, while sampling only the parameters of the new incoming user.

4 Technical details of experiments

4.1 Evaluation of models by Gibbs sampling

The parameters of all the models were computed with Gibbs sampling. As a very brief tutorial, the model parameters are sampled one at a time, conditional on the current values of all the other parameters. One iteration step consists of sampling all the parameters once, in a fixed order. The observed data consists of triplets (u, d, r). For each iteration step m of sampling, we get a sample of all the parameters of the model, denoted by $\boldsymbol{\psi}^{(m)}$. Asymptotically, the sampled parameters $\boldsymbol{\psi}^{(m)}$ satisfy $\boldsymbol{\psi}^{(m)} \sim P(\boldsymbol{\psi} \mid \mathcal{D})$.

Each sample of parameters generates a matrix of probabilities $\mathrm{P}(r, u, d \mid \boldsymbol{\psi}^{(m)})$. The prediction of relevance, $\mathrm{P}(r \mid u, d, \boldsymbol{\psi}^{(m)})$, can be computed from these by the Bayes rule. As the final prediction we use the mean of the predictions over the M Gibbs iterations,

$$\begin{aligned} \mathrm{P}(r \mid u, d, \mathcal{D}) &= \mathbb{E}_{P(\boldsymbol{\psi}\mid\mathcal{D})}\left[\mathrm{P}(r \mid u, d, \boldsymbol{\psi})\right] \\ &\approx \frac{1}{M}\sum_{m=1}^{M} \mathrm{P}(r \mid u, d, \boldsymbol{\psi}^{(m)}). \end{aligned} \tag{1}$$

[3]The complexity of Document URP is analogous to User URP, where the role of users and documents is reversed, that is, $\mathcal{O}(NK_D^2 + N_D K_D + N_U K_D)$ per iteration.

4.2 Metropolis-Hastings sampling of priors

The Dirichlet priors of multinomials that generate user groups or document clusters, were sampled with the Metropolis-Hastings algorithm (Hastings 1970; Metropolis et al. 1953) with a flat prior in the interval [1, 10] and a Gaussian proposition distribution.

4.3 Monitoring of convergence

We sampled three MCMC chains in parallel and monitored the convergence of the predictions. First, each of the chains were run for 100 iterations of burn-in, with tempering like in (Koivisto 2004). After that, the burn-in period was continued for another 50 iterations without the tempering, to burn in the actual posterior distribution.

Convergence of the Markov chain Monte Carlo simulations was measured as follows. At the end of the burn-in period, the squared Hellinger distance H^2 between the chains (details below) was used as a convergence check: it was required to achieve the limit of 10^{-3}. The definition of squared Hellinger distance between two discrete probability distributions p and q is

$$\mathrm{H}^2(p,q) = \frac{1}{2}\sum_i \left(\sqrt{p_i} - \sqrt{q_i}\right)^2 = 1 - \sum_i \sqrt{p_i}\sqrt{q_i}, \tag{2}$$

where p_i and q_i denote the probabilities of the elementary events. We evaluated the squared Hellinger distance between the conditional distributions $P(r \mid u, d)$ produced by the 3 chains, pairwise between all chain pairs. The average of the Hellinger distances between the conditional distributions given by the chains measures the expected uncertainty in the prediction of relevance. Various figures of realized convergence of the chains are presented in Sect. 5.3.

After the burn-in period, each chain was run for another n iterations, and finally those $3 \times n$ samples were averaged to estimate expectations of $P(r \mid u, d)$. The number of needed iterations n depended on the data set and the complexity of the model.

4.4 Baseline models

We implemented three simple models to give baseline results. The *Naive Model* always predicts the same relevance value for r, according to the more frequent value in the training set.

The *Document Frequency Model* does not take into account differences between users or user groups at all. It simply models the probability of a document being relevant as the frequency of $r = 1$ in the training data for the document:

$$P(r = 1 \mid d) = \frac{\sum_u \#(u, d, r = 1)}{\sum_{u,r} \#(u, d, r)}. \tag{3}$$

The *User Frequency Model*, in its turn, does not take into account differences between documents or document groups. It simply models the probability of a user assigning documents relevant as the frequency of $r = 1$ in the training data for the user:

$$P(r = 1 \mid u) = \frac{\sum_d \#(u, d, r = 1)}{\sum_{d,r} \#(u, d, r)} \quad . \tag{4}$$

4.5 Experimental scenarios

In this section we introduce different types of experimental scenarios that were studied both with artificial and real data. The scenarios have various levels of difficulty for models that group only users, only documents, or that group both. The first two cases favor models that either group only users or only documents. The next two cases favor models that group both users and documents.

4.5.1 Only "new" documents

We constructed this experiment to correspond to the prediction of relevances for new documents in information retrieval. We took care that each of the randomly selected test documents had only few relevance judgments, in our case 3 ratings, in the training data. (Tests with other numbers of known ratings were run, too.) The rest of the ratings for these documents were left to the test set. For the rest of the documents, all the ratings were included in the training set. These documents represented the "older" documents in an information retrieval application; many users have already seen and revealed their opinion on them. This way all the tested documents are "new" in the sense that we only know 3 users' opinions of them. However, we are able to use "older" documents (for which users' opinions are already known) for training the user groups and document clusters.

We expected that this experiment should be hard for the URP models that group users, but on the other hand easy for the document-based URP that groups documents. The two-way grouping models' performance should lie between the user-based and document-based URP models.

4.5.2 Only "new" users

The experimental setting for new users was constructed in exactly the same way as the setting for new documents but the roles of users and documents were reversed.

We expected that this experiment should be hard for the document-based URP that groups documents but on the other hand easy for the user-based URP models that group users. Also in this case the two-way grouping models' performance should lie between the user-based and the document-based URP models.

4.5.3 Scenario where either user or document is "new"

In an even more general scenario either the users or the documents can be "new." We constructed a setting in which the test set consists of user-document pairs where either the user is "new" and the document is "old" or vice versa. Again, "new" meant having 3 ratings in the training set.

We expected that this scenario should bring out the need for two-way generalization; the two-way models were expected to perform better than URP models of any kind.

4.5.4 Scenario where both user and document are "new"

In this setting all the users and documents appearing in the test set were "new," having only 3 ratings in the training set. This case is similar to the previous setting, but much harder, even for the two-way grouping models.

We expected that this experiment should resemble the results of the previous case but with considerably worse levels of performance.

4.6 Measures of performance

For all the models, except the naive model, we used the log-likelihood of the test data set as a measure of performance, written in the form of perplexity,

$$\text{perplexity} = e^{-\frac{\mathcal{L}}{N}}, \quad \text{where } \mathcal{L} = \sum_{i=1}^{N} \log P(r_i \mid u_i, d_i, \mathcal{D}). \tag{5}$$

Here $\mathcal{D}$ denotes the training set data, and N is the size of the test set. Gibbs sampling gives an estimate for the table of relevance probabilities over all (u, d) pairs, $P(r \mid u, d, \mathcal{D})$, from which the likelihood of each test pair (u_i, d_i) can be estimated as $P(r_i \mid u_i, d_i, \mathcal{D})$. The best possible performance yields perplexity $= 1$ and binary random guessing (coin flipping) yields perplexity $= 2$. If perplexity is greater than 2 the model is performing worse than random guessing. Theoretically, perplexity can grow without a limit if the model predicts zero probability for some element in the test data set. However, we clipped the probabilities to the range $[e^{-10}, 1]$ implying maximum perplexity of $e^{10} \approx 22{,}000$.

We further computed the accuracy, that is, the fraction of the triplets in the test data set for which the prediction was correct. For the naive model, the prediction accuracy was the only performance measure used since, unlike the other models, it does not produce a probability of the relevance. For the other models we took the predicted relevance to be

$$\arg\max_{r \in \{0,1\}} P(r \mid u, d, \mathcal{D}), \tag{6}$$

where $P(r \mid u, d, \mathcal{D})$ is the probability of relevance given by (1). In all the experiments, statistical significance was tested with the Wilcoxon signed rank test.

5 Demonstrations with artificial data

We present two different demonstrations with artificial data to show the effect of choices in model selection. Table 2 summarizes the choices made in the models.

1. **Demonstration 1**. We demonstrate the performance of the different models in the task of finding biclusters when there are new users, new documents, or both. We expect all the topic models to succeed in this task but when there are new documents(/users/both) generalization over documents(/users/both) should be helpful.
2. **Demonstration 2**. We show how a misleading data set can favor a model which does not generate users and documents, while we normally would expect the models that do generate them to perform better.

Table 2 Summary of the models (**u** = user, **d** = document)

Model Abbreviation	Generates **u**, **d**	Groups **u**	Groups **d**
Two-Way Model	•	•	•
Two-Way NO-GEN	–	•	•
User URP	–	•	–
User URP-GEN	•	•	–
Variational URP	–	•	–
Document URP	–	–	•

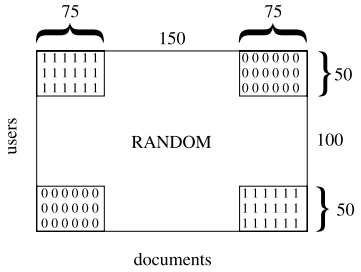

Fig. 5 *Focused biclusters* data. The matrix consists of 200 users and 300 documents and the ratings are missing for 70% of the (user, document) pairs. The corners of the matrix are clearly distinguishable biclusters and in the middle the ratings are uniformly random noise. Most of the data is in the corners: 2/3 of all the ratings are included in the corners while only 1/4 of the possible (user, document) pairs lie there. The density of observed ratings is six times as high in the corners as in the rest of the matrix

5.1 Demonstration 1: when do different generalizations help

We demonstrate the task of finding biclusters with an artificial dataset that has clear known cluster boundaries, sketched and explained in Fig. 5. By biclusters we mean a division of users and documents into subsets such that the users within each subset have similar ratings within each subset of documents. We expect that when there are new documents generalization over documents should be helpful, and analogously for new users. So, when there are both new users and new documents, two-way models should outperform URP models, both user-based and document-based.

We constructed four different settings as described in Sect. 4.5, namely

i) *Only New Documents Case*,
ii) *Only New Users Case*,
iii) *Either New User or New Document Case*, and
iv) *Both New User and New Document Case*.

We produced 10 artificial data sets of 18,000 ratings each, that followed the pattern of Fig. 5. All the models were trained with the known true numbers of clusters ($K_U = K_D = 3$). Details about the demonstration experiments can be found in Appendix B, Sect. B.2.

5.1.1 Results of demonstration 1

In the "New Documents" case, there is reason to believe that there would not be enough training data to learn the relevances correctly without generalization over documents. As expected, the Two-Way Model and Document URP performed better than the user-based URP models (see Table 3, column *New d*). The same phenomenon can be also seen between the baseline models. The Document Frequency Model clearly performs worse than the User Frequency Model in the "New Documents" case, since it tries to model each document individually, whereas the User Frequency Model generalizes over documents.

In the "New Users" case, there is reason to believe that there would not be enough training data to learn the relevances correctly without generalization over users. As expected, the Two-Way Model and the user-based URP models achieved better results than Document URP (see Table 3, column *New u*). Also, the roles of the baseline models have become interchanged.

Table 3 Perplexity of the various models in Demonstration 1. Smaller perplexity is better and 2.0 corresponds to random guessing. The best result of each column is underlined. The best result differs statistically significantly with P-value ≤ 0.01 from the second best one ($\mathbf{u}$ = user, $\mathbf{d}$ = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	1.52	1.54	1.53	1.70
Two-Way NO-GEN	1.91	1.88	1.89	1.97
User URP	1.69	1.68	1.68	1.88
User URP-GEN	1.68	1.58	1.62	1.83
Variational URP	6.74	2.23	3.50	10.7
Document URP	1.64	1.74	1.68	1.86
User Freq.	2.02	5.65	3.25	4.99
Document Freq.	5.29	2.01	3.21	5.92

In the "Either New" case, we expected the difference between one-way grouping and two-way grouping to become obvious. The difference is rather subtle, however, though statistically significant (see Table 3, column *Either New*).

Finally, in the "Both New" case we expected the two-way grouping to again improve results compared to one-way grouping URPs, but with the results being worse overall. This effect can be seen in Table 3, column *Both New*.

Variational URP[4] has the property of overestimating the confidence in its predictions, resulting in extreme probabilities near 0 or 1. When the prediction is incorrect, this is strongly penalized in the perplexity, and hence the performance of Variational URP is quite unstable when measured with perplexity. This is seen in Table 3.

It can also be noticed that the generation of users and documents seems useful with this kind of data; Two-Way Model has consistently lower perplexity than Two-Way NO-GEN (P-value $= 0.002$ in all cases), and User URP-GEN has lower perplexity than User URP (P-value $= 0.01$ in the case *New d*, P-value $= 0.002$ in other cases).

The prediction accuracy of the best model varied between 83–84%, while the prediction accuracy of the best baseline model varied between 50–52%. The accuracies can be found in the Appendix C, in Table 14, Sect. C.1.

5.1.2 Effect of the amount of rating information about new users and documents

The differences between models are naturally largest when a very small amount of information about "new" users and documents is available. When the amount of information is increased, more complex models are expected to gradually become better than models that try to generalize over many users and documents. With increasing amount of information, one is able to train a model for each user (or document) separately with less need for generalization. In Figs. 6–8 we show how the differences gradually change as a function of number of known ratings for "new" users/documents, in one of the datasets from demonstration 1. The number of known ratings for "new" users/documents was varied within $\{3, 5, 10, 20\}$. All the perplexity and accuracy values can be found in the Appendix C, Sect. C.2, Tables 17–22. As we here know the true model that was used to generate the data, we can also compare the results to the theoretically derived optimal perplexity (1.26) achieved with the true model, shown as a horizontal line in the figures.

[4]We only show the performance of Variational URP for the artificial data since our implementation is too inefficient for larger data sets.

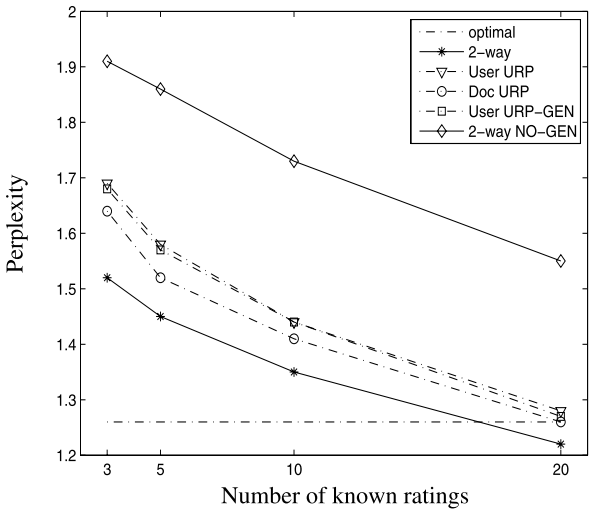

Fig. 6 Perplexity as a function of the amount of rating information about "new" documents in the "New Documents" case

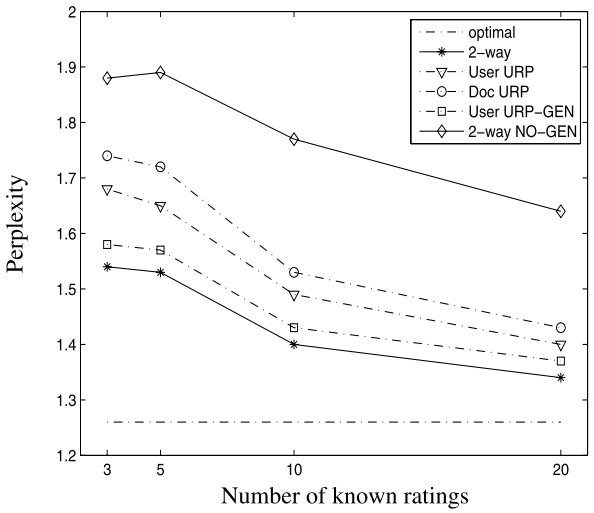

Fig. 7 Perplexity as a function of the amount of rating information about "new" users in the "New Users" case

5.2 Demonstration 2: when does generation of users/documents not help

We demonstrate that there are cases where generation of users and documents is not helpful because the data is misleading. For this, we use artificial data, which is intended to mislead the models which generate users/documents; a large proportion of the data lies in an area where the relevances are not predictable (random). The data is sketched and explained in Fig. 10. We apply the two user-based URP models, User URP and User URP-GEN, to demonstrate the effect of generating the marginals in this case.

We generated 10 artificial data sets of 18,000 ratings each, and in each set we randomly assigned one half of the data triplets to the test set and the rest to the training set. Both models were trained with the known true number of clusters ($K_U = 4$).

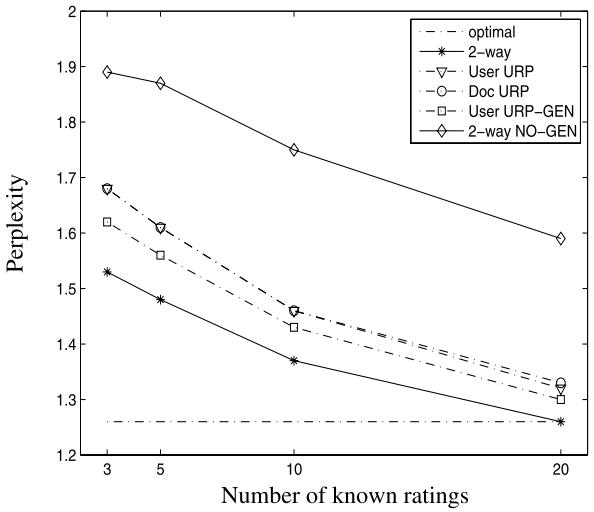

Fig. 8 Perplexity as a function of the amount of rating information about "new" users/documents in the "Either New" case

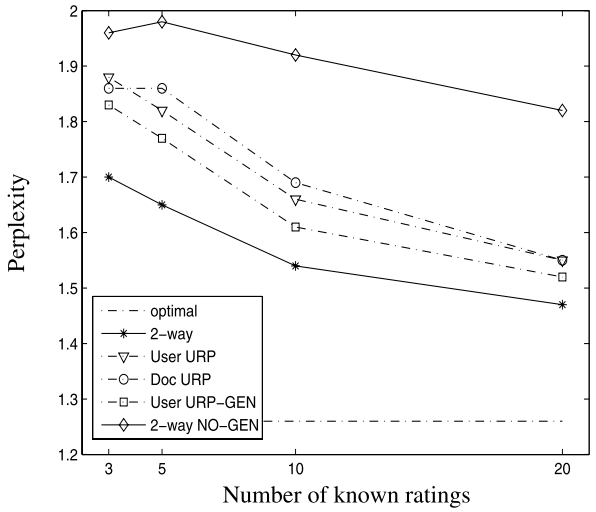

Fig. 9 Perplexity as a function of the amount of rating information about "new" users/documents in the "Both New" case

The difference between this data and the *Focused biclusters* type of data (Sect. 5.1) is that the density of (u, d) pairs is very different in the "clusterable" area; in a *Focused biclusters* type of data set, the relevant information about the biclusters lies in the same area where most of the data lies, whereas in a *Misleading biclusters* type of data set, the relevant information is in the area where the data is at its sparsest. With this type of data set, it makes a difference whether the model generates the users and documents from the marginals, because ignoring the generation essentially equals to assuming that all (u, d) pairs carry equal amount of information about the relevance r.

We expect the models with user/document generation to be misled by the large random clusters, and hence assume the non-generating models to improve their relative performance with *Misleading biclusters* data.

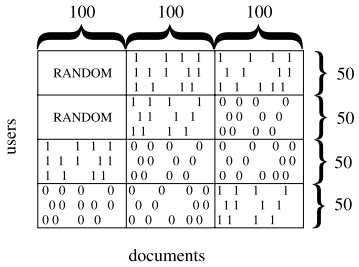

Fig. 10 *Misleading biclusters* data. The matrix consists of 200 users and 300 documents and the ratings are missing for 70% of the (user, document) pairs. The "random corner" misleads the models: 1/3 of all the ratings are in the random corner while only 1/6 of the possible (user, document) pairs lie there. Only 40% of the ratings are missing in the random corner while 76% of the ratings are missing in the rest of the data matrix

Table 4 Demonstration 2. The results of the two URP models differ from each other statistically significantly with P-value < 0.01. Small perplexity and large accuracy are better. The best result of each column is underlined (**u** = user, **d** = document)

Method	Perplexity	Accuracy %
User URP	<u>1.54</u>	<u>83</u>
User URP-GEN	1.61	78
User Frequency Model	1.93	63
Document Frequency Model	2.04	49
Naive Model	–	50

5.2.1 Results of demonstration 2

As expected, this misleading data caused the non-generative version of URP to outperform the generative one (Table 4). On the other hand, in normal circumstances, generation of users and documents should be beneficial, which is shown nicely in all the results of the first demonstration, in Table 3.

5.3 Convergence of sampling

As described in Sect. 4.3, we used a Hellinger distance based measure for monitoring the convergence of MCMC chains. Figure 11 shows an example of how the convergence measure behaved during sampling. These values are from a dataset in demonstration 1. In general, generation of users and documents improved convergence compared to the non-generating counterpart. Additionally, one-way grouping models converged faster than two-way grouping models.

In order to assess how fast the chains actually converged to the final results (as opposed to converging only to each other) we show in Fig. 12 how the perplexity of the test set evolved with increasing number of samples for the Two-way model. Each of the ten curves shows one independent sampling chain.

Since we average over parallel chains, we finally studied the effect of the number of chains in Fig. 13. As we can see, averaging over only three chains achieves almost the same result as averaging over ten chains in this case.

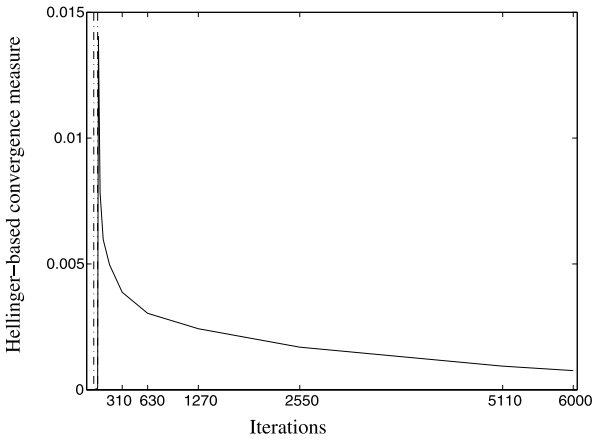

Fig. 11 Hellinger-based convergence measure between 3 chains as a function of sampling iterations for Two-Way Model with a dataset from demonstration 1. This measure was used for convergence monitoring. The *first vertical dotted line* shows when tempering was finished, and the *second vertical dotted line* shows when the burn-in period was finished and the gathering of samples was started. Smaller values correspond to better convergence

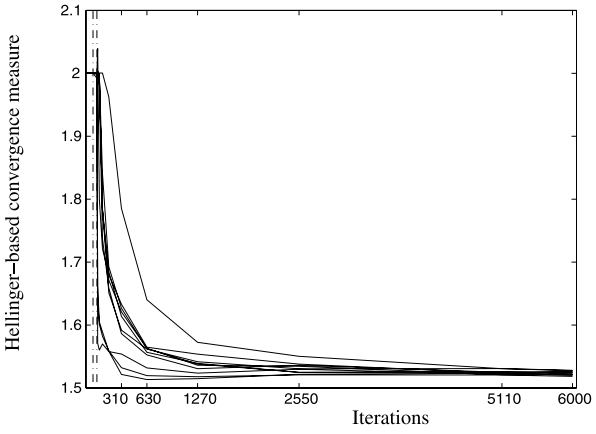

Fig. 12 Perplexity of the test set as a function of samples gathered from 10 independent chains. The *first vertical dotted line* shows when tempering was finished, and the *second vertical dotted line* shows when burn-in period was finished and the gathering of samples was started. Smaller perplexity is better

6 Case studies

In this chapter we show, with two case studies, how the Two-Way Model compares with the other models on real-world data, in the same way we compared them in Demonstration 1.

6.1 Experimental setup in case studies

To validate the parameters, we constructed the validation set and the initial training set for the validation phase, in the same way as described in Sect. 4.5. The new documents or users included in the final test set were not used in the validation phase. Details of the experimental setup can be found in the Appendix B, Sect. B.3.

The latent topic models were evaluated in the validation phase for a range of cluster numbers. The trained models were tested on the validation set, and the lowest perplexity was used as a performance criterion for choosing the cluster numbers. For the final results the models were trained with all the training data with the validated cluster numbers, and tested with the final test data set.

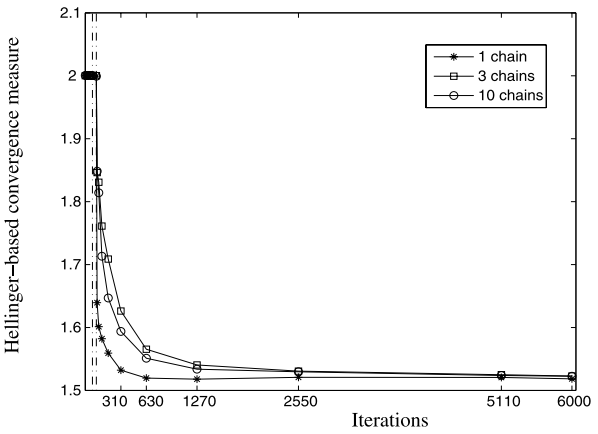

Fig. 13 Perplexity of the test set for averages over different numbers of parallel chains. The *first vertical dotted line* shows when tempering was finished, and the *second vertical dotted line* shows when burn-in period was finished and the gathering of samples was started. Smaller perplexity is better

Table 5 The validated cluster numbers in the parliament case study

Method	K_U	K_D
Two-Way Model	4	2
Two-Way NO-GEN	4	2
User URP	3	–
User URP-GEN	3	–
Document URP	–	2

6.2 Case study I: parliament data

We predicted votes of the British Parliament using a publicly available data set (British Parliament data 1997–2001). The data set consisted of 514,983 votes given by the members of parliament on 1,272 issues. We predicted the votes for previously unseen (user, issue) pairs, where the "users" are the members of the parliament (MP).

We selected the cluster numbers using a validation set described in more detail in the Appendix B, Sect. B.3. The choices from which the cluster numbers were selected were $K_U \in \{1, 2, 3, 4, 5, 10, 20, 50\}$ for the user groups and $K_D \in \{1, 2, 3, 4, 5, 10, 20\}$ for the document clusters. The selected cluster numbers are shown in Table 5. These values were used in all experimental scenarios.

6.2.1 Missing votes in parliament data

The parliament data does not contain a "yes" or "no" answer for all the (user, issue) pairs; about 40% of all possible votes are missing. The fact that a member of parliament has not voted "yes" or "no" on a particular issue may either be due to her not being present at the voting, or her tactical reasons not to take a stand in the matter. In either case, the vote is not missing at random. From the modeling perspective we could assign a "rating," say -1, to all the missing votes and make the sparse data matrix full, which would, however, notably increase the computational load. Fortunately, some of this information is effectively taken into account by modeling the user margins $\mathrm{P}(u \mid \mathcal{D})$ and the document margins $\mathrm{P}(d \mid \mathcal{D})$.

Table 6 Parliament Data. Comparison between the models by perplexity over the test set. In each column, the best model (underlined) differs statistically significantly from the second best one (P-value ≤ 0.01). Smaller perplexity is better; 2.0 corresponds to binary random guessing and 1.0 to perfect prediction (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	1.37	<u>1.40</u>	<u>1.38</u>	<u>1.62</u>
Two-Way NO-GEN	1.62	1.62	1.62	1.86
User URP	1.48	1.45	1.46	1.73
User URP-GEN	1.50	1.44	1.47	1.74
Document URP	<u>1.32</u>	1.53	1.42	1.67
User Freq.	2.00	5.68	3.32	4.78
Document Freq.	5.36	1.76	3.12	5.85

In collaborative filtering setting the missing at random assumption has been studied by Marlin et al. (2005, 2007), and it has been found that when users give ratings to music pieces the missing ratings have a different distribution than the ratings that were actually given.

6.2.2 Results with parliament data

The results are summarized in Table 6. The prediction accuracy of the best model varied between 86–95%, while the prediction accuracy of the best baseline model varied between 64–71%. The accuracies can be found in the Appendix C (Sect. C.1, Table 15). We discuss the findings together with the second case study in Sect. 6.4.

6.3 Case study II: scientific articles data

In our second case study we used data gathered in a controlled experiment, where the test subjects browsed through a set of titles of scientific articles and chose the most interesting ones via a web form (Puolamäki et al. 2005). The test subjects were shown 80 lists with six article titles each. The subjects participating in the experiment were researchers in either vision research, artificial intelligence, or machine learning. The browsed lists consisted of titles of scientific articles published during autumn 2004 in major journals in the fields of vision research, artificial intelligence, machine learning, and general science. On each page there was a randomly generated list of titles always containing titles from each discipline. On each page the subjects were to choose the two most interesting titles according to their own preferences. Altogether, the data consisted of 25 users' opinions on 480 titles of scientific articles ("documents").

Again, we selected the cluster numbers using a validation set described in more detail in the Appendix B, Sect. B.3. The choices from which the cluster numbers were selected for the scientific articles data were $K_U \in \{1, 2, 3, 4, 5, 10, 15, 20\}$ for the user groups and $K_D \in \{1, 2, 3, 4, 5, 10, 20, 50\}$ for the document clusters. The selected cluster numbers are shown in Table 7. These values were used in all experimental scenarios.

6.3.1 Results with scientific articles data

The results of comparing the perplexities of the models on test data set are summarized in Table 8. Since there was such a small number of users, we proceeded in leave-one-out fashion, making one user at a time the "new" user who had only 3 ratings in the training set, and the rest of her ratings in the test set. The results are averages over these 25 leave-one-out runs.

Table 7 The validated cluster numbers in the scientific articles case study

Method	K_U	K_D
Two-Way Model	5	3
Two-Way NO-GEN	5	3
User URP	2	–
User URP-GEN	2	–
Document URP	–	2

Table 8 Scientific articles data. Comparison between the models by perplexity. Average over 25 test set likelihoods, presented here in the form of perplexity. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value $\leq$0.01) are marked with boldface. Smaller perplexity is better; 2.0 corresponds to binary random guessing and 1.0 to perfect prediction (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	1.74	**1.73**	1.73	1.85
Two-Way NO-GEN	1.80	**1.77**	1.79	1.88
User URP	1.76	1.69	1.73	1.85
User URP-GEN	1.77	**1.70**	1.73	1.85
Document URP	1.76	1.89	1.82	1.92
User Freq.	1.89	5.39	3.18	4.84
Document Freq.	3.74	1.78	2.59	3.76

The averaged prediction accuracy of the best model varied between 67–74%, while the prediction accuracy of the best baseline model varied between 67–70%. The accuracies can be found in the Appendix C (Sect. C.1, Table 16).

6.4 Conclusions of the case studies

6.4.1 Generalization over "new" users or documents is needed

In the *Only New Documents Case* for the parliament case study, the document-based URP model performs best, as expected (see Table 6, column *New d*). The reason for the clearly worse performance of the user-based URP models can be explained with the fact that the number of "bins" is large compared to the number of training data samples. User-based one-way models do not generalize over documents, so for each test document they may have, for instance, $K_U = 3$ bins and 3 data samples. Thus, for each test document there are only 3 training samples to estimate the parameters of the K_U bins, which is generally not enough. In document-based URP the bins are distributed in the other direction, so each "new" document gets grouped with other similar documents. Therefore, there would only be problems in the case of new users for this kind of model.

In the *Only New Users Case* in both case studies, the user-based URP models (User URP and User URP-GEN) clearly perform better than the document-based URP (see Tables 6 and 8, column *New u*). The reason is the same as in the previous paragraph, only this time the experiment favors grouping of users.

In our Two-Way Model, there are K_U bins per document cluster, not per document, which makes the model more robust to variation caused by the small number of training data samples, and this works regardless of whether it is the documents or users that are "new."

6.4.2 Two-way generalization is needed when both users and documents are "new"

Our model is at least as good as one-way grouping models in the cases where we expected two-way generalization to be needed, namely columns *Either New* and *Both New* in Tables 6 and 8. In the parliament case study, the expected differences are shown clearly. In the scientific articles case, the user-based one-way models (User URP and User URP-GEN) reached the same level of performance.

7 Discussion

We have introduced a latent grouping framework which extends a state-of-the-art method, the User Rating Profile model (URP), by introducing a two-way latent grouping. Our Two-Way Model assumes that both the users and the documents have a latent group structure. We compared the model against the user-based URP (which groups users) and the document-based URP (which groups documents) on two real-world data sets. The predictions of all these models were computed with Gibbs sampling. In addition, we analyzed the structural choices in URP and our model and demonstrated the effects of the different choices with artificial data sets.

We compared the models in different types of tasks. In the case of *New Documents*, the task was to predict users' subjective relevances for new documents, with only very few existing ratings—still being able to utilize information about relevances of documents seen earlier by a mass of users. The task resembles collaborative filtering where relevance of a new document is to be predicted. In this task, the available information about the attitudes of users towards the new documents is very limited, and generalizing over similar documents proved to be beneficial, as was expected. Obviously, this kind of generalization is available in the document-based URP model, but the document group structure of our Two-Way Model enables such generalization, too. In the case of *New User*, the task was to predict relatively new users' subjective relevances for documents, with only very few existing ratings from this particular user. The task turns the roles of users and documents the other way around, and hence the user-based URP model outperformed the document-based URP in this case. For the Two-Way Model there is no difference in these situations, because the model is symmetric with respect to users and documents.

The case *New Users and Documents* demonstrated a task where prediction of new users' relevances for new documents makes generalization in both ways necessary. Our Two-Way Model gives better relevance predictions than either of the one-way URP models in this task.

Acknowledgements The authors would like to thank Dr. Jacob Goldberger for valuable discussions. This work was supported by the Academy of Finland, decision no. 79017, and by the IST Programme of the European Community, under the PASCAL Network of Excellence, IST-2002-506778. This publication only reflects the authors' views. Access rights to the data sets and other materials are restricted due to other commitments. This work was carried out in the Adaptive Informatics Research Centre, a Centre of Excellence of the Academy of Finland.

Appendix A: Generative process and sampling formulas of the models

In this appendix we give a detailed description of the generative models presented in this article, and the Gibbs sampling formulas for the posterior distributions for each variable relating to the user clusters. The formulas are analogous for document clusters. In our notation n denotes an index for the observed (user, document, rating) triplets ($n \in \{1, 2, \ldots, N\}$), $\mathcal{D}$ denotes all the observed data, and $\boldsymbol{\psi}$ denotes all the parameters of the model.

A.1 User URP with Gibbs

The generative process proceeds according to the following steps (see also Fig. 1 and summary of notation in Tables 1 and 9):

1) For each user u, a vector of multinomial parameters $\boldsymbol{\beta}_U(u)$ is drawn from Dirichlet($\boldsymbol{\alpha}_U$). This is denoted by the plate with N_U repetitions in the graphical model representation. Each parameter vector $\boldsymbol{\beta}_U(u)$ contains the probabilities for a user u to have different attitudes u^*, that is, to belong to different user groups u^*.
2) For each (user group, document) pair (u^*, d), a vector of Bernoulli parameters $\boldsymbol{\theta}_R(u^*, d)$ is drawn from Dirichlet($\boldsymbol{\alpha}_R(u^*, d)$). This is denoted by the plate with $K_U \times N_D$ repetitions in the graphical model representation. Each parameter vector $\boldsymbol{\theta}_R(u^*, d)$ defines the probability of the user group u^* to consider document d relevant (or irrelevant).

The rest of the steps are repeated for each of the N rating triplets:

3) From the given set of N (u, d)-pairs, user u and document d are picked.
4) As the user u and document d are fixed, a user group or attitude u^* is drawn for document d, from the selected user's Multinomial($\boldsymbol{\beta}_U(u)$). The (u^*, d)-pair in effect selects the parameter vector $\boldsymbol{\theta}_R(u^*, d)$ from the set of $K_U \times N_D$ vectors in the node labeled by $\boldsymbol{\theta}_R$ in Fig. 1.
5) For the generated pair (u^*, d), a binary relevance value r is drawn from the Bernoulli($\boldsymbol{\theta}_R(u^*, d)$).

A.1.1 Likelihood and posterior of user URP model

The likelihood function of the model is

$$P(\mathcal{D} \mid \boldsymbol{\psi}) = \prod_n P(r_n \mid u_n, d_n, \boldsymbol{\psi}) = \prod_n \sum_{u^*} P(u^* \mid u_n, \boldsymbol{\beta}_U) P(r_n \mid u^*, d_n, \boldsymbol{\theta}_R), \tag{7}$$

Table 9 Notation specific to URP model

Symbol	Description
$\boldsymbol{\beta}_U(u)$	Vector of multinomial parameters defining the probabilities of certain user u to belong to different user groups
$\boldsymbol{\theta}_R(u^*, d)$	Vector of Bernoulli parameters defining the probabilities of certain user group u^* to consider document d relevant or irrelevant
$\boldsymbol{\alpha}_U$	Dirichlet prior parameters for all $\boldsymbol{\beta}_U$
$\boldsymbol{\alpha}_R$	Dirichlet prior parameters for all $\boldsymbol{\theta}_R$

where the distributions are

$$\begin{cases} P(u^* \mid u) \sim \text{Multinomial}(\boldsymbol{\beta}_U(u)) \\ P(r \mid u^*, d) \sim \text{Bernoulli}(\boldsymbol{\theta}_R(u^*, d)). \end{cases} \tag{8}$$

The posterior probability is proportional to the product of the likelihood and the priors,

$$\begin{aligned} P(\boldsymbol{\psi} \mid \mathcal{D}, \text{priors}) &= P(\boldsymbol{\beta}_U, \boldsymbol{\theta}_R \mid \mathcal{D}, \boldsymbol{\alpha}_U, \boldsymbol{\alpha}_R) \\ &\propto P(\boldsymbol{\beta}_U \mid \boldsymbol{\alpha}_U) P(\boldsymbol{\theta}_R \mid \boldsymbol{\alpha}_R) P(\mathcal{D} \mid \boldsymbol{\psi}), \end{aligned} \tag{9}$$

where the prior distributions are

$$\begin{cases} P(\boldsymbol{\beta}_U(u)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_U) \\ P(\boldsymbol{\theta}_R(u^*, d)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_R). \end{cases} \tag{10}$$

A.1.2 Sampling formulas of user URP moel

Sampling formula for user group u^* is

$$P(u_n^* \mid u_n, d_n, r_n, \boldsymbol{\psi}) \propto \frac{\boldsymbol{\theta}_R(u_n^*, d_n)_{r_n} \boldsymbol{\beta}_U(u_n)_{u_n^*}}{\sum_{u^*} \boldsymbol{\theta}_R(u^*, d_n)_{r_n} \boldsymbol{\beta}_U(u_n)_{u^*}}. \tag{11}$$

Sampling formula for each parameter vector $\boldsymbol{\beta}_U$ in the users vs. user groups matrix $[\boldsymbol{\beta}_U]$ is

$$\begin{aligned} &P(\boldsymbol{\beta}_U(u) | \{u_n\}, \{u_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(nuu^*1 + \boldsymbol{\alpha}_U(u)_1, \ldots, nuu^*K_U + \boldsymbol{\alpha}_U(u)_{K_U}), \end{aligned} \tag{12}$$

where $nuu^*k = \#\{\text{Ratings with } u_n = u \wedge u_n^* = k\}$.

Sampling formula for each Bernoulli parameter vector $\boldsymbol{\theta}_R(u^*, d)$ is

$$\begin{aligned} &P(\boldsymbol{\theta}_R(u^*, d) | \{d_n\}, \{r_n\}, \{u_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(\boldsymbol{\alpha}_R(0) + nu^*d0, \boldsymbol{\alpha}_R(1) + nu^*d1), \end{aligned} \tag{13}$$

where $nu^*dr = \#\{\text{Ratings with } r_n = r \wedge u_n^* = u^* \wedge d_n = d\}$.

A.2 User URP with generation of users and documents

The generative process of URP-GEN proceeds is given below (see Fig. 3 and summary of notation in Tables 1 and 10).

Note, that since parameters α_D and θ_D are conditionally independent of all other parameters given document d, they have no effect on the predictions of relevance $P(r \mid u, d)$ in this model. So, θ_D is not sampled when modeling the conditional distribution $P(r \mid u, d)$. However, for completeness we describe the full generative process of the model.

1) For each user group u^*, a vector of multinomial parameters $\boldsymbol{\beta}_U(u^*)$ is drawn from Dirichlet($\boldsymbol{\alpha}_U$). This is denoted by the plate with K_U repetitions in the graphical model representation. Each parameter vector $\boldsymbol{\beta}_U(u^*)$ contains the probability for the users to belong to a user group u^*.

2A) For the whole user collection, a vector of multinomial parameters $\boldsymbol{\theta}_U$ is drawn from Dirichlet($\boldsymbol{\alpha}_{u^*}$). The parameter vector $\boldsymbol{\theta}_U$ contains the probabilities of different user groups u^* to occur.

Table 10 Notation specific to URP Model with Generation of Users/Documents (URP-GEN)

Symbol	Description
$\boldsymbol{\beta}_U(u^*)$	Vector of multinomial parameters defining the probabilities of certain user group u^* to contain each user
$\boldsymbol{\theta}_U$	Multinomial probabilities of user groups u^* to occur
$\boldsymbol{\theta}_D$	Multinomial probabilities of documents d to occur (needed only for the generative process)
$\boldsymbol{\theta}_R(u^*, d)$	Vector of Bernoulli parameters defining the probabilities of certain user group u^* to consider document d relevant or irrelevant
$\boldsymbol{\alpha}_U$	Dirichlet prior parameters for all $\boldsymbol{\beta}_U$
$\boldsymbol{\alpha}_{u^*}$	Dirichlet prior parameters for $\boldsymbol{\theta}_U$
$\boldsymbol{\alpha}_D$	Dirichlet prior parameters for $\boldsymbol{\theta}_D$ (needed only for the generative process)
$\boldsymbol{\alpha}_R$	Dirichlet prior parameters for all $\boldsymbol{\theta}_R$

2B) Symmetrically, for the whole document collection, a vector of multinomial parameters $\boldsymbol{\theta}_D$ is drawn from Dirichlet($\boldsymbol{\alpha}_D$). The parameter vector $\boldsymbol{\theta}_D$ contains the probability for each document d to occur.

3) For each (u^*, d) pair, a vector of Bernoulli parameters $\boldsymbol{\theta}_R(u^*, d)$ is drawn from Dirichlet($\boldsymbol{\alpha}_R$). This is denoted by the plate with $K_U \times N_D$ repetitions in the graphical model representation. Each parameter vector $\boldsymbol{\theta}_R(u^*, d)$ defines the probability of the user group u^* to consider document d relevant (or irrelevant).

The rest of the steps are repeated for each of the N rating triplets:

4A) A user group u^* is drawn from Multinomial($\boldsymbol{\theta}_U$). As the user group is fixed the corresponding multinomial parameter vector $\boldsymbol{\beta}_U(u^*)$ can be selected from the set of K_U vectors in the node labeled by $\boldsymbol{\beta}_U$ in Fig. 3. Then, a user u is drawn from Multinomial($\boldsymbol{\beta}_U(u^*)$).

4B) A document d is drawn from Multinomial($\boldsymbol{\theta}_D$).

5) For the generated pair (u^*, d), a binary relevance r is drawn from Bernoulli($\boldsymbol{\theta}_R(u^*, d)$).

A.2.1 Likelihood and posterior of user URP with generation of users/documents

The likelihood function of the model is

$$\begin{aligned} P(\mathcal{D} \mid \boldsymbol{\psi}) &= \prod_n P(r_n \mid u_n, d_n, \boldsymbol{\psi}) P(u_n \mid \boldsymbol{\psi}) P(d_n \mid \boldsymbol{\psi}) \\ &= \prod_n P(d_n \mid \boldsymbol{\theta}_D) \sum_{u^*} P(r_n \mid u^*, d_n, \boldsymbol{\theta}_R) P(u_n \mid u^*, \boldsymbol{\beta}_U) P(u^* \mid \boldsymbol{\theta}_U), \end{aligned} \tag{14}$$

where the distributions are

$$\begin{cases} P(u^*) \sim \text{Multinomial}(\boldsymbol{\theta}_U) \\ P(d) \sim \text{Multinomial}(\boldsymbol{\theta}_D) \\ P(u \mid u^*) \sim \text{Multinomial}(\boldsymbol{\beta}_U(u^*)) \\ P(r \mid u^*, d) \sim \text{Bernoulli}(\boldsymbol{\theta}_R(u^*, d)). \end{cases} \tag{15}$$

The posterior probability is proportional to the product of the likelihood and the priors,

$$\begin{aligned} P(\boldsymbol{\psi} \mid \mathcal{D}, \text{priors}) &= P(\boldsymbol{\beta}_U, \boldsymbol{\theta}_U, \boldsymbol{\theta}_D, \boldsymbol{\theta}_R \mid \mathcal{D}, \boldsymbol{\alpha}_U, \boldsymbol{\alpha}_D, \boldsymbol{\alpha}_{u^*}, \boldsymbol{\alpha}_R) \\ &\propto P(\boldsymbol{\beta}_U \mid \boldsymbol{\alpha}_U) P(\boldsymbol{\theta}_U \mid \boldsymbol{\alpha}_{u^*}) P(\boldsymbol{\theta}_D \mid \boldsymbol{\alpha}_D) P(\boldsymbol{\theta}_R \mid \boldsymbol{\alpha}_R) P(\mathcal{D} \mid \boldsymbol{\psi}), \end{aligned} \tag{16}$$

where the prior distributions are

$$\begin{cases} P(\boldsymbol{\theta}_U) \sim \text{Dirichlet}(\boldsymbol{\alpha}_{u^*}) \\ P(\boldsymbol{\theta}_D) \sim \text{Dirichlet}(\boldsymbol{\alpha}_D) \\ P(\boldsymbol{\beta}_U(u^*)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_U) \\ P(\boldsymbol{\theta}_R(u^*, d)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_R). \end{cases} \tag{17}$$

A.2.2 Sampling formulas of user URP with generation of users/documents

Sampling formula for user group u^* is

$$P(u_n^* \mid u_n, d_n, r_n, \boldsymbol{\psi}) \propto \frac{\boldsymbol{\beta}_U(u_n^*)_{u_n} \boldsymbol{\theta}_R(u_n^*, d_n)_{r_n} \boldsymbol{\theta}_U(u_n^*)}{\sum_{u^*} \boldsymbol{\beta}_U(u^*)_{u_n} \boldsymbol{\theta}_R(u^*, d_n)_{r_n} \boldsymbol{\theta}_U(u^*)}. \tag{18}$$

Sampling formula for each parameter vector $\boldsymbol{\beta}_U$ in the users vs. user groups matrix $[\boldsymbol{\beta}_U]$ is

$$\begin{aligned} &P(\boldsymbol{\beta}_U(u^*) \mid \{u_n\}, \{u_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(nu^*u1 + \boldsymbol{\alpha}_U(u^*)_1, \ldots, nu^*uN_U + \boldsymbol{\alpha}_U(u^*)_{N_U}), \end{aligned} \tag{19}$$

where $nu^*uq = \#\{\text{Samples with } u_n^* = u^* \wedge u_n = q\}$.

Sampling formula for the parameter vector of user group probabilities $\boldsymbol{\theta}_U$ is

$$P(\boldsymbol{\theta}_U \mid \{u_n^*\}, \boldsymbol{\psi}) \propto \text{Dir}(nu^*1 + \boldsymbol{\alpha}_{u^*}(1), \ldots, nu^*K_U + \boldsymbol{\alpha}_{u^*}(K_U)), \tag{20}$$

where $nu^*k = \#\{\text{Samples with } u_n^* = k\}$.

Sampling formula for each Bernoulli parameter vector $\boldsymbol{\theta}_R(u^*, d)$ is

$$\begin{aligned} &P(\boldsymbol{\theta}_R(u^*, d) \mid \{d_n\}, \{r_n\}, \{u_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(\boldsymbol{\alpha}_R(0) + nu^*d0, \boldsymbol{\alpha}_R(1) + nu^*d1), \end{aligned} \tag{21}$$

where $nu^*dr = \#\{\text{Samples with } u_n^* = u^* \wedge d_n = d \wedge r_n = r\}$.

A.3 Two-way model without generation of users/documents

The model generates ratings r given pairs of users and documents, or $(r \mid u, d)$, with binary relevances r as follows (see Fig. 4 and summary of notation in Tables 1 and 11):

1A) For each user u, a vector of multinomial parameters $\boldsymbol{\beta}_U(u)$ is drawn from Dirichlet($\boldsymbol{\alpha}_U$). This is denoted by the plate with N_U repetitions in the graphical model representation. Each parameter vector $\boldsymbol{\beta}_U(u)$ contains the probabilities for a user u to have different attitudes u^*, that is, to belong to different user groups u^*.

1B) Symmetrically, for each document d, a vector of multinomial parameters $\boldsymbol{\beta}_D(d)$ is drawn from Dirichlet($\boldsymbol{\alpha}_D$). This is denoted by the plate with N_D repetitions in the graphical model representation. Each parameter vector $\boldsymbol{\beta}_D(d)$ contains the probabilities for a document d to belong to different document clusters d^*.

Table 11 Notation specific to Two-Way Grouping Model without generation of users and documents (Two-Way NO-GEN)

Symbol	Description
$\boldsymbol{\beta}_U(u)$	Vector of multinomial parameters defining the probabilities of certain user u to belong to different user groups
$\boldsymbol{\beta}_D(d)$	Vector of multinomial parameters defining the probabilities of certain document d to belong to different document clusters
$\boldsymbol{\theta}_R(u^*, d^*)$	Vector of Bernoulli parameters defining the probability of user group u^* to consider document cluster d^* relevant or irrelevant
$\boldsymbol{\alpha}_U$	Dirichlet prior parameters for all $\boldsymbol{\beta}_U$
$\boldsymbol{\alpha}_D$	Dirichlet prior parameters for all $\boldsymbol{\beta}_D$
$\boldsymbol{\alpha}_R$	Dirichlet prior parameters for all $\boldsymbol{\theta}_R$

2) For each cluster pair (u^*, d^*), a vector of Bernoulli parameters $\boldsymbol{\theta}_R(u^*, d^*)$ is drawn from Dirichlet($\boldsymbol{\alpha}_R$). This is denoted by the plate with $K_U \times K_D$ repetitions in the graphical model representation. Each parameter vector $\boldsymbol{\theta}_R(u^*, d^*)$ defines the probability of the user group u^* to consider the document cluster d^* relevant (or irrelevant).

The rest of the steps are repeated for each of the N rating triplets:

3) From the given set of (u, d)-pairs user u and document d are picked. (The set consists of N such pairs.)

4A) As the user u is fixed the corresponding parameter vector $\boldsymbol{\beta}_U(u)$ can be selected from the set of N_U vectors in the node labeled by $\boldsymbol{\beta}_U$ in Fig. 4. Then, a user group u^* is drawn from Multinomial($\boldsymbol{\beta}_U(u)$).

4B) As the document d is fixed the corresponding parameter vector $\boldsymbol{\beta}_D(d)$ can be selected from the set of N_D vectors in the node labeled by $\boldsymbol{\beta}_D$ in Fig. 4. Then, a document cluster d^* is drawn from Multinomial($\boldsymbol{\beta}_D(d)$).

5) For the generated cluster pair (u^*, d^*), a binary relevance r is drawn from Bernoulli($\boldsymbol{\theta}_R(u^*, d^*)$).

A.3.1 Likelihood and posterior of two-way model without generation of users/documents

The likelihood function of the model is

$$\begin{aligned} P(\mathcal{D} \mid \boldsymbol{\psi}) &= \prod_n P(r_n \mid u_n, d_n, \boldsymbol{\psi}) \\ &= \prod_n \sum_{u^*} P(u^* \mid u_n, \boldsymbol{\beta}_U) \sum_{d^*} P(d^* \mid d_n, \boldsymbol{\beta}_D) P(r_n \mid u^*, d^*, \boldsymbol{\theta}_R), \end{aligned} \tag{22}$$

where the distributions are

$$\begin{cases} P(u^* \mid u) \sim \text{Multinomial}(\boldsymbol{\beta}_U(u)) \\ P(d^* \mid d) \sim \text{Multinomial}(\boldsymbol{\beta}_D(d)) \\ P(r \mid u^*, d^*) \sim \text{Bernoulli}(\boldsymbol{\theta}_R(u^*, d^*)). \end{cases} \tag{23}$$

The posterior probability is proportional to the product of the likelihood and the priors,

$$\begin{aligned} P(\boldsymbol{\psi} \mid \mathcal{D}, \text{priors}) &= P(\boldsymbol{\beta}_U, \boldsymbol{\beta}_D, \theta_R \mid \mathcal{D}, \boldsymbol{\alpha}_U, \boldsymbol{\alpha}_D, \boldsymbol{\alpha}_R) \\ &\propto P(\boldsymbol{\beta}_U \mid \boldsymbol{\alpha}_U) P(\boldsymbol{\beta}_D \mid \boldsymbol{\alpha}_D) P(\boldsymbol{\theta}_R \mid \boldsymbol{\alpha}_R) P(\mathcal{D} \mid \boldsymbol{\psi}), \end{aligned} \tag{24}$$

where the prior distributions are

$$\begin{cases} P(\boldsymbol{\beta}_U(u)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_U) \\ P(\boldsymbol{\beta}_D(d)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_D) \\ P(\boldsymbol{\theta}_R(u^*, d^*)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_R). \end{cases} \tag{25}$$

A.3.2 Sampling formulas of two-way model without generation of users/documents

Sampling formula for user group u^* is

$$P(u_n^* \mid u_n, r_n, d_n^*, \boldsymbol{\psi}) \propto \frac{\boldsymbol{\theta}_R(u_n^*, d_n^*)_{r_n} \boldsymbol{\beta}_U(u_n)_{u_n^*}}{\sum_{u^*} \boldsymbol{\theta}_R(u^*, d_n^*)_{r_n} \boldsymbol{\beta}_U(u_n)_{u^*}}. \tag{26}$$

Sampling formula for each parameter vector $\boldsymbol{\beta}_U$ in the users vs. user groups matrix $[\boldsymbol{\beta}_U]$ is

$$\begin{aligned} &P(\boldsymbol{\beta}_U(u) | \{u_n\}, \{u_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(nuu^*1 + \boldsymbol{\alpha}_U(u)_1, \ldots, nuu^*K_U + \boldsymbol{\alpha}_U(u)_{K_U}), \end{aligned} \tag{27}$$

where $nuu^*k = \#\{\text{Ratings with } u_n = u \wedge u_n^* = k\}$.

Sampling formula for each Bernoulli parameter vector $\boldsymbol{\theta}_R(u^*, d^*)$ is

$$\begin{aligned} &P(\boldsymbol{\theta}_R(u^*, d^*) | \{r_n\}, \{u_n^*\}, \{d_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(\boldsymbol{\alpha}_R(0) + nu^*d^*0, \boldsymbol{\alpha}_R(1) + nu^*d^*1), \end{aligned} \tag{28}$$

where $nu^*d^*r = \#\{\text{Ratings with } r_n = r \wedge u_n^* = u^* \wedge d_n^* = d^*\}$.

Sampling formulas for the document-related variables d^* and $\boldsymbol{\beta}_D(d)$ can be derived analogously ($u \leftrightarrow d$).

A.4 Two-way latent grouping model

The generative process proceeds according to the following steps (see also Fig. 2 and summary of notation in Tables 1 and 12):

1A) For the whole user collection, a vector of multinomial parameters $\boldsymbol{\theta}_U$ is drawn from Dirichlet($\boldsymbol{\alpha}_{u^*}$). The parameter vector $\boldsymbol{\theta}_U$ contains the probabilities of different user groups u^* to occur.

2A) For each user group u^*, a vector of multinomial parameters $\boldsymbol{\beta}_U(u^*)$ is drawn from Dirichlet($\boldsymbol{\alpha}_U$). This is denoted by the node $\boldsymbol{\beta}_U$ in Fig. 2. The parameter vector $\boldsymbol{\beta}_U(u^*)$ contains the probability for each user to belong to user group u^*.

1B) Symmetrically, for the whole document collection, a vector of multinomial parameters $\boldsymbol{\theta}_D$ is drawn from Dirichlet($\boldsymbol{\alpha}_{d^*}$). The parameter vector $\boldsymbol{\theta}_D$ contains the probabilities of different document clusters d^* to occur.

2B) For each document cluster d^*, a vector of multinomial parameters $\boldsymbol{\beta}_D(d^*)$ is drawn from Dirichlet($\boldsymbol{\alpha}_D$). The parameter vector $\boldsymbol{\beta}_D(d^*)$ contains the probability for each document to belong to the document cluster d^*.

Table 12 Notation specific to Two-Way model

Symbol	Description
$\boldsymbol{\theta}_U$	Multinomial probabilities of user groups u^* to occur
$\boldsymbol{\beta}_U(u^*)$	Vector of multinomial parameters defining the probabilities of certain user group u^* to contain each user
$\boldsymbol{\theta}_D$	Multinomial probabilities of document clusters d^* to occur
$\boldsymbol{\beta}_D(d^*)$	Vector of multinomial parameters defining the probabilities of certain document cluster d^* to contain each document
$\boldsymbol{\theta}_R(u^*, d^*)$	Vector of Bernoulli parameters defining the probabilities of certain user group u^* to consider document cluster d^* relevant or irrelevant
$\boldsymbol{\alpha}_U$	Dirichlet prior parameters for all $\boldsymbol{\beta}_U$
$\boldsymbol{\alpha}_{u^*}$	Dirichlet prior parameters for $\boldsymbol{\theta}_U$
$\boldsymbol{\alpha}_D$	Dirichlet prior parameters for all $\boldsymbol{\beta}_D$
$\boldsymbol{\alpha}_{d^*}$	Dirichlet prior parameters for $\boldsymbol{\theta}_D$
$\boldsymbol{\alpha}_R$	Dirichlet prior parameters for all $\boldsymbol{\theta}_R$

3) For each cluster pair (u^*, d^*), a vector of Bernoulli parameters $\theta_R(u^*, d^*)$ is drawn from Dirichlet($\boldsymbol{\alpha}_R$). This is denoted by $\boldsymbol{\theta}_R$ residing within both the plate of K_U and repetitions and the plate of K_D repetitions, thus going through all the $K_U \times K_D$ cluster pairs. Each parameter vector $\boldsymbol{\theta}_R(u^*, d^*)$ defines the probability of the user group u^* to consider the document cluster d^* relevant (or irrelevant).

The rest of the steps are repeated for each of the N rating triplets:

4A) A user group u^* is drawn from Multinomial($\boldsymbol{\theta}_U$). As the user group is fixed the corresponding parameter vector $\boldsymbol{\beta}_U(u^*)$ can be selected from the set of K_U vectors in the node labeled by $\boldsymbol{\beta}_U$ in Fig. 2. Then, a user u is drawn from Multinomial($\boldsymbol{\beta}_U(u^*)$).

4B) A document cluster d^* is drawn from Multinomial($\boldsymbol{\theta}_D$). As the document cluster is fixed the corresponding parameter vector $\boldsymbol{\beta}_D(d^*)$ can be selected from the set of K_D vectors in the node labeled by $\boldsymbol{\beta}_D$ in Fig. 2. Then, a document d is drawn from Multinomial($\boldsymbol{\beta}_D(d^*)$).

5) For the generated cluster pair (u^*, d^*), a binary relevance r is drawn from Bernoulli($\boldsymbol{\theta}_R(u^*, d^*)$).

A.4.1 Likelihood and posterior of two-way latent grouping model

The likelihood function of the model is

$$P(\mathcal{D} \mid \boldsymbol{\psi}) = \prod_n P(r_n \mid u_n, d_n, \boldsymbol{\psi}) P(u_n \mid \boldsymbol{\psi}) P(d_n \mid \boldsymbol{\psi}) \tag{29}$$

$$= \prod_n \sum_{u^*} P(u_n \mid u^*, \boldsymbol{\beta}_U) P(u^* \mid \boldsymbol{\theta}_U) \times \sum_{d^*} P(d_n \mid d^*, \boldsymbol{\beta}_D) P(d^* \mid \boldsymbol{\theta}_D) P(r_n \mid u^*, d^*, \boldsymbol{\theta}_R), \tag{30}$$

where the distributions are

$$\begin{cases} P(u^*) \sim \text{Multinomial}(\boldsymbol{\theta}_U) \\ P(d^*) \sim \text{Multinomial}(\boldsymbol{\theta}_D) \\ P(u \mid u^*) \sim \text{Multinomial}(\boldsymbol{\beta}_U(u^*)) \\ P(d \mid d^*) \sim \text{Multinomial}(\boldsymbol{\beta}_D(d^*)) \\ P(r \mid u^*, d^*) \sim \text{Bernoulli}(\boldsymbol{\theta}_R(u^*, d^*)). \end{cases} \tag{31}$$

The posterior probability is proportional to the product of the likelihood and the priors,

$$\begin{aligned} P(\boldsymbol{\psi} \mid \mathcal{D}, \text{priors}) &= P(\boldsymbol{\beta}_U, \boldsymbol{\beta}_D, \boldsymbol{\theta}_U, \boldsymbol{\theta}_D, \boldsymbol{\theta}_R \mid \mathcal{D}, \text{priors}) \\ &\propto P(\boldsymbol{\beta}_U \mid \boldsymbol{\alpha}_U) P(\boldsymbol{\theta}_U \mid \boldsymbol{\alpha}_{u^*}) P(\boldsymbol{\beta}_D \mid \boldsymbol{\alpha}_D) P(\boldsymbol{\theta}_D \mid \boldsymbol{\alpha}_{d^*}) P(\boldsymbol{\theta}_R \mid \boldsymbol{\alpha}_R) P(\mathcal{D} \mid \boldsymbol{\psi}), \end{aligned} \tag{32}$$

where the prior distributions are

$$\begin{cases} P(\boldsymbol{\theta}_U) \sim \text{Dirichlet}(\boldsymbol{\alpha}_{u^*}) \\ P(\boldsymbol{\theta}_D) \sim \text{Dirichlet}(\boldsymbol{\alpha}_{d^*}) \\ P(\boldsymbol{\beta}_U(u^*)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_U) \\ P(\boldsymbol{\beta}_D(d^*)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_D) \\ P(\boldsymbol{\theta}_R(u^*, d^*)) \sim \text{Dirichlet}(\boldsymbol{\alpha}_R). \end{cases} \tag{33}$$

A.4.2 Sampling formulas of two-way model latent grouping model

Sampling formula for user group u^* is

$$P(u_n^* \mid u_n, r_n, d_n^*, \boldsymbol{\psi}) \propto \frac{\boldsymbol{\beta}_U(u_n^*)_{u_n} \boldsymbol{\theta}_R(u_n^*, d_n^*)_{r_n} \boldsymbol{\theta}_U(u_n^*)}{\sum_{u^*} \boldsymbol{\beta}_U(u^*)_{u_n} \boldsymbol{\theta}_R(u^*, d_n^*)_{r_n} \boldsymbol{\theta}_U(u^*)}. \tag{34}$$

Sampling formula for each parameter vector $\boldsymbol{\beta}_U$ in the users vs. user groups matrix $[\boldsymbol{\beta}_U]$ is

$$\begin{aligned} &P(\boldsymbol{\beta}_U(u^*) \mid \{u_n\}, \{u_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(nu^*u1 + \boldsymbol{\alpha}_U(u^*)_1, \ldots, nu^*uN_U + \boldsymbol{\alpha}_U(u^*)_{N_U}), \end{aligned} \tag{35}$$

where $nu^*uq = \#\{\text{Samples with } u_n^* = u^* \wedge u_n = q\}$. Sampling formula for the user group probability parameters $\boldsymbol{\theta}_U$ is

$$P(\boldsymbol{\theta}_U \mid \{u_n^*\}, \boldsymbol{\psi}) \propto \text{Dir}(nu^*1 + \boldsymbol{\alpha}_{u^*}(1), \ldots, nu^*K_U + \boldsymbol{\alpha}_{u^*}(K_U)), \tag{36}$$

where $nu^*k = \#\{\text{Samples with } u_n^* = k\}$.

Sampling formula for each Bernoulli parameter vector $\boldsymbol{\theta}_R(u^*, d^*)$ is

$$\begin{aligned} &P(\boldsymbol{\theta}_R(u^*, d^*) | \{r_n\}, \{u_n^*\}, \{d_n^*\}, \boldsymbol{\psi}) \\ &\quad \propto \text{Dir}(\boldsymbol{\alpha}_R(0) + nu^*d^*0, \boldsymbol{\alpha}_R(1) + nu^*d^*1), \end{aligned} \tag{37}$$

where $nu^*d^*r = \#\{\text{Samples with } u_n^* = u^* \wedge d_n^* = d^* \wedge r_n = r\}$.

Sampling formulas for the document-related variables d^*, $\boldsymbol{\beta}_D(d^*)$, and $\boldsymbol{\theta}_D$ can be derived analogously ($u \leftrightarrow d$).

Table 13 The numbers of documents and users in different data sets

Data	N_{dtest}	N_{dtrain}	N_{utest}	N_{utrain}	N_D	N_U
Artificial	15	285	10	190	300	200
Parliament	65	1207	35	644	1272	679
Scientific articles	22	458	1	24	480	25

Appendix B: Details of experiments

B.1 Construction of test set in "new" users and "new" documents cases

We randomly selected N_{dtest} documents to be the "new" documents. Of the ratings for these documents, we randomly selected 3 ratings per document to be assigned to the training set and the rest of the ratings were assigned to the test set. The other N_{dtrain} documents appeared only in the training set. In the same way, we randomly selected N_{utest} users to be the "new" users. Of the ratings of these users, 3 randomly selected ratings per user were assigned to the training set and the rest of the ratings were left to the test set. The other N_{utrain} users appeared only in the training set.

- *Only New Documents Case*. Those ratings where the user was new were discarded from the test set.
- *Only New Users Case*. Those ratings where the document was new were discarded from the test set.
- *Either New User or New Document Case*. Those ratings where both user and document were new were discarded from the test set.
- *Both New User and New Document Case*. Only those ratings where both user and document were new were included in the test set.

The rest of the preliminary test set became the final test set.

B.2 Demonstrations with artificial data

We produced 10 artificial data sets in both demonstrations. All the models were trained with the known true numbers of clusters. The trained models were tested with separate test sets to produce 10 perplexity values, and the final result was the mean of the perplexities.

We sampled three MCMC chains in parallel and monitored the convergence as described in Sect. 4.3. After the burn-in period, each chain was run for another 400 or more iterations,[5] and finally the samples of all three chains were averaged to estimate expectations of $P(r \mid u, d)$.

B.3 Experimental setup in case studies

For the validation of cluster numbers we used the training set to construct a validation set and a preliminary training set, in a similar manner as for the artificial data above. "New" documents or users included in the test set were not used in the validation. From the rest of the documents we again randomly selected N_{dvalid} documents to be the "new" documents of the validation set ($N_{dvalid} = 65$ in parliament data and $N_{dvalid} = 48$ in the scientific articles data). In the same way we randomly selected N_{uvalid} users to be the "new" users ($N_{uvalid} = 35$ in the parliament data and $N_{uvalid} = 2$ in the scientific articles data).

[5] At most 20,000 iterations were run, depending on the convergence.

Our Two-Way Model, user-based User URP and document-based Document URP were trained with the training set from the validation phase for a range of cluster numbers. The trained models were tested with the validation set, and the lowest perplexity was used as the performance criterion for choosing the cluster numbers. For the final results, the models were trained with all the training data using the validated cluster numbers and tested with the test data set.

We sampled three MCMC chains in parallel and required the convergence check described in Sect. 4.3. After the burn-in period, each chain was run for another $n = 400$ or more iterations, and finally all the $3 \times n$ samples were averaged to estimate expectations of $P(r \mid u, d)$.

The prediction of the naive model was $r = 0$ for the scientific articles and $r = 1$ for the parliament votings.

Appendix C: The supplementary results

C.1 Prediction accuracies

Table 14 Demonstration 1. Accuracy, large values are better. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value ≤ 0.01) are marked with boldface (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	**83**	**83**	**84**	84
Two-Way NO-GEN	60	67	64	56
User URP	77	**83**	81	72
User URP-GEN	78	**83**	81	**75**
Variational URP	76	**79**	78	75
Document URP	**83**	**78**	**81**	**77**
User Freq.	45	50	47	52
Document Freq.	50	49	49	50
Naive Model	50	50	50	48

Table 15 Parliament Data. Comparison between the models by prediction accuracy over the test set. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value ≤ 0.01) are marked with boldface. Large accuracy is better (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	95	95	**95**	**86**
Two-Way NO-GEN	94	71	83	69
User URP	90	**94**	**92**	82
User URP-GEN	91	**96**	93	83
Document URP	**97**	84	91	**87**
User Freq.	54	50	52	52
Document Freq.	66	71	68	64
Naive Model	54	52	53	55

Table 16 Scientific articles data. Comparison between the models by prediction accuracy over the test set. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value ≤ 0.01) are marked with boldface. Large accuracy is better (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	**72**	**71**	**72**	**67**
Two-Way NO-GEN	68	**70**	69	65
User URP	**73**	**74**	**74**	**66**
User URP-GEN	**73**	**74**	**73**	**67**
Document URP	**72**	64	68	59
User Freq.	67	57	62	57
Document Freq.	65	**70**	68	**63**
Naive Model	67	67	67	65

C.2 Performance with more information about new users/documents

In Sect. 5.1.2 we showed graphs about how the results change when varying the amount of information about "new" users or documents. In this section we list the actual result values for those curves. The perplexities are shown in Tables 17–19. Additionally, the corresponding accuracy values are listed in Tables 20–22.

Table 17 With information about 5 ratings for new users/documents: Perplexity of the various models in Demonstration 1. Smaller perplexity is better and 2.0 corresponds to random guessing. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value ≤ 0.01) are marked with boldface (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	1.45	1.53	1.48	1.65
Two-Way NO-GEN	1.86	1.89	1.87	1.98
User URP	1.58	1.65	1.61	1.82
User URP-GEN	1.57	1.57	1.56	1.77
Document URP	1.52	1.72	1.61	1.86
User Freq.	2.02	3.76	2.66	4.63
Document Freq.	2.64	2.02	**2.00**	3.05

Table 18 With information about 10 ratings for new users/documents: Perplexity of the various models in Demonstration 1. Smaller perplexity is better and 2.0 corresponds to random guessing. The best result differs statistically significantly with P-value ≤ 0.01 from the second best one (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	1.35	1.40	1.37	1.54
Two-Way NO-GEN	1.73	1.77	1.75	1.92
User URP	1.44	1.49	1.46	1.66
User URP-GEN	1.44	1.43	1.43	1.61
Document URP	1.41	1.53	1.46	1.69
User Freq.	2.01	2.20	2.09	2.16
Document Freq.	2.15	2.03	2.09	2.14

Table 19 With information about 20 ratings for new users/documents: Perplexity of the various models in Demonstration 1. Smaller perplexity is better and 2.0 corresponds to random guessing. The best result differs statistically significantly with P-value ≤ 0.01 from the second best one ($\mathbf{u}$ = user, $\mathbf{d}$ = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	1.22	1.34	1.26	1.47
Two-Way NO-GEN	1.55	1.64	1.59	1.82
User URP	1.28	1.40	1.32	1.55
User URP-GEN	1.27	1.37	1.30	1.52
Document URP	1.26	1.43	1.33	1.55
User Freq.	2.02	2.07	2.04	2.06
Document Freq.	2.06	2.03	2.05	2.07

Table 20 With information about 5 ratings for new users/documents: Accuracy over test set of the various models in Demonstration 1. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value ≤ 0.01) are marked with boldface. Large accuracy is better ($\mathbf{u}$ = user, $\mathbf{d}$ = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	83	79	81	77
Two-Way NO-GEN	**66**	**63**	65	**57**
User URP	**81**	**78**	**80**	**75**
User URP-GEN	**81**	**79**	**81**	**75**
Document URP	**83**	**75**	**80**	**67**
User Freq.	46	50	48	48
Document Freq.	50	47	54	48
Dumb	50	50	55	47

Table 21 With information about 10 ratings for new users/documents: Accuracy over test set of the various models in Demonstration 1. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value ≤ 0.01) are marked with boldface. Large accuracy is better ($\mathbf{u}$ = user, $\mathbf{d}$ = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	85	**82**	**84**	**77**
Two-Way NO-GEN	**85**	83	**84**	**74**
User URP	**85**	**82**	**84**	**78**
User URP-GEN	**85**	81	**83**	78
Document URP	85	**82**	84	**75**
User Freq.	47	49	48	46
Document Freq.	49	45	47	53
Dumb	49	50	49	49

Table 22 With information about 20 ratings for new users/documents: Accuracy over test set of the various models in Demonstration 1. The best result of each column is underlined and the values that do not differ from the best value statistically significantly (P-value ≤ 0.01) are marked with boldface. Large accuracy is better (**u** = user, **d** = document)

Method	New **d**	New **u**	Either New	Both New
Two-Way Model	**90**	**83**	**87**	**75**
Two-Way NO-GEN	**90**	**83**	**87**	79
User URP	90	**83**	**87**	**78**
User URP-GEN	**90**	**83**	**87**	**78**
Document URP	**90**	84	88	**78**
User Freq.	43	50	46	50
Document Freq.	49	46	48	49
Dumb	50	50	50	54

References

Blei, D. M., & Jordan, M. I. (2003). Modeling annotated data. In *Proceedings of the 26th annual international ACMSIGIR conference on research and development in information retrieval* (pp. 127–134). New York: Assoc. Comput. Mach..

Blei, D., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

British Parliament data. *Votings of the British Parliament*. (1997–2001). http://www.publicwhip.org.uk/project/data.php.

Buntine, W. (2002). Variational extensions to EM and multinomial PCA. In T. Elomaa, H. Mannila, & H. Toivonen (Eds.), *Proceedings of the thirteenth European conference on machine learning, ECML'02* (Vol. 2430, pp. 23–34). Berlin: Springer.

Buntine, W., & Jakulin, A. (2006). Discrete components analysis. In C. Saunders, M. Grobelnik, S. Gunn, & J. Shawe-Taylor (Eds.), *Subspace, latent structure and feature selection techniques*. Berlin: Springer.

Erosheva, E., Fienberg, S., & Lafferty, J. (2004). Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences*, *101*, 5220–5227.

Hastings, W. K. (1970). Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, *57*(1), 97–109.

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

Jin, R., & Si, L. (2004). A Bayesian approach towards active learning for collaborative filtering. In *Proceedings of the twentieth conference on uncertainty in artificial intelligence, UAI'04* (pp. 278–285). AUAI Press.

Keller, M., & Bengio, S. (2004). Theme topic mixture model: A graphical model for document representation. In *PASCAL workshop on text mining and understanding*.

Koivisto, M. (2004). *Sum-product algorithms for the analysis of genetic risks*. Doctoral dissertation, Department of Computer Science, University of Helsinki.

Konstan, J., Miller, B., & Maltz, D., Herlocker, J. (1997). GroupLens: Applying collaborative filtering to usenet news. *Communications of the ACM*, *40*(3), 77–87.

Madeira, S. C., & Oliveira, A. L. (2004). Biclustering algorithms for biological data analysis: a survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *1*, 24–45.

Marlin, B. (2004). Modeling user rating profiles for collaborative filtering. In *Advances in neural information processing systems* (Vol. 16, pp. 627–634). Cambridge: MIT Press.

Marlin, B., Roweis, S. T., & Zemel, R. S. (2005). Unsupervised learning with non-ignorable missing data. In R.G. Cowell, & Z. Ghahramani (Eds.), *Proceedings of the tenth international workshop on artificial intelligence and statistics, AISTATS'05* (pp. 222–229). Society for Artificial Intelligence and Statistics. (Available electronically at http://www.gatsby.ucl.ac.uk/aistats/).

Marlin, B., Zemel, R. S. (2004). The multiple multiplicative factor model for collaborative filtering. In *ICML'04: Proceedings of the 21th international conference on machine learning* (p. 73). New York: Assoc. Comput. Mach. Press.

Marlin, B. M., Zemel, R. S., Roweis, S., & Slaney, M. (2007). Collaborative filtering and the missing at random assumption. In *Proceedings of the 23rd conference on uncertainty in artificial intelligence, UAI'07*.

McCallum, A., Corrada-Emmanuel, A., & Wang, X. (2004). *The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and Academic Email* (Technical Report).

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., & Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, *21*(6), 1087–1092.

Popescul, A., Ungar, L., Pennock, D., & Lawrence, S. (2001). Probabilistic models for unified collaborative and content-based recommendation in sparse-data environments. In *Proceedings of the 17th conference on uncertainty in artificial intelligence, UAI'01* (pp. 437–444). San Mateo: Morgan Kaufmann.

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, *155*, 945–959.

Puolamäki, K., Salojärvi, J., Savia, E., Simola, J., & Kaski, S. (2005). Combining eye movements and collaborative filtering for proactive information retrieval. In G. Marchionini, A. Moffat, J. Tait, R. Baeza-Yates, & N. Ziviani (Eds.), *SIGIR'05: proceedings of the 28th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 146–153). New York: Assoc. Comput. Mach. Press.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th conference on uncertainty in artificial intelligence, UAI'04* (pp. 487–494). AUAI Press.

Savia, E., Puolamäki, K., Sinkkonen, J., & Kaski, S. (2005). Two-way latent grouping model for user preference prediction. In F. Bacchus, & T. Jaakkola (Eds.), *Proceedings of the 21st conference on uncertainty in artificial intelligence, UAI'05* (pp. 518–525). AUAI Press.

Shardanand, U., & Maes, P. (1995). Social information filtering: Algorithms for automating 'word of mouth'. In *Proceedings of the ACM CHI95 human factors in computing systems conference* (pp. 210–217). Cambridge: Assoc. Comput. Mach.

Si, L., & Jin, R. (2003). Flexible mixture model for collaborative filtering. In T. Fawcett & N. Mishra (Eds.), *Proceedings of the twentieth international conference on machine learning, ICML'03* (pp. 704–711). Menlo Park: AAAI Press.

Tanay, A., Sharan, R., & Shamir, R. (2006). Biclustering algorithms: A Survey. In *Handbook of computational molecular biology*. London: Chapman & Hall.

Wettig, H., Lahtinen, J., Lepola, T., Myllymäki, P., & Tirri, H. (2003). Bayesian analysis of online newspaper log data. In *Proceedings of the 2003 symposium on applications and the Internet workshops* (pp. 282–278). Los Alamitos: IEEE Comput. Soc.

Yu, K., Schwaighofer, A., Tresp, V., Xu, X., & Kriegel, H.-P. (2004). Probabilistic memory-based collaborative filtering. *IEEE Transactions on Knowledge and Data Engineering*, *16*(1), 56–69.

Yu, K., Yu, S., & Tresp, V. (2005a). Dirichlet enhanced latent semantic analysis. In R.G. Cowell, & Z. Ghahramani (Eds.), *Proceedings of the tenth international workshop on artificial intelligence and statistics, AISTATS'05* (pp. 437–444). Society for Artificial Intelligence and Statistics.

Yu, S., Yu, K., Tresp, V., & Kriegel, H.-P. (2005b). A probabilistic clustering-projection model for discrete data. In A. Jorge, L. Torgo, P. Brazdil, R. Camacho, & J. Gama (Eds.), *Proceedings of the 9th European conference on principles and practice of knowledge discovery in databases, PKDD'05* (Vol. 3721, pp. 417–428). Berlin: Springer.

Zitnick, C., & Kanade, T. (2004). Maximum entropy for collaborative filtering. In *Proceedings of the 20th conference on uncertainty in artificial intelligence, UAI'04* (pp. 636–643). AUAI Press.