TKK Reports in Information and Computer Science

Espoo 2009 TKK-ICS-R15

# ON TWO-WAY GROUPING BY ONE-WAY TOPIC MODELS

Eerika Savia, Kai Puolamäki and Samuel Kaski

TEKNILLINEN KORKEAKOULU
TEKNISKA HÖGSKOLAN
HELSINKI UNIVERSITY OF TECHNOLOGY
TECHNISCHE UNIVERSITÄT HELSINKI
UNIVERSITE DE TECHNOLOGIE D'HELSINKI

TKK Reports in Information and Computer Science

Espoo 2009 TKK-ICS-R15

# ON TWO-WAY GROUPING BY ONE-WAY TOPIC MODELS

Eerika Savia, Kai Puolamäki and Samuel Kaski

Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science

Teknillinen korkeakoulu
Informaatio- ja luonnontieteiden tiedekunta
Tietojenkäsittelytieteen laitos

Distribution:
Helsinki University of Technology
Faculty of Information and Natural Sciences
Department of Information and Computer Science
P.O.Box 5400
FI-02015 TKK
FINLAND
URL: http://ics.tkk.fi
Tel. +358 9 451 1
Fax +358 9 451 3369
E-mail: series@ics.tkk.fi

ISBN 978-951-22-9946-1 (Print)
ISBN 978-951-22-9947-8 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)
URL: http://lib.tkk.fi/Reports/2009/isbn9789512299478.pdf

TKK ICS
Espoo 2009

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. The same applies for documents in the case of new users. We have shown earlier that if there are both new users and new documents, two-way generalization becomes necessary, and introduced a probabilistic Two-Way Model for the task. The task of finding a two-way grouping is a non-trivial combinatorial problem, which makes it computationally difficult. We suggest approximating the Two-Way Model with two URP models; one that groups users and one that groups documents. Their two predictions are combined using a product of experts model. This combination of two one-way models achieves even better prediction performance than the original Two-Way Model. This article contains the full technical details of the conference article [22].

KEYWORDS: latent topic model, collaborative filtering, cold-start problem, product of experts

# CONTENTS

# 1 INTRODUCTION

This paper considers models for the task of predicting relevance values for user–item pairs based on a set of observed ratings of users for the items. In particular, we concentrate on the task of predicting relevance when very few ratings are known for each user or item.[1]

In so-called collaborative filtering methods 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 enough data available for each user individually. The early collaborative filtering methods were memory-based (see, e.g., [12, 25]). Model-based approaches are justified by the poor scaling of the memory-based techniques. Recent work includes probabilistic and information-theoretic models, see for instance [9, 10, 27, 30].

A family of models most related to our work are the latent topic models, which have been successfully used in document modeling but also in collaborative filtering [3, 5, 6, 11, 15, 16, 18, 19, 21, 28, 29].

The closest related models include probabilistic Latent Semantic Analysis (pLSA; [9]), Latent Dirichlet Allocation (LDA; [2, 4]), and User Rating Profile model (URP; [14]), which all assume a one-way grouping. In addition, there is a two-way grouping model, called Flexible Mixture Model (FMM; [26]). We have discussed the main differences between our Two-Way Model and these related models in [24].

## 1.1 Cold-Start Problem

Since a collaborative filtering system has to rely on the past experiences of the users, it will have problems when assessing new documents that have not yet been 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 imply the problem where new users that have only few ratings will get poor predictions. This problem of unseen or almost unseen users and documents is generally referred to as the *cold-start problem* in recommender system literature, see for instance [13]. The Two-Way Model was proposed to tackle this problem of either new users or new documents [24, 23].

## 1.2 Approximating Two-Way Model with Two One-Way Models

It has been shown for hard biclustering of binary data matrices, that clustering the marginals independently to produce a check-board-like biclustering is guaranteed to achieve fairly good results compared to the NP-hard optimal solution. An approximation ratio for the crossing of two one-way clusterings has been proven [20, 1]. Inspired by this theoretical guarantee, we suggest approximating the Two-Way Model with two User Rating Profile models (URP, [14]); one that groups users and one that groups documents. The combination of the two Gibbs-sampled probabilistic predictions is made using a product of experts model [8].

[1] 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.

We have followed the experimental setups of our earlier study [24] in order to be able to compare the results in a straightforward manner. We briefly describe the experimental scenarios, the performance measures and the baseline models in Sect. 3. In Sect. 6.1 we demonstrate with clearly clustered toy data how the product of two URP models improves the relevance predictions of the corresponding one-way models. Finally, in Sect. 6.2 we show in a real-world case study from our earlier paper that the proposed method works as expected also in practice.

We expected the proposed method to have the advantage of giving better predictions than the individual one-way models with the computational complexity of the one-way model. The one-way grouping models are faster and more reliable in their convergence than the Two-Way Model, basically because of the difference in the intrinsic complexity of the tasks they are solving.

## 2 METHODS

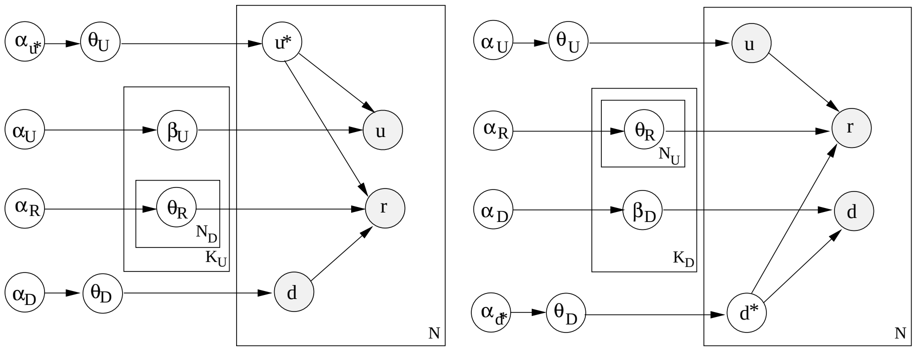
Originally, User Rating Profile model was suggested to be estimated by variational approximation (variational URP, [14]), but we have introduced also Gibbs-sampled variants of the model in [24, 23] (Gibbs URP and Gibbs URP-GEN). The difference between a one-way model and the Two-Way Model is whether to cluster only users (documents) or to cluster both users and documents. Another difference between URP and the Two-Way Model is whether the users and documents are assumed to be generated by the model or treated as covariates of the model. In our earlier study [24] it was found that unless the data marginals are especially misleading about the full data, it is always useful to design the model to be fully generative, in contrast to seeing users and documents as given covariates of the model. Therefore, we have only included the generative variants of Gibbs URP models is this study (Gibbs URP-GEN).

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 |

## 2.1 One-Way Grouping Models

In Fig. 1 we show graphical representations of the generative Gibbs URP model introduced in [24] (User Gibbs URP-GEN), and the corresponding document-grouping variant (Doc Gibbs URP-GEN). They are the one-way grouping models used as the basis of our suggested method. Our main notations are summarized in Table 1 and a table listing distributions of all the random variables can be found in the Appendix.



(a) User Gibbs URP-GEN groups only users and assumes that the relevance depends solely on the user group and the document.

(b) Doc Gibbs URP-GEN groups only documents and assumes that the relevance depends solely on the document cluster and the user.

Figure 1: Graphical model representations of the generative Gibbs URP models with user grouping (User Gibbs URP-GEN) and with document grouping (Doc Gibbs URP-GEN). 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 $K_U$ (or $K_D$) represents the different user groups (or document clusters), and $\boldsymbol{\beta}_U$ (or $\boldsymbol{\beta}_D$) denotes the vector of multinomial parameters for each user group (or document cluster). The plate labeled with $N_D$ (or $N_U$) represents the documents (or users). In the intersection of these plates there is a Bernoulli-model for each of the $K_U \times N_D$ (or $K_D \times N_U$) combinations of user group and document (or document cluster and user). Since $\alpha_D$ and $\theta_D$ (or $\alpha_U$ and $\theta_U$ ) are conditionally independent of all other parameters given document $d$ (or user $u$), they have no effect on the predictions of relevance $P(r \mid u, d)$ in these models. They only describe how documents $d$ (or users $u$) are assumed to be generated. A table listing distributions of all the random variables can be found in the Appendix.

## 2.2 Two-Way Grouping Model

In Fig. 2 we show a graphical representation of the Two-Way Model that our suggested method approximates. A table listing distributions of all the random variables can be found in the Appendix. The Two-Way Model generalizes the generative user-grouping URP by grouping both users and documents. It has been shown to predict relevance more accurately than one-way models 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. Finally, Table 2 summarizes the differences between the models.

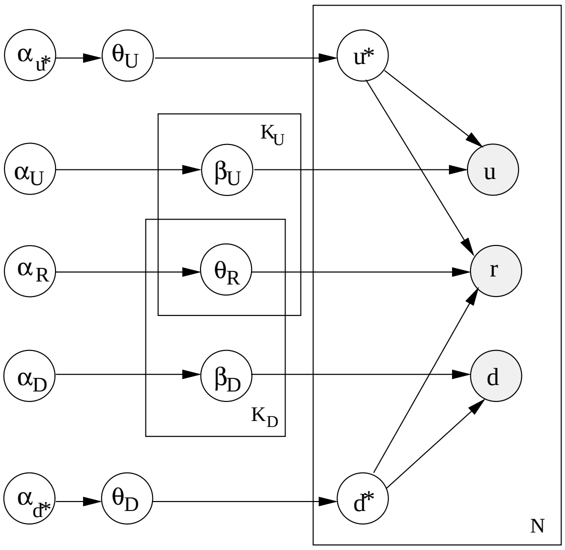


Figure 2: Graphical model representation of the Two-Way Model, which groups both users and documents and assumes that the relevance depends only on the user group and the document cluster instead of individual users/documents. The rightmost plate represents the repeated choice of $N$ (user, document, 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. A table listing distributions of all the random variables can be found in the Appendix.

## 2.3 Approximation of Two-Way Model by Product of Experts

We propose a model where we estimate predictive Bernoulli distributions separately with user-based URP and document-based URP and combine their results with a product of experts model [8]. To be exact, we took the product of the Bernoulli relevance probabilities given by the user-based URP ($P_U(r = 1|u, d)$) and the document-based URP ($P_D(r = 1|u, d)$) and normalized the product distributions, as follows:

$$P_{PoE}(r = 1|u, d) = \frac{P_U(r = 1|u, d)\ P_D(r = 1|u, d)}{\sum_{r=0,1} P_U(r|u, d)\ P_D(r|u, d)} \quad . \tag{1}$$

Table 2: Summary of the models (**u**=user, **d**=document). The column "**Generative**" indicates which of the models generate users and/or documents. The column "**Gibbs**" indicates which of the models are estimates by Gibbs sampling, in contrast to variational approximation. Prefix "2-way" stands for combination of two one-way models.

| **Model Abbreviation** | **Generative** | **Gibbs** | **Groups u** | **Groups d** |
|---|---|---|---|---|
| Two-Way Model | ● | ● | ● | ● |
| 2-way Gibbs URP-GEN | ● | ● | ● | ● |
| 2-way Gibbs URP | – | ● | ● | ● |
| 2-way Variational URP | – | – | ● | ● |
| 1-way User Gibbs URP-GEN | ● | ● | ● | – |
| 1-way User Gibbs URP | – | ● | ● | – |
| 1-way User Varl URP | – | – | ● | – |
| 1-way Doc Gibbs URP-GEN | ● | ● | – | ● |
| 1-way Doc Gibbs URP | – | ● | – | ● |
| 1-way Doc Var URP | – | – | – | ● |

### 2.4 Baseline Models

We compared our results to two simple baseline models. These models mainly serve as an estimate of the lower bound of performance by making an assumption that the data comes from one cluster only. 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:

$$P(r = 1 \mid d) = \frac{\sum_u \#(u, d, r = 1)}{\sum_{u,r} \#(u, d, r)} \quad . \tag{2}$$

The *User Frequency Model*, on the other hand, does not take into account differences between documents or document groups. It is the analogue of Document Frequency Model, where the roles of users and documents have been interchanged.

## 3 EXPERIMENTAL SETTING

### 3.1 Experimental Scenarios

In this section we describe the different types of experimental scenarios that were studied with both data sets. The training and test sets were taken from the earlier study [24]. The scenarios have various levels of difficulty for models that group only users, only documents, or that group both.

- **Only "New" Documents.** This scenario had been constructed to correspond to prediction of relevances for new documents in information retrieval. It had been taken care that each of the randomly selected test documents had only 3 ratings in the training data. The rest of the ratings for these documents had been left to the test set. For the rest of the documents, all the ratings were included in the training set. Hence, the models were able to use "older" documents (for which users' opinions are already known) for training the user groups and document clusters. This scenario favors models that cluster documents.

- **Only "New" Users.** The experimental setting for new users had been constructed in exactly the same way as the setting for new documents but with the roles of users and documents reversed. This scenario favors models that cluster users.
- **Either User or Document is "New".** In an even more general scenario either the users or the documents can be "new." In this setting the test set consisted of user-document pairs where either the user is "new" and the document is "old" or vice versa. This scenario brings out the need for two-way generalization.
- **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.

### 3.2 Sampling

We sampled three MCMC chains in parallel with Gibbs sampling and monitored the convergence as described in[24]. After the burn-in each chain was run for another 400 or more iterations[2], and finally the samples of all three chains were averaged to estimate expectations of $P(r \mid u, d)$.

The Dirichlet priors of multinomials that generate user groups or document clusters, were sampled with the Metropolis-Hastings algorithm [7, 17] with a flat prior in the interval $[1, 10]$ and a Gaussian proposition distribution.

### 3.3 Combining One-Way Models with Variational URP

According to our earlier studies, the variational URP generally seems to produce extreme predictions, near either 0 or 1. Therefore, the variational URP models (User Var URP and Doc Var URP) were combined as a hard biclustering model, as follows. The MAP estimates for cluster belongings from the distributions of the one-way variational URP models were used to divide all the users and documents into bins to produce a hard check-board-like biclustering. In each bicluster the $P(r = 1|u, d)$ was set to the mean of the training data points that lay in the bicluster.

### 3.4 Measures of Performance

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

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

$$\text{perplexity} = e^{-\frac{\mathcal{L}}{N}} \ , \ \text{where } \mathcal{L} = \sum_{i=1}^{N} \log P(r_i \mid u_i, d_i, \mathcal{D}) \quad . \tag{3}$$

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 $P(r_i \mid u_i, d_i, \mathcal{D})$.[3]

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 probability for the relevance. We took the predicted relevance to be

$$\arg \max_{r \in \{0,1\}} P(r \mid u, d, \mathcal{D}) \ , \tag{4}$$

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

# 4 DEMONSTRATION WITH ARTIFICIAL DATA

The artificial data sets were taken from the earlier study [24], and the experimental setting is described in detail in the Appendix. All the models were trained with the known true numbers of clusters ($K_U = K_D = 3$). For each of the 10 data sets the models were trained with a training set and tested with a separate test set, and the final result was the mean of the 10 test set perplexities.

## 4.1 Description of the Data

The data was designed such that it contained bicluster structure with $K_U = K_D = 3$. There were 10 artificial data sets of size 18,000, that all followed the pattern of Fig. 3.

[3] Theoretically, perplexity can grow without a limit if the model predicts zero probability for some element in the test data set, so in practice, we clipped the probabilities to the range $[e^{-10}, 1]$.

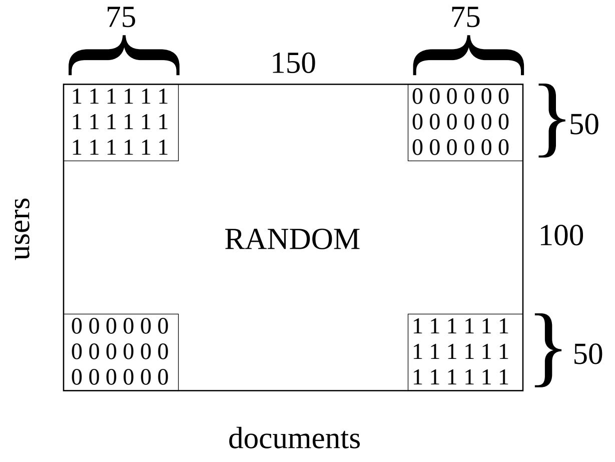


Figure 3: *Focused biclusters* data. The matrix consists of 200 users and 300 documents and the ratings are missing for 70% of the (user, document) pairs. The corners of the matrix are clearly distinguishable biclusters and in the middle the ratings are uniformly random noise. Most of the data is in 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 ratings is 6-fold in the corners compared to the middle of the (user, document) matrix.

We constructed four different settings as described in Sect. 3.1, namely

1. *Only New Documents Case,*
2. *Only New Users Case,*
3. *Either New User or New Document Case,* and
4. *Both New User and New Document Case.*

## 5 EXPERIMENTS WITH PARLIAMENT DATA

We selected the cluster numbers using a validation set described in [24]. The selected cluster numbers are shown in Table 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 3. These values were used in all experimental scenarios.

Table 3: The validated cluster numbers in the parliament case study.

| **Method** | $K_U$ | $K_D$ |
|---|---|---|
| Two-Way Model | 4 | 2 |
| User Gibbs URP-GEN | 2 | – |
| Doc Gibbs URP-GEN | – | 2 |

# 6 RESULTS

## 6.1 Results with Artificial Data

The results of the experiment with artificial data are shown in Table 4. The proposed product of two generative Gibbs URP models outperformed even the Two-Way Model in all the scenarios, being the best in all but the "both new" case, where the hard clustering of MAP estimates of variational URP models was the best. The hard biclustering model worked very well for the variational URP (See Table 4), in contrast to the product of experts - combination, which did not perform well for variational URP. 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 full results with all the accuracy values can be found in the Appendix.

Table 4: Perplexity of the various models in experiments with artificial data. In each column, the best model (underlined) differs statistically significantly from the second-best one (P-value $\leq$ 0.01). Small perplexity is better; 2.0 corresponds to binary random guessing and 1.0 to perfect prediction.

| | New | New | Either | Both |
|---|---|---|---|---|
| **Method** | **Doc** | **User** | **New** | **New** |
| Two-Way Model | 1.52 | 1.54 | 1.53 | 1.70 |
| 2-way Gibbs URP-GEN | 1.46 | 1.47 | 1.45 | 1.70 |
| 2-way Var URP | 1.55 | 1.57 | 1.54 | 1.52 |
| User Gibbs URP-GEN | 1.68 | 1.57 | 1.62 | 1.83 |
| User Var URP | 7.03 | 2.07 | 3.45 | 9.27 |
| Doc Gibbs URP-GEN | 1.56 | 1.69 | 1.62 | 1.81 |
| Doc Var URP | 1.86 | 5.99 | 3.08 | 6.90 |
| User Freq. | 2.02 | 5.65 | 3.25 | 4.99 |
| Document Freq. | 5.29 | 2.01 | 3.21 | 5.92 |

## 6.2 Results with Parliament Data

The product of two generative Gibbs URP models outperformed even the Two-Way Model in all the scenarios, being the best in all cases (see Table 5). The prediction accuracy of the best model varied between 93–97%, while the prediction accuracy of the best baseline model varied between 64–71%. The full results can be found in the Appendix.

Table 5: 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 $\leq 0.01$). Small perplexity is better; 2.0 corresponds to binary random guessing and 1.0 to perfect prediction.

| Method | New Doc | New User | Either New | Both New |
|---|---|---|---|---|
| Two-Way Model | 1.37 | 1.40 | 1.38 | 1.62 |
| 2-way Gibbs URP-GEN | 1.19 | 1.22 | 1.20 | 1.45 |
| User Gibbs URP-GEN | 1.47 | 1.34 | 1.41 | 1.64 |
| Doc Gibbs URP-GEN | **1.34** | 1.54 | 1.43 | 1.68 |
| User Freq. | 2.00 | 5.68 | 3.32 | 4.78 |
| Document Freq. | 5.36 | 1.76 | 3.12 | 5.85 |

## 7 DISCUSSION

We have tackled the problem of new users or documents in collaborative filtering. We have shown in our previous work that if there are both new users and new documents, two-way generalization becomes necessary, and introduced a probabilistic Two-Way Model for the task in [24].

In this paper we suggest an approximation for the Two-Way Model with two User Rating Profile models — one that groups users and one that groups documents — which are combined as a product of experts (PoE). We show with two data sets from the earlier study [24], that the PoE model achieves the performance level of the more principled Two-Way Model and even outperforms it.

The task of finding such a two-way grouping that best predicts the relevance is a difficult combinatorial problem, which makes convergence of the sampling hard to achieve. This work was motivated by the finding that hard biclustering of binary data can be approximated using two one-way clusterings with a proven approximation ratio.

The main advantage of the proposed method, compared to earlier works, is the ability to make at least as good predictions as the Two-Way Model but with the computational complexity of the one-way model. The one-way grouping models are faster and more reliable in their convergence than the Two-Way Model, basically because of the difference in the intrinsic complexity of the tasks they are solving.

### Acknowledgements

This work was supported in part by the PASCAL2 NoE of the EC. This publication only reflects the authors' views. Access rights to the data sets and are restricted due to other commitments. This work was done in the Adaptive Informatics Research Centre, a CoE of the Academy of Finland.

## REFERENCES

[1] A. Anagnostopoulos, A. Dasgupta, and R. Kumar. Approximation algorithms for co-clustering. In *Proceedings of the Twenty-Seventh ACM SIGMOD-SIGACT-SIGART symposium on Principles of database systems*, pages 201–210, New York, NY, USA, 2008. ACM.

[2] D. Blei, A. Y. Ng, and M. I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[3] David M. Blei and Michael I. Jordan. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 127–134. ACM Press, 2003.

[4] Wray Buntine. Variational extensions to EM and multinomial PCA. In T. Elomaa, H. Mannila, and H. Toivonen, editors, *Proceedings of the Thirteenth European Conference on Machine Learning, ECML'02*, volume 2430 of *Lecture Notes in Artificial Intelligence*, pages 23–34. Springer-Verlag, 2002.

[5] Wray Buntine and Aleks Jakulin. Discrete components analysis. In C. Saunders, M. Grobelnik, S. Gunn, and J. Shawe-Taylor, editors, *Subspace, Latent Structure and Feature Selection Techniques*. Springer-Verlag, 2006.

[6] Elena Erosheva, Stephen Fienberg, and John Lafferty. Mixed membership models of scientific publications. *Proceedings of the National Academy of Sciences*, 101:5220–5227, 2004.

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

[8] Geoffrey E. Hinton. An approximation ratio for biclustering. *Neural Computation*, 14(8):1771–1800, 2002.

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

[10] R. Jin and L. Si. A Bayesian approach towards active learning for collaborative filtering. In *Proceedings of the Twentieth Conference on Uncertainty in Artificial Intelligence, UAI'04*, pages 278–285. AUAI Press, 2004.

[11] Mikaela Keller and Samy Bengio. Theme topic mixture model: A graphical model for document representation. In *PASCAL Workshop on Text Mining and Understanding*, 2004.

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

[13] Xuan Nhat Lam, Thuc Vu, Trong Duc Le, and Anh Duc Duong. Addressing cold-start problem in recommendation systems. In *ICUIMC'08: Proceedings of the 2nd international conference on Ubiquitous information management and communication*, pages 208–211, New York, NY, USA, 2008. ACM.

[14] Benjamin Marlin. Modeling user rating profiles for collaborative filtering. In *Advances in Neural Information Processing Systems 16*, pages 627–634. MIT Press, 2004.

[15] Benjamin Marlin and Richard S. Zemel. The multiple multiplicative factor model for collaborative filtering. In *ICML'04: Proceedings of the 21th International Conference on Machine Learning*, page 73. ACM Press, 2004.

[16] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. The author-recipient-topic model for topic and role discovery in social networks: Experiments with Enron and Academic Email. Technical report, University of Massachusetts, December 2004.

[17] Nicholas Metropolis, Arianna W. Rosenbluth, Marshall N. Rosenbluth, Augusta H. Teller, and Edward Teller. Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092, 1953.

[18] A. Popescul, L.H. Ungar, D.M. Pennock, and S. Lawrence. 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*, pages 437–444. Morgan Kaufmann, 2001.

[19] Jonathan K. Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155:945–59, 2000.

[20] Kai Puolamäki, Sami Hanhijärvi, and Gemma C. Garriga. An approximation ratio for biclustering. *Information Processing Letters*, 108:45–49, 2008.

[21] Michal Rosen-Zvi, Tom Griffiths, Mark Steyvers, and Padhraic Smyth. The author-topic model for authors and documents. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI'04*, pages 487–494. AUAI Press, 2004.

[22] E. Savia, K. Puolamäki, and S. Kaski. Two-way grouping by one-way topic models. In *Proceedings of the 8th International Symposium on Intelligent Data Analysis (IDA)*. Springer-Verlag, 2009.

[23] E. Savia, K. Puolamäki, J. Sinkkonen, and S. Kaski. Two-way latent grouping model for user preference prediction. In F. Bacchus and T. Jaakkola, editors, *Uncertainty in Artificial Intelligence 21*, pages 518–525. AUAI Press, Corvallis, Oregon, 2005.

[24] Eerika Savia, Kai Puolamäki, and Samuel Kaski. Latent grouping models for user preference prediction. *Machine Learning*, 74(1):75–109, 2009. Published online: 3 September 2008.

[25] U. Shardanand and P. Maes. Social information filtering: Algorithms for automating 'word of mouth'. In *Proceedings of the ACM CHI 95 Human Factors in Computing Systems Conference*, pages 210–217. ACM/Addison-Wesley, 1995.

[26] Luo Si and Rong Jin. Flexible mixture model for collaborative filtering. In T. Fawcett and N. Mishra, editors, *Proceedings of the Twentieth International Conference on Machine Learning, ICML'03*, pages 704–711. AAAI Press, 2003.

[27] H. Wettig, J. Lahtinen, T. Lepola, P. Myllymäki, and H. Tirri. Bayesian analysis of online newspaper log data. In *Proceedings of the 2003 Symposium on Applications and the Internet Workshops*, pages 282–278, Los Alamitos, California, 2003. IEEE Computer Society.

[28] Kai Yu, Shipeng Yu, and Volker Tresp. Dirichlet enhanced latent semantic analysis. In Robert G. Cowell and Zoubin Ghahramani, editors, *Proceedings of the Tenth International Workshop on Artificial Intelligence and Statistics, AISTATS'05*, pages 437–444. Society for Artificial Intelligence and Statistics, 2005.

[29] Shipeng Yu, Kai Yu, Volker Tresp, and Hans-Peter Kriegel. A probabilistic clustering-projection model for discrete data. In Alípio Jorge, Luís Torgo, Pavel Brazdil, Rui Camacho, and João Gama, editors, *Proceedings of the 9th European Conference on Principles and Practice of Knowledge Discovery in Databases, PKDD'05*, volume 3721 of *Lecture Notes in Computer Science*, pages 417–428. Springer, 2005.

[30] C.L. Zitnick and T. Kanade. Maximum entropy for collaborative filtering. In *Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence, UAI'04*, pages 636–643. AUAI Press, 2004.

# 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 ($n \in \{1, 2, \dots, N\}$), $\mathcal{D}$ denotes all the observed data, and $\boldsymbol{\psi}$ denotes all the parameters of the model.

## A.1 Generative User URP (User Gibbs URP-GEN)

The generative process of URP-GEN proceeds is given below (See Fig. 1(a) and summary of notation in Tables 1 and 6).

Note, that since parameters $\alpha_D$ and $\theta_D$ 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, $\theta_D$ 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.

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. 1(a). 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($\boldsymbol{\theta}_R(u^*, d)$).

**Likelihood and Posterior of Generative User URP.**
The likelihood function of the model is

$$\begin{aligned} P(\mathcal{D} \mid \boldsymbol{\psi}) &= \prod_n P(r_n \mid u_n, d_n, \boldsymbol{\psi})\, P(u_n \mid \boldsymbol{\psi})\, P(d_n \mid \boldsymbol{\psi}) \qquad (5) \\ &= \prod_n P(d_n \mid \boldsymbol{\theta}_D) \sum_{u^*} P(r_n \mid u^*, d_n, \boldsymbol{\theta}_R)\, P(u_n \mid u^*, \boldsymbol{\beta}_U)\, P(u^* \mid \boldsymbol{\theta}_U)\ , \end{aligned}$$

where the distributions are

$$\begin{cases} P(u^*) & \sim \text{Multinomial}(\boldsymbol{\theta}_U) \\ P(d) & \sim \text{Multinomial}(\boldsymbol{\theta}_D) \\ P(u \mid u^*) & \sim \text{Multinomial}(\boldsymbol{\beta}_U(u^*)) \\ P(r \mid u^*, d) & \sim \text{Bernoulli}(\boldsymbol{\theta}_R(u^*, d))\ . \end{cases} \qquad (6)$$

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

$$\begin{aligned} P(\boldsymbol{\psi} \mid \mathcal{D}, \text{priors}) &= P(\boldsymbol{\beta}_U, \boldsymbol{\theta}_U, \boldsymbol{\theta}_D, \boldsymbol{\theta}_R \mid \mathcal{D}, \boldsymbol{\alpha}_U, \boldsymbol{\alpha}_D, \boldsymbol{\alpha}_{u^*}, \boldsymbol{\alpha}_R) \qquad (7) \\ \propto\ & P(\boldsymbol{\beta}_U \mid \boldsymbol{\alpha}_U)\, P(\boldsymbol{\theta}_U \mid \boldsymbol{\alpha}_{u^*})\, P(\boldsymbol{\theta}_D \mid \boldsymbol{\alpha}_D)\, P(\boldsymbol{\theta}_R \mid \boldsymbol{\alpha}_R)\, P(\mathcal{D} \mid \boldsymbol{\psi})\ , \end{aligned}$$

where the prior distributions are

$$\begin{cases} P(\boldsymbol{\theta}_U) & \sim \text{Dirichlet}(\boldsymbol{\alpha}_{u^*}) \\ P(\boldsymbol{\theta}_D) & \sim \text{Dirichlet}(\boldsymbol{\alpha}_D) \\ P(\boldsymbol{\beta}_U(u^*)) & \sim \text{Dirichlet}(\boldsymbol{\alpha}_U) \\ P(\boldsymbol{\theta}_R(u^*, d)) & \sim \text{Dirichlet}(\boldsymbol{\alpha}_R)\ . \end{cases} \qquad (8)$$

Table 6: Notation specific to Generative User URP Model (User Gibbs 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 $\boldsymbol{\theta}_R$ |

**Sampling Formulas of Generative User URP.**
Sampling formula for user group $u^*$ is

$$P(u_n^* \mid u_n, d_n, r_n, \boldsymbol{\psi}) \propto \frac{\boldsymbol{\beta}_U(u_n^*)_{u_n}\ \boldsymbol{\theta}_R(u_n^*, d_n)_{r_n}\ \boldsymbol{\theta}_U(u_n^*)}{\sum_{u^*} \boldsymbol{\beta}_U(u^*)_{u_n}\ \boldsymbol{\theta}_R(u^*, d_n)_{r_n}\ \boldsymbol{\theta}_U(u^*)}\ . \qquad (9)$$

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

$$\begin{aligned} P(\boldsymbol{\beta}_U(u^*) \quad &| \quad \{u_n\}, \{u_n^*\}, \boldsymbol{\psi}) \propto \\ \text{Dir} \quad &(nu^*u1 + \boldsymbol{\alpha}_U(u^*)_1\,,\ \ldots,\ nu^*uN_U + \boldsymbol{\alpha}_U(u^*)_{N_U}) \ ,(10) \end{aligned}$$

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

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

$$P(\boldsymbol{\theta}_U \mid \{u_n^*\}, \boldsymbol{\psi}) \propto \text{Dir}(nu^*1 + \boldsymbol{\alpha}_{u^*}(1)\,,\ \ldots,\ nu^*K_U + \boldsymbol{\alpha}_{u^*}(K_U)) \ \ , \quad (11)$$

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

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

$$\begin{aligned} P(\boldsymbol{\theta}_R(u^*, d) \quad &| \quad \{d_n\}, \{r_n\}, \{u_n^*\}, \boldsymbol{\psi}) \propto \\ \text{Dir} \quad &(\boldsymbol{\alpha}_R(0) + nu^*d0,\ \boldsymbol{\alpha}_R(1) + nu^*d1) \ \ , \qquad (12) \end{aligned}$$

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

## A.2 Two-Way Model

The generative process proceeds according to the following steps (see also Fig. 2 and summary of notation in Tables 1 and 7):

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 $d^*$.

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($\boldsymbol{\theta}_R(u^*, d^*)$).

**Likelihood and Posterior of Two-Way Model.**
The likelihood function of the model is

$$
\begin{aligned}
P(\mathcal{D} \mid \boldsymbol{\psi}) &= \prod_n P(r_n \mid u_n, d_n, \boldsymbol{\psi})\, P(u_n \mid \boldsymbol{\psi})\, P(d_n \mid \boldsymbol{\psi}) && (13)\\
&= \prod_n \sum_{u^*} P(u_n \mid u^*, \boldsymbol{\beta}_U)\, P(u^* \mid \boldsymbol{\theta}_U) \,\cdot && (14)\\
&\quad \sum_{d^*} P(d_n \mid d^*, \boldsymbol{\beta}_D)\, P(d^* \mid \boldsymbol{\theta}_D)\, P(r_n \mid u^*, d^*, \boldsymbol{\theta}_R)\ ,
\end{aligned}
$$

where the distributions are

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

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

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

where the prior distributions are

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

Table 7: 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 $\boldsymbol{\theta}_R$ |

**Sampling Formulas of Two-Way Model.**
Sampling formula for user group $u^*$ is

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

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

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

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

Sampling formula for the user group probability parameters $\boldsymbol{\theta}_U$ is

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

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

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

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

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 $(u \leftrightarrow d)$.

# 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 taken to the training set and the rest of the ratings were left to the test set. The other $N_{dtrain}$ documents only appeared 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, randomly selected 3 ratings per user were taken to the training set and the rest of the ratings were left to the test set. The other $N_{utrain}$ users only appeared in the training set.

- *Only New Documents Case.* Those ratings where user was new were discarded from the test set.
- *Only New Users Case.* Those ratings where 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.

Table 8: The numbers of documents and users in different data sets.

| **Data** | $\mathbf{N_{dtest}}$ | $\mathbf{N_{dtrain}}$ | $\mathbf{N_{utest}}$ | $\mathbf{N_{utrain}}$ | $\mathbf{N_D}$ | $\mathbf{N_U}$ |
|---|---|---|---|---|---|---|
| Artificial | 15 | 285 | 10 | 190 | 300 | 200 |
| Parliament | 65 | 1207 | 35 | 644 | 1272 | 679 |

## B.2 Details of Experimental Setup

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 $N_{dvalid} = 65$ documents to be the "new" documents of the validation set. In the same way we randomly selected $N_{uvalid} = 35$ users to be the "new" users.

Our Two-Way Model, user-based User URP-GEN and document-based Doc URP-GEN were trained with the training set of 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 with the validated cluster numbers and tested with the test data set.

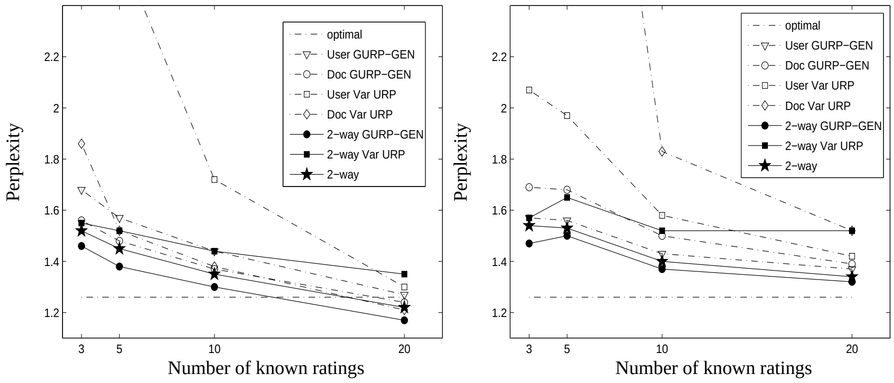
We sampled three MCMC chains in parallel and required the convergence check described in [24]. After the burn-in each chain was run for another $n$ =400 or more iterations, and finally all the $3 \times n$ samples were averaged to estimate expectations of $P(r \mid u, d)$.

We also implemented a very simple baseline model for the accuracy results. The *Naive Model* always predicts the same relevance value for $r$, according to the more frequent value in the training set. The prediction of the naive model was $r = 0$ for the scientific articles and $r = 1$ for the parliament votings.

# C SUPPLEMENTARY RESULTS

## C.1 Effect of the Amount of Rating Information about New Users and Documents.

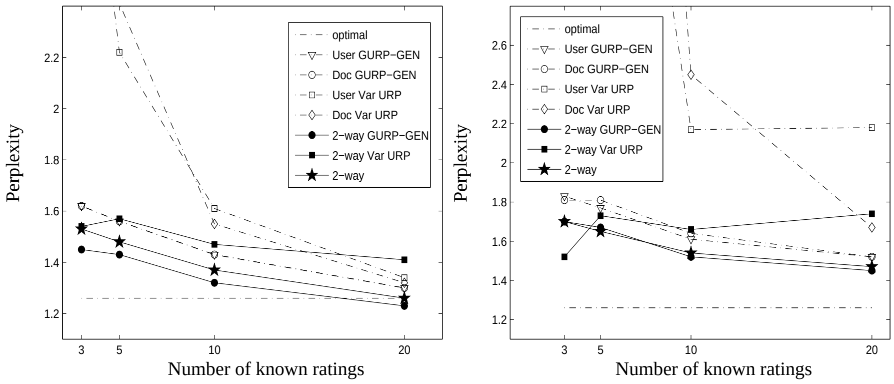
The number of known ratings for "new" users/documents was varied within $\{3, 5, 10, 20\}$. As we here know the true model that was used to generate the data, we can also compare the results to the theoretically derived optimal perplexity (1.26) achieved with the true model, shown as a horizontal line in the figures.



(a) "New Documents" case.

(b) "New Users" case.

Figure 4: Perplexity as a function of the amount of rating information about "new" users/documents.



(a) "Either New" case.

(b) "Both New" case.

Figure 5: Perplexity as a function of the amount of rating information about "new" users/documents.

## C.2 Prediction Accuracies

Table 9: Demonstration with artificial data. 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 $\leq$ 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 |
| | | | | |
| 2-way Gibbs URP-GEN | **83** | 84 | **84** | **79** |
| 2-way Var URP | **82** | **80** | **82** | **81** |
| 2-way User/Doc Freq. | 50 | 50 | 50 | 52 |
| | | | | |
| User Gibbs URP-GEN | 78 | **83** | **81** | **75** |
| User Var URP | 76 | **81** | 79 | **75** |
| | | | | |
| Doc Gibbs URP-GEN | **83** | **78** | **80** | **79** |
| Doc Var URP | **81** | **77** | 79 | **78** |
| | | | | |
| User Freq. | 45 | 50 | 47 | 52 |
| Document Freq. | 50 | 49 | 49 | 50 |
| Naive Model | 50 | 50 | 50 | 48 |

Table 10: 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 $\leq$ 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** |
| | | | | |
| 2-way Gibbs URP-GEN | **97** | **94** | **96** | 93 |
| 2-way User/Doc Freq. | 66 | 64 | 65 | 64 |
| | | | | |
| User Gibbs URP-GEN | **89** | **96** | **92** | **83** |
| | | | | |
| Doc Gibbs URP-GEN | **97** | **86** | **91** | **87** |
| | | | | |
| User Freq. | 54 | 50 | 52 | 52 |
| Document Freq. | 66 | 71 | 68 | 64 |
| Naive Model | 54 | 52 | 53 | 55 |

TKK REPORTS IN INFORMATION AND COMPUTER SCIENCE

| | |
|---|---|
| TKK-ICS-R5 | Matti Koskimies<br>Applying Model Checking to Analysing Safety Instrumented Systems. June 2008. |
| TKK-ICS-R6 | Alexander Ilin, Tapani Raiko<br>Practical Approaches to Principal Component Analysis in the Presence of Missing Values. June 2008. |
| TKK-ICS-R7 | Kai Puolamäki, Samuel Kaski<br>Bayesian Solutions to the Label Switching Problem. June 2008. |
| TKK-ICS-R8 | Abhishek Tripathi, Arto Klami, Samuel Kaski<br>Using Dependencies to Pair Samples for Multi-View Learning. October 2008. |
| TKK-ICS-R9 | Elia Liitiäinen, Francesco Corona, Amaury Lendasse<br>A Boundary Corrected Expansion of the Moments of Nearest Neighbor Distributions. October 2008. |
| TKK-ICS-R10 | He Zhang, Markus Koskela, Jorma Laaksonen<br>Report on forms of enriched relevance feedback. November 2008. |
| TKK-ICS-R11 | Ville Viitaniemi, Jorma Laaksonen<br>Evaluation of pointer click relevance feedback in PicSOM. November 2008. |
| TKK-ICS-R12 | Markus Koskela, Jorma Laaksonen<br>Specification of information interfaces in PinView. November 2008. |
| TKK-ICS-R13 | Jorma Laaksonen<br>Definition of enriched relevance feedback in PicSOM. November 2008. |
| TKK-ICS-R14 | Jori Dubrovin<br>Checking Bounded Reachability in Asynchronous Systems by Symbolic Event Tracing. April 2009. |

ISBN 978-951-22-9946-1 (Print)
ISBN 978-951-22-9947-8 (Online)
ISSN 1797-5034 (Print)
ISSN 1797-5042 (Online)