Neural Relational Topic Models for Scientific Article Analysis

Haoli Bai
The Chinese University of Hong Kong
Hong Kong
Shenzhen Research Institute
The Chinese University of Hong Kong
Shenzhen, China
hlbai@cse.cuhk.edu.hk

Zhuangbin Chen
The Chinese University of Hong Kong
Hong Kong
Shenzhen Research Institute
The Chinese University of Hong Kong
Shenzhen, China
zbchen@cse.cuhk.edu.hk

Michael R. Lyu
The Chinese University of Hong Kong
Hong Kong
Shenzhen Research Institute
The Chinese University of Hong Kong
Shenzhen, China
lyu@cse.cuhk.edu.hk

Irwin King
The Chinese University of Hong Kong
Hong Kong
Shenzhen Research Institute
The Chinese University of Hong Kong
Shenzhen, China
king@cse.cuhk.edu.hk

Zenglin Xu
SMILE Lab
University of Electronic Science and
Technology of China
Chengdu, China
zenglin@gmail.com

ABSTRACT

Topic modelling and citation recommendation of scientific articles are important yet challenging research problems in scientific article analysis. In particular, the inference on coherent topics can be easily affected by irrelevant contents in articles. Meanwhile, the extreme sparsity of citation networks brings difficulty to a valid citation recommendation. Intuitively, articles with similar topics are more likely to cite each other, and cited articles tend to share similar themes. Motivated from this intuition, we aim to boost the performance of both topic modelling and citation recommendation by effectively leverage this underlying correlation between latent topics and citation networks. To this end, we propose a novel Bayesian deep generative model termed as Neural Relational Topic Model (NRTM), which is composed with a Stacked Variational Auto-Encoder (SVAE) and a multilayer perception (MLP). Specifically, the SVAE utilizes an inference network to learn more representative topics of document contents, which can help to enrich the latent factors in collaborative filtering of citations. Furthermore, the MLP network conducts nonlinear collaborative filtering of citations, which can further benefit the inference of topics by leveraging the knowledge of citation networks. Extensive experiments on two real-world datasets demonstrate that our model can effectively take advantages of the coherence between topic learning and citation recommendation, and significantly outperform the state-of-the-art methods on both tasks.

CIKM '18, October 22–26, 2018, Torino, Italy

ACM ISBN 978-1-4503-6014-2/18/10…$15.00
https://doi.org/10.1145/3269206.3271696

CCS CONCEPTS

• **Information systems → Document topic models**; **Recommender systems**; *Content analysis and feature selection*; *Query suggestion*; *Probabilistic retrieval models*; *Document filtering*; Top-k retrieval in databases; • **Computing methodologies → Topic modeling**; **Latent Dirichlet allocation**; *Bayesian network models*; • **Mathematics of computing →** *Bayesian networks*; *Computing most probable explanation*;

KEYWORDS

Topic Modelling, Citation Recommendation, Deep Learning

ACM Reference Format:
Haoli Bai, Zhuangbin Chen, Michael R. Lyu, Irwin King, and Zenglin Xu. 2018. Neural Relational Topic Models for Scientific Article Analysis. In *The 27th ACM International Conference on Information and Knowledge Management (CIKM '18), October 22–26, 2018, Torino, Italy*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3269206.3271696

1 INTRODUCTION

Topic modelling and citation recommendation are two widely concerned yet challenging tasks for scientific article analysis [6, 8]. Specifically, with the increasing diversity of interdisciplinary scientific articles, the inference of coherent topics are significantly challenged by irrelevant contents in articles. Meanwhile, the growing size of citation networks leads to extreme sparsity of citations, which poses challenges to a valid recommendation.

Intuitively, when documents display similar topics, they tend to cite each other with a higher probability. Similarly, articles where there are citations tend to show some closeness in content themes. Based on this observation, researchers resort to Bayesian latent variable models [6, 8, 14, 20, 26] to jointly infer latent topics and predict citations, such that this underlying correlation can be helpful for both topic discovery and citation prediction.

Despite the success of these inspiring works, previous methods mainly suffer from three drawbacks. First, the shallow structure in previous methods may be insufficient for modeling the complex correlation between latent topics and citation networks, and thus may lead to suboptimal solutions. Although more recent works [18, 35, 36] can learn latent representations of documents via deep auto-encoders, the collaborative filtering based on linear matrix factorization is still insufficient to exploit the underlying correlation. Second, previous Bayesian models either suffer from the tricky derivations of variational inference, or the computational burden of sampling techniques during the approximation of posteriors. Although neural topic models [24, 25] are recently proposed to address the inference problem, they are designed purely for text analysis without considering the information from citation networks. Third, the topic coherence with the presence of citation network has not been explored in past literature. Topic coherence has been a long standing problem in topic modelling, which could also affect the performance of recommendation through the underlying correlation in our setting.

In order to address the above problems, we propose a novel Bayesian deep generative model named Neural Relational Topic Model (NRTM), which is composed of a Stacked Variational Auto-Encoder (SVAE) and a multilayer perception (MLP). Specifically, to resolve the first problem, we incorporate SVAE with MLP to collectively enhance the model capacity. The topics of learned by SVAE can enrich the latent factors for MLP, and MLP is able to perform nonlinear collaborative filtering for citation recommendation. Conversely, the knowledge from citation networks can also be transferred back to topic modelling sufficiently via the powerful multilayer perceptions. Therefore, the correlation between two sources are mutually leveraged in our deep neural architecture. To address the second problem, we adopt neural variational inference [15] for our Bayesian generative model, leading to an efficient and black-box learning scheme. For the third challenge, we directly investigate the topic coherence with the recently proposed evaluation metrics [29]. The coherence of topics can be improved by both the expressive SVAE as well as the mutual exploitation of the correlation between two domains.

In summary, the contributions of our paper are as follows:

- We propose a Bayesian deep generative model named Neural Relational Topic Model, which can mutually leverage the underlying correlation between latent topics and citation networks.
- We adopt neural variational inference for the inference of document latent topics in our model, leading to an efficient and black box learning scheme.
- We investigate the coherence of learned topics with the presence of citation networks, which has not been explored in past literature to the best of our knowledge. An illustrative example of learned coherent topics can be found in Table 4.
- Extensive experiments on two real-world datasets demonstrate that our method effectively leverages the correlation via deep architectures, and achieves the state-of-the-art performance on both topic modelling and citation recommendation tasks.

2 PRELIMINARIES

In this section, we cover through the necessary background knowledge for our problem. The notations used throughout the paper are as follows: we denote the citation network as $Y \in \mathrm{R}^{N \times N}$, where $y_{ij} = 1$ stands for a citation (a positive link) between document i and j, and 0 (a negative link) otherwise. For convenience, we denote $\mathcal{Y}^{(+)}$ as the set of positive links, while $\mathcal{Y}^{(-)}$ is the set of negative links. For document i, the content can be represented by $w_i = [w_{i1}, ..., w_{iD_i}] \in \{0,1\}^{D_i \times V}$, where $w_{id} \in \{0,1\}^V$ is the d-th word of document i represented by a one hot vector, D_i is the length of the document and V is the vocabulary size of the corpora shared by all documents.

We have two tasks to consider: First, we aim at discovering latent topics from document contents; second, we seek to recommend relevant citations based on collaborative filtering of historical references among documents. We first show how classical methods tackle the two tasks collectively, then we introduce mean filed and neural variational inference, the latter of which acts as the building block to extend classical models to our proposed method.

2.1 Relational Topic Model

Relational Topic Model (RTM) [6] is a traditional Bayesian latent variable model for joint learning of topics and citation networks. RTM unifies Latent Dirichlet Allocation (LDA) [3] of topic learning with matrix factorization of network modelling. Specifically, assume there are T topics, and $\beta \in \mathrm{R}^{T \times V}$ is the topic-word matrix where each row $\beta_t \in \mathrm{R}^V$ is the word distribution of topic t lying on a simplex. For document i, $\theta_i \in \mathrm{R}^T$ is the topic latent variable, while $z_{id} \in \mathrm{R}^T$ is the one-hot encoded topic indicator variable for the d-th word in document i.

The generative process of RTM is as follows:

(1) For each document i:
  (a) Draw the topic variable $\theta_i | \alpha \sim \mathrm{Dir}(\alpha)$;
  (b) For each word w_{id} of the document:
    (i) Draw the topic indicator variable $z_{id} | \theta_i \sim \mathrm{Multi}(\theta_i)$;
    (ii) Draw the word $w_{id} | z_{id}, \beta \sim \mathrm{Multi}(\beta_{z_{id}})$;
(2) For each observed link y_{ij} between document i and j:
  (a) Draw $y_{ij} | z_i, z_j \sim \mathrm{Bernoulli}(\phi(\eta^\top z_i \circ z_j + \nu))$;

where $z_i = \frac{1}{D_i} \sum_d z_{id}$, and $\phi(\cdot)$ is the function parameterized by η and ν for the Bernoulli distribution. The prediction of links comes from bilinear matrix factorization with a nonlinear transformation $\phi(\cdot)$. Nevertheless, the variational inference in RTM is non-trivial: the analytical update of variational parameters highly depends on the choice of $\phi(\cdot)$ (e.g., probit function or sigmoid function). Furthermore, the shallow structure of the model may not be able to capture the underlying coherence of the two tasks.

2.2 Mean Field and Neural Variational Inference

Variational inference techniques are famous approximation tools for probabilistic graphical models. Consider a probabilistic model $p_\lambda(x, \theta)$ with data x, latent variables θ and parameters λ, we seek to find the posterior $p_\lambda(\theta|x) = p_\lambda(x, \theta)/p(\theta)$ according to Bayes rules. Most posteriors are intractable, and a substitute way is to by minimizing the Kullback-Leibler (KL) divergence $\mathrm{KL}(q_\gamma(\theta) \| p_\lambda(\theta))$ between approximated posterior $q_\gamma(\theta)$ parameterized by γ and true posterior $p_\lambda(\theta|x)$. However, the mean-filed [32] assumption in VI highly restricts the flexibility of approximated posterior. Meanwhile, the parameter update usually requires tricky approximations (as appeared in RTM), which may not be easily implemented in practical applications.

Recently proposed Neural Variational Inference (NVI) [15] addresses the above challenges elegantly with the help of Variational Auto-Encoders (VAE). Specifically, NVI learns variational posterior $q_\gamma(\theta|x)$ and the observation likelihood $p_\lambda(x|\theta)$ based on the encoding and decoding process of VAE. The reparameterization trick in VAE allows the gradient of variational parameters to back-propagate through the random variables [15, 28]. For example, a univariate Gaussian distributed variable $\theta \sim \mathcal{N}(\mu, \sigma^2 \mathrm{I})$) can be reparameterized via a location-scale transformation, i.e., $\theta = \mu + \epsilon \cdot \sigma$, where $\epsilon \sim \mathcal{N}(0, \mathrm{I})$. where ϵ is a randomly injected noise, and parameters μ and σ are obtained via the encoder parameterized by γ. NVI turns to optimize the Evidence Lower Bound (ELBO) as follows:

$$\mathcal{L}(\lambda, \gamma; x) = \mathrm{E}[\log p_\lambda(x|\theta)] - \mathrm{KL}(q_\gamma(\theta) \| p_\lambda(\theta)), \tag{1}$$

where the first term is the reconstruction loss, and the KL-divergence is as described before. NVI allows a flexible posterior approximation without the mean-field assumption, and with the back-propagation of gradients, the update of variational parameters can be done in a black-box way.

3 NEURAL RELATIONAL TOPIC MODEL

In this section, we first show how we enhance NVI to learn latent topics from document contents and develop a neural topic model. Then we demonstrate how we resort to multilayer perceptions to capture pairwise interactions among topics, which are then utilized for citation recommendation. The architecture of our model is presented in Figure 1, in which two stacked variational auto-encoders are deployed to learn latent topics for two documents x_i and x_j respectively, and their latent topics θ_i and θ_j are concatenated and fed into a multilayer perception to output the predicted citation link $\tilde{y}_{ij}$.

3.1 Stacked Variational Auto-Encoders for Neural Topic Modelling

Previous neural topic models [5, 25, 30] generate document contents with only one-layer decoder which may not fully take the advantages of deep structure of auto-encoders. In order to obtain richer representations for the latent topics, we apply a stacked architecture to the variational auto-encoder as presented in Figure 1, which we name as the Stacked Variational Auto-Encoder (SVAE). With a similar spirit to stacked auto encoder [2], the greedy layer-wise approach to pretrain the SVAE is adopted to help model learn more informative features of document contents. Nevertheless, to fully integrate topic models into the Bayesian framework, the following issues need to be addressed:

Reparameterization of neural topic models As described in Section 2.2, a key point that allows the back propagation of variational parameters is the reparameterization trick. In neural topic modelling, we have two set of latent variables: topic indicator variable z and topic variable θ. Since z is discrete, its reparameterization could be troublesome in nature. Thanks to the conjugate Dirichlet-Multinomial distributions, we can analytically sum out z as

$$\begin{aligned} p(w_{id}|\theta_i, \beta) &= \sum_{z_{id}=1}^{T} p(w_{id}|z_{id}, \beta) p(z_{id}|\theta_i) \\ &= \sum_{z_{id}=1}^{T} \mathrm{Multi}(\beta_{z_{id}}) \mathrm{Dir}(\alpha) = \mathrm{Multi}(\theta_i \beta), \end{aligned} \tag{2}$$

leading to a topic-document model [30].

However, the reparameterization for the Dirichlet distributed topic variables θ is more challenging. To deal with this, we choose Normal distribution as a substitute to Dirichlet distribution, i.e., for document i,

$$\theta_i | w_i \sim \mathrm{Dir}(\alpha_i) \quad \Longrightarrow \quad \theta_i | w_i \sim \mathcal{N}(\mu_i, \sigma_i \mathrm{I}), \tag{3}$$

and the reparameterization can be performed via the location-scale transformation as described in Section 2.2. Note that unlike LDA, we do not impose simplex constraints on topic variables θ, but allow them to be fully inferred by deep neural networks. Although one can assign softmax transformation over the topic variables θ to ensure the constraint, such design may restrict the model capacity and affect the performance of citation recommendation[1].

Representation of the topic-word matrix β The representation of β is a key factor for the quality of topic coherence. Instead of linearly generating the document contents based on β in LDA, in NTMs it is done in a nonlinear manner via the decoding process of VAE, where β is interpreted as the weights of the decoder. Previous methods such as NVDM [25] and ProdLDA [30] only deploy one layer on the decoder which may suffer from limited expressiveness of the structure. Different from them, in our model multiple hidden layers are stacked for the decoder, hence β can be represented by the layer-wise product of decoder weights, i.e., $\beta = \prod_{l=L/2}^{L} W^l$, where L is the number of layers of the auto-encoder. With such design, the learning of topic-word matrix β can be fully incorporated into backpropagation updates of the neural structure.

According to Equation 1, the ELBO of our neural topic model is formulated as:

$$\mathcal{L}_1(w_i) = \frac{1}{S} \sum_{s=1}^{S} \sum_{d}^{D_i} \ln p_\lambda(w_{id}|\theta_i^{(s)}, \beta) - \mathrm{KL}(q_\gamma(\theta_i) \| p_\lambda(\theta_i)), \tag{4}$$

in which the reconstruction loss and the KL-divergence are:

$$\ln p_\lambda(w_{id}|\theta_i^{(s)}, \beta) = \quad w_{id} \ln \tilde{w}_{id} + (1 - w_{id}) \ln(1 - \tilde{w}_{id}), \tag{5}$$

$$\mathrm{KL}(q_\gamma(\theta_i) \| p_\lambda(\theta_i)) = \quad \frac{1}{2} \sum_{t=1}^{T} (\mu_{it}^2 + \sigma_{it}^2 - (1 + \ln(\sigma_{it}^2))), \tag{6}$$

respectively. Here $\tilde{w}_{id} = f(w_{id}, \lambda)$ is the recovered word for w_{id} by the SVAE. Since SGVB estimator is generally unbiased with low variance comparing to other black-box inference methods, e.g., BBVI [27], a common practice is to set the number of samples $S = 1$ for the simplification of computation.

[1]We conducted experiments and found topic variables with softmax transformation generally decrease the performance on both topic learning and citation recommendation.

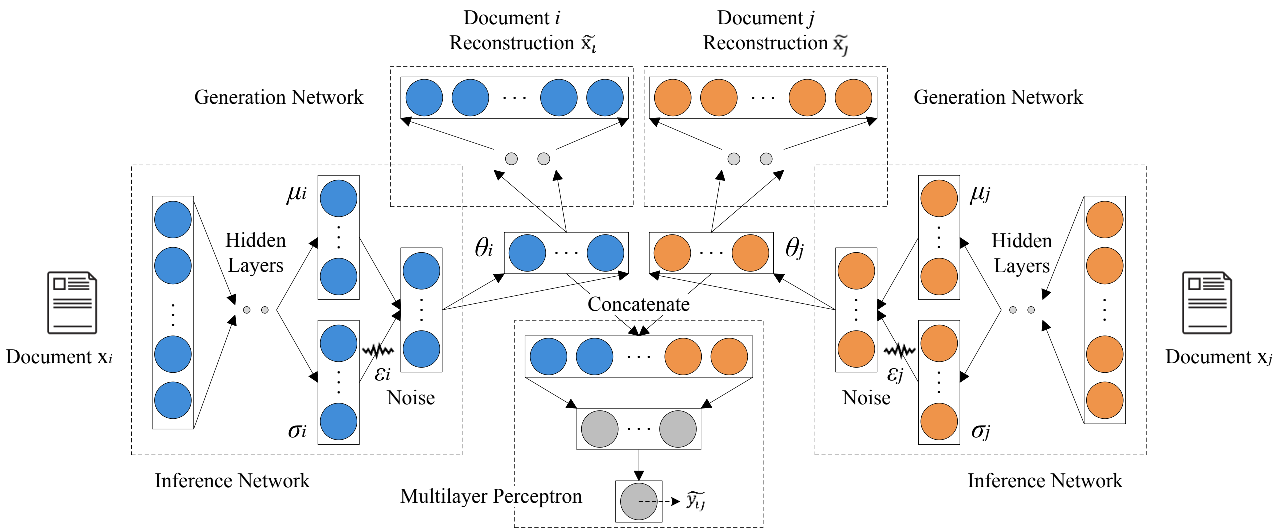


Figure 1: The neural architecture of the proposed neural variational topic model. The document contents x_i and x_j are forwarded in the stacked variational auto-encoders and their topic variables θ_i and θ_j are generated. Then, the topic variables θ_i and θ_j are concatenated and fed into the multilayer perception to predict the citation $\tilde{y}_{ij}$.

3.2 MLP: Combining Topic Modelling with Collaborative Filtering

In collaborative filtering, when it comes to modelling the interaction between two documents, previous works still rely on matrix factorization and simply apply inner product on the latent features. In order to efficiently fuse the information from document contents and citation networks, we adopt multilayer perceptions to replace the inner product and capture the complex interactions among latent topics. Specifically, for each pair of documents (i, j), we first concatenate the latent representation $\theta = [\theta_i, \theta_j]$ as the input to the MLP, and perform forward propagation to generate $\tilde{y}_{ij}$ which indicates the probability that document i cites j, as shown in Figure 1. As in RTM, we assume each y_{ij} is Bernoulli distributed, so the log likelihood function for y_{ij} can be written as

$$\begin{aligned}\mathcal{L}_2 = & \ \mathrm{E}[\log p(y_{ij}|\theta_i, \theta_j)] \\ = & \sum_{(i,j)\in\mathcal{Y}^{(+)}} \log \tilde{y_{ij}} + \sum_{(i,k)\in\mathcal{Y}^{(-)}} \log(1-\tilde{y_{ij}}),\end{aligned} \tag{7}$$

where $\tilde{y}_{ij} = f_\eta(\theta_i, \theta_j)$, and $f_\eta(\theta_i, \theta_j)$ is the MLP approximator parameterized by η. To ensure $f_\eta(\theta_i, \theta_j) \in (0, 1)$, we apply the sigmoid activation function to the output layer.

The MLP architecture can simultaneously benefit topic learning and citation recommendation. First, the concatenated topic variable θ conveys the representations of documents to the collaborative filtering of citations. Second, the back propagation of gradients can transfer the knowledge of the citation network to update latent representations of topics. Therefore, the MLP architecture bridges topic modelling and collaborative filtering in a fully deep way which can better leverage the coherence between the two tasks compared to previous MF-based approaches [6, 35, 36].

Combining MLP with SVAE together, the full ELBO for NRTM can be formulated as:

$$\begin{aligned}\mathcal{L} = & \underbrace{\frac{1}{S}\sum_{s=1}^{S}\sum_{d}^{D_i} \ln p_\lambda(w_{id}|\theta_i^{(s)}, \beta) - \mathrm{KL}(q_\gamma(\theta_i)\|p_\lambda(\theta_i))}_{\mathcal{L}_1} \\ & + \underbrace{\frac{\alpha}{S}\sum_{s=1}^{S} \log p(y_{ij}|\theta_i^{(s)}, \theta_j^{(s)})}_{\mathcal{L}_2} + \underbrace{\beta(\sum_{l=1}^{L}\|W^{(l)}\|_2^2 + \sum_{l'=1}^{L'}\|W^{(l')}\|_2^2)}_{\mathcal{L}_3},\end{aligned} \tag{8}$$

where $\mathcal{L}_1$ is the ELBO for topic modelling, $\mathcal{L}_2$ is the log likelihood for collaborative filtering, and L' is number of layers in MLP. To prevent over-fitting, we also assign l_2 norms over the weights of SVAE and MLP denoted as $\mathcal{L}_3$. α and β are regularizers for $\mathcal{L}_2$ and $\mathcal{L}_3$ respectively. The SAVE and MLP can be learned jointly to minimize the ELBO in Equation 8.

Overall, the generative process formulated in the neural framework of our model can be summarized as follows:

(1) For each document i:
  (a) Input the word vector w_{id};
  (b) Forward calculation of μ_i and σ_i via the encoder;
  (c) Draw the noise $\epsilon_i \sim \mathcal{N}(0, \mathrm{I})$ and compute $\theta_i = \mu_i + \sigma_i\epsilon_i$;
  (d) Recover the word $\tilde{w}_{id}$ via the decoder;
(2) For each observed citation y_{ij} between document i and j:
  (a) Draw $y_{ij}|\theta_i, \theta_j \sim \mathrm{Bernoulli}(f_\eta(\theta_i, \theta_j))$;

Aside from MLP, our model can be easily extended to other neural architectures. For example, convolutional neural networks can be performed on the outer product of θ_i and θ_j so as to model the pairwise correlations of latent topic spaces [10]. We leave them as future work.

Table 1: Statistics of real-world datasets.

Datasets	Documents	Vocab. Size	Citations	Sparsity
Citeulike-a	16,980	8,000	44,709	99.974%
Cora	13,147	17,059	57,018	99.967%

3.3 Implementation

Efficient Computation of $\mathcal{L}$ Since each document contains thousands of words, sequentially passing each word w_{id} of document i and sample *epsilon* to calculate $\mathcal{L}$ is computationally expensive. With the exchangeability of w_{id}, for $d = 1, ..., D_i$, a more efficient way is to equivalently represent the document content as $x_i = \sum_{d=1}^{D_i} w_{id}$. Moreover, since z_{id} has been summed out in Equation 2, we can pass x_i directly to SVAE and sample ϵ once for the entire content of document i. In order to meet the input requirements of sigmoid cross entropy function in Equation 5, we further normalize x_i as is done in [35].

Negative Sampling For each pair of positive link y_{ij}, we uniformly select N_{neg} negative links associated with document i which indicate that i does not cite those articles. With negative samples, the training of the network can acquire both positive and negative feedback from the observations, leading to a more comprehensive representation for the documents. In practice we find that negative sampling can contribute to better performance on both topic modelling and collaborative filtering. Other studies show that non-uniform sampling [12] may further improve the performance of the model.

4 EXPERIMENT

In this section, we present the experiments of NRTM. We evaluate NRTM for citation recommendation and topic learning separately, and then provide some qualitative analysis. The code is available at https://github.com/zbchern/Neural-Relational-Topic-Models.

4.1 Dataset

We evaluate our model on two real world datasets: Citeulike-a[2] and Cora [23]. Citeulike-a is from an online repository that allows users to create their own collections of papers, and Cora is another widely used dataset for citations of scientific papers. For both datasets, we preprocess the data in a similar way to [33, 35], i.e., lemmatize redundant terms, remove infrequent words and delete documents with less than 10 citations. We concatenate the title and abstract as the content for each article. Note that for a fair comparison, we follow [35] to convert all links to undirected links, although our method can also be directly applied to directed networks.

The statistics of the proprocessed datasets are summarized in Table 1. It can be observed that both datasets are extremely sparse. For each document, we randomly select 80% of the observed citations for training, and the remaining 20% citations for testing. For each positive instance in the training set, we randomly draw N_{neg} negative samples. Since all documents appear in both training and testing set, collaborative filtering can be carried out effectively.

[2] http://www.citeulike.org

Table 2: Summarization of the baselines.

Methods	Topic Modelling	Recommendation
LDA	PGM	/
RTM	PGM	MF
ProdLDA	Deep	/
ProdLDA+MF	Deep	MF
NCF	/	Deep
RDL	Deep	MF
SVAE	Deep	/
NRTM	Deep	Deep

4.2 Experiment Setup

Baselines The baselines listed below are highly related to our model:

- **LDA** [3] is a well-known Bayesian topic model.
- **RTM** [6] is a classical method that combines LDA with MF to jointly learn document topics which also generates recommendations for documents.
- **ProdLDA** [30] is the state-of-the art neural topic model that can learn coherent topic words with efficient computational estimations for parameters.
- **ProdLDA+MF** initializes the low rank latent factors of matrices by topic embeddings learned by ProdLDA to perform collaborative filtering on the citation network data.
- **NCF** [11] is the state-of-the art method that performs collaborative filtering via neural network. However, NCF does not leverage document contents as auxiliary information for citation recommendation.
- **RDL** [35] is the state-of-the-art auto-encoder based method for network modelling. The prediction of links is similar to that in RTM. Although RDL is not originally proposed for learning topics, we adopt the same methodologies as used in RTM to yield a topic-word matrix β to measure the topic quality.

Table 2 is a summarization of the baselines in terms of the approach used for different tasks, where SVAE and NRTM are the proposed methods in this paper. "PGM" denotes probabilistic graphical models, "MF" denotes matrix factorization, "Deep" denotes deep neural network based approaches and "/" indicates the incapability on the task.

Evaluation Metrics For citation recommendation, we report two measures of ranking quality: *Hit Ratio* (HR) and *Normalized Discounted Cumulative Gain* (NDCG). HR@K indicates whether the target is in the top-K of the recommendation list, while NDCG@K is accumulated from the top of the result top-K list to the bottom, with the gain of each result discounted at lower ranks. We adopt the leave-one-out evaluation scheme, and randomly draw 99 negative samples that are not cited by the document i for each positive link y_{ij}. All experiments are repeated for 5 times and the average values of HR and NDCG scores are reported.

For topic modelling, we report the C_V score [29] computed by Palmetto toolkit[3]. The C_V score is to measure how close the meanings of words are clustered in the topic, and is shown to be more consistent to human judgment compared to other widely used metrics such as PMI and NPMI [16]. We report the C_V score averaged over all topics. Aside from quantitative evaluation of topic modelling, we also present some top-K citation recommendations for each topic, as well as some cases of the topical indications of the document contexts.

Parameter Settings For LDA, we use the internal packages provided by scikit-learn[4], and adopt the variational inference for parameter estimation. For RTM, we search through the concentration parameter α of Dirichlet distribution and find $\alpha = 0.1$ can provide robust and good results across the two datasets. We implement ProdLDA based on the released code[5] of [30], and keep the network architecture unchanged. For ProdLDA+MF, we first pretrain the ProdLDA model, and then train the neural network with MF jointly with the regularizer for the MF loss set as 1000. In terms of RDL, we set the network structure as *V-200-100-T-100-200-V*. Recall that V and T are the vocabulary size and the number of topics respectively. The regularizers in RDL are set to yield its best performance, i.e., $\lambda_p = 0.1, \lambda_e = 50, \lambda_w = 0.00001, \lambda_s = 0.1$ and λ_n. For NRTM, we set the hidden layers of SVAE as the same as those in RDL, while for MLP, we assign 4 fully connected layers (T-$\frac{T}{2}$-$\frac{T}{4}$-1) as default. We also vary the layer depth of both SVAE and MLP to verify the effectiveness of our deep neural structure in Section 4.5. We set $\alpha = 1000, \beta = 0.0001$ for all experiments. Before we jointly train the SVAE and the MLP, we first pretrain the network layer-wisely for SVAE as described in Section 3.1.

We use the Adam optimizer with the learning rate of 0.001. For negative sampling, we consistently draw 5 negative links for each positive link for all network based approaches. We also report the effect of the number of negative samples in Section 4.5.

4.3 Citation Recommendation

The HR and NDCG scores are reported in Figure 2, according to which we have the following observations:

- NRTM generally outperforms all the baselines by a large margin, especially when K is large, which demonstrates the superiority of NRTM in citation recommendation. The performance of NRTM indicates that our model can indeed effectively fuse the information of latent topics to help the collaborative filtering for citations, even though the citation network is extremely sparse.
- RDL underperforms NRTM in most cases, which demonstrates that the matrix factorization may not be effective for the collaborative filtering for citations. Nevertheless, since RDL is also a deep auto-encoder based approach, the learned document embeddings could still benefit the collaborative filtering to a great extend compared to ProdLDA+MF and RTM.
- NCF also achieves comparable performance to RDL, even though NCF only utilizes the network structure. This indicates that the neural collaborative filtering framework can effectively learn the pairwise interactions of document embeddings.
- ProdLDA+MF does not show competitive performance compared to NRTM, RDL and NCF, which suggests that the latent topics inferred by the in ProdLDA cannot benefit collaborative filtering. This phenomenon can be attributed to that one-layer decoder prevents it from learning informative representations from documents.
- RTM has the lowest scores among all methods on both datasets. The phenomenon can be explained by its shallow structure which may not be able to fully learn the document embeddings and fuse them into the collaborative filtering task.

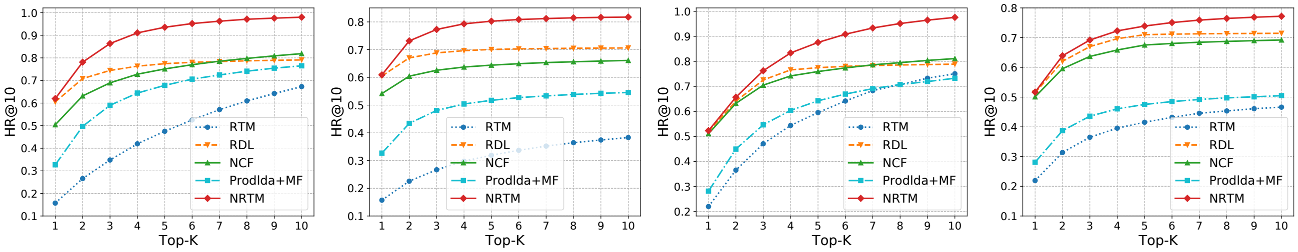
To further verify the effectiveness of our model, we vary the number of topics from 10 to 50 with an interval of 10, and corresponding results of HR@10 and NDCG@10 are shown in Figure 3. It can be observed that our NRTM achieves the highest scores for different number of topics on both datasets, and it is not very sensitive to the number of topics. For instance, NRTM with 20 topics is enough to achieve competitive results on both datasets. The excellent performance and robustness of our model indicate that the neural architecture grants NRTM with the superior learning ability even with small number of latent dimensions. RDL again ranks in the second place, while it can be more easily affected by the number of topics. NCF is still competitive to RDL, which demonstrates the power of neural network for collaborative filtering. ProdLDA+MF outperformes RTM by a large margin on Citeulike-a, but is slightly better on Cora. Overly, RTM ranks the last and is very sensitive to the choice of the topic number.

4.4 Topic Modelling

Next, we present the results of topic modelling. The C_V scores are shown in Table 3. We can observe that:

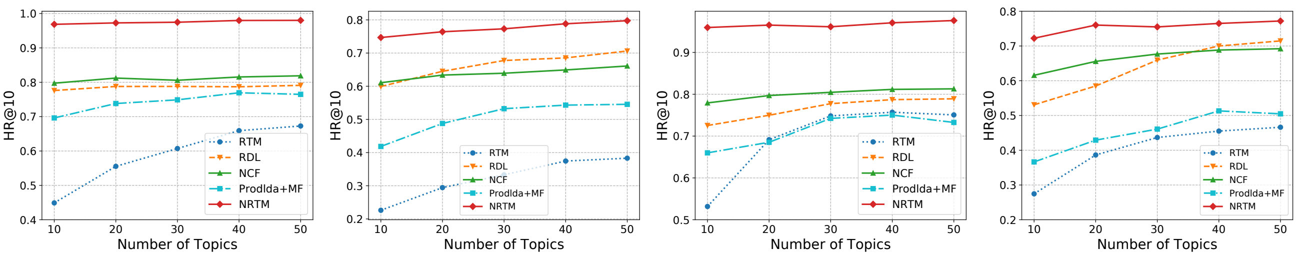
- Our NRTM achieves the highest C_V scores with both 20 and 50 topics on two datasets. Compared to SVAE, there is an improvement of the coherence score, especially on Cora, which verifies NRTM can indeed effectively utilizes the information from the citation network to learn topics with better qualities.
- Even without the network knowledge, SVAE also shows competitive performance to joint learning methods such as RDL and RTM, especially on Citeulike-a. Compared to ProdLDA, the higher C_V scores of SVAE indicate the advantages of its stacked architecture over the one-layer decoder in ProdLDA.
- ProdLDA+MF even performs worse than the ProdLDA on both datasets, which is beyond our expectation. We attribute the decrease of C_V score to the incomplete representation of latent topics in ProdLDA, which may not effectively leverage the knowledge from the citation network, and vice versa as illustrated in Figure 2 and Figure 3.
- In comparison with LDA, RTM only outperforms it on Citeulike-a with 50 topics. On the other hand, LDA even achieves the

[3]https://github.com/dice-group/Palmetto
[4]http://scikit-learn.org
[5]https://github.com/akashgit/autoencoding_vi_for_topic_models



(a) HR@K on Citeulike-a (b) NDCG@K on Citeulike-a (c) HR@K on Cora (d) NDCG@K on Cora

Figure 2: Performance of HR@K and NDCG@K on the two datasets with K ranging from 1 to 10.



(a) HR@10 on Citeulike-a (b) NDCG@10 on Citeulike-a (c) HR@10 on Cora (d) NDCG@10 on Cora

Figure 3: Performance of HR@10 and NDCG@10 with number of topics varying from 10 to 50.

Table 3: C_V scores.

	Citeulike-a		Cora	
	20 topics	50 topics	20 topics	50 topics
LDA	0.4081	0.4308	0.4183	0.4308
RTM	0.4501	0.4267	0.3902	0.3844
ProdLDA	0.4509	0.423	0.4295	0.3927
ProdLDA+MF	0.4095	0.3315	0.3997	0.3757
RDL	0.4857	0.4632	0.4663	0.4516
SVAE	0.4957	0.4819	0.4285	0.4084
NRTM	**0.5104**	**0.5028**	**0.4762**	**0.4833**

highest C_V scores among content-only based methods, indicating it as a competitive method in topic modeling.

Qualitative Study To further analyze the topic coherence, we show the top-10 words concerning the topic "gene" generated by different approaches as shown in Table 4. We can see that despite off-topic word "motif" generated by NRTM, our method can still produce more coherent words for the topic as indicated by the C_V score. SVAE realizes the highest coherence score among the methods that do not consider the network information. Other approaches such as LDA produce more irrelevant words like "regulatory", "regulation", leading to the drop of the C_V score.

In Table 5, we also demonstrate the top 3 topics and top 5 recommended citations associated with the most frequently cited article in the testing set of Citeulike-a. Due to the limited space, we only report the results obtained by NRTM and RDL. It can be observed that topics produced by NRTM generally covers "gene" (topic 1), "rna" (topic 2) and "bioinformation" (topic 3), all of which are related to the document intuitively as revealed by the title. Besides, the coherence score also shows that top words within each topic are highly correlated. RDL also discovers two closely related topics (topic 1 and topic 2), but topic 3 is less related to the document. Furthermore, the C_V scores of RDL also show that the words are less correlated compared to those produced by NRTM. In terms of the recommended citations. NRTM hits 5 out of 5 ground truth citations with an accuracy of 100%, while RDL only hits 2.

4.5 Further Analysis

In this subsection, we further evaluate the NRTM by considering the following two questions:

i) Are deep neural networks helpful for leveraging the coherence of the two tasks?

To demonstrate the effectiveness of the deep neural architecture, we consider varying the number of layers for both SVAE and MLP. Table 6 shows the results for varying layers in SVAE. Note that we keep the number of topics as 50, and choose a four layer MLP. SVAE-1, SVAE-2 and SVAE-3 denote the encoders with structures V-100-50, V-200-100-50, and V-400-200-100-50 respectively, and decoders are constructed in a symmetric way. It can be observed that with 2 hidden layers, the model achieves the best performance regarding to both topic learning and citation recommendation. However, SVAE-1 and SVAE-3 show less competitive results. We attribute the drop of SVAE-1 to the limited expressiveness of the structure, and SVAE-3 to the over-fitting of document contents.

Table 4: Top 10 words for topic "gene".

Models	Top-10 words	*CV* Score
LDA	gene, expression, genes, cell, cells, biological, expressive, regulatory, regulation, function	0.4865
RTM	human, selection, genetic, evolutionary, genes, evolution, genome, disease, sequence, rate	0.4960
ProdLDA	genome, approximately, genomic, brain, regions, proteins, neurons, gene, evolutionary, noncoding	0.4923
ProdLDA+MF	gene, mirna, genes, development, genomes, sequence, genomic, context, sequences, needs	0.4879
RDL	sequential, genome, genes, human, evolution, evolutionary, dna, genomes, identify, functional	0.5367
SVAE	gene, evolutionary, metabolic, genomescale, evolution, motifs, genomic, transcription, cortex, transcriptional	0.5067
NRTM	gene, phylogenetic, genes, motifs, protein, proteins, transcription, transcriptional, motif, genomes	**0.6232**

Table 5: Top 10 recommended citations.

Document entitled "Cytoscape: a software environment for integrated models of biomolecular interaction networks"	**NRTM**
Top 3 topics	*CV* **Score**
1. genes, gene, protein, rna, genomic, genome, yeast, genomes, dna, mrna	0.7532
2. rna, protein, transcription, microbial, genomes, microarray, pfam, proteins, phylogenetic, eukaryotic	0.6278
3. bioinformatics, databases, habitat, user, functionality, metadata, photographs, communities, indoor, opensource	0.4868
Top 5 documents	**cite or not?**
1. The Biomolecular Interaction Network Database and related tools 2005 update	**yes**
2. Probabilistic model of the human protein-protein interaction network	**yes**
3. Transcriptional Regulatory Networks in Saccharomyces cerevisiae	**yes**
4. The Gaggle: An open-source software system for integrating bioinformatics software and data sources	**yes**
5. Understanding biological functions through molecular networks	**yes**
Document entitled "Cytoscape: a software environment for integrated models of biomolecular interaction networks"	**RDL**
Top 3 topics	*CV* **Score**
1. rna, protein, transcription, microarray, microbial, genomes, proteins, eukaryotic, pfam, genes	0.6505
2. protein, proteins, rna, cancer, chipseq, genome, genomes, sequencing, genomic, genes	0.5493
3. kernels, queries, basal, retrieval, tasks, software, phylogenetic, swarm, institutions, memory	0.4154
Top 5 documents	**cite or not?**
1. Predicting transcription factor binding sites using local overrepresentation and comparative genomics	no
2. Pathway analysis using random forests classification and regression	**yes**
3. Broad network-based predictability of Saccharomyces cerevisiae gene loss-of-function phenotypes	**yes**
4. Comparative protein modelling by satisfaction of spatial restraints	no
5. Fast and reliable prediction of noncoding RNAs	no

In terms of varying the layers of MLP, we keep the number of topics to 50, and adopt the SVAE-2 structure. The results are shown in Table 7, where MLP-1 represents a direct output layer, while MLP-n denotes $n-1$ hidden layers plus the output layer. It can be observed that with only one layer for the MLP, the performance significantly decreases. However, as the number of layer increases, the scores of HR, NDCG as well as topic coherence increase in general. Although the performance of MLP-2 is slightly better than that of MLP-3 on Citeulike-a, the best results for different scores are all attained at MLP-4, while for MLP-5 the performance is also competitive. Actually when we take more than 6 layers for the MLP, the performance saturates or even drops slightly due to over fitting, and we omit the detailed performance due to the limited page. From Table 7, we know that with proper layer depth for neural structures, the correlations between latent topics and citation networks can be well fused and utilized, leading to a better performance on both tasks.

ii) How does the number of negative samples affect the performance of NRTM?

We now investigate the effect of negative sampling on the performance of the model. Table 8 lists the experimental results with varying number of negative samples N_{neg} from 1 to 9 at an interval of 2. It can be observed that, the best performance of citation recommendation can be achieved at 5 and 7 negative samples respectively on the two datasets, while more negative samples lead to a longer training time with little increase on HR and NDCG scores. In terms of the C_V score, we find an intermediate number (i.e., 3, 5 and 7) of negative samples on the two datasets achieves the highest topic coherence. However, choosing only 13 negative samples generally lead to an apparent drop on both the ranking quality and the topic

Table 6: Performance of NRTM w.r.t. different number of layers for SVAE.

Datasets	SVAE-1	SVAE-2	SVAE-3
HR@10			
Citeulike-a	0.9695	**0.9801**	0.9614
Cora	0.6140	**0.9762**	0.9582
NDCG@10			
Citeulike-a	0.7519	**0.8175**	0.7264
Cora	0.7392	**0.7719**	0.7102
***CV*-Score**			
Citeulike-a	0.4768	**0.5028**	0.4893
Cora	0.4698	**0.4833**	0.4801

Table 7: Performance of NRTM w.r.t. different number of layers for MLP.

Datasets	MLP-1	MLP-2	MLP-3	MLP-4	MLP-5
HR@10					
Citeulike-a	0.4896	0.9681	0.9743	**0.9801**	0.9746
Cora	0.4208	0.8860	0.9642	**0.9762**	0.9601
NDCG@10					
Citeulike-a	0.3075	0.7489	0.7565	**0.8175**	0.8106
Cora	0.2681	0.6995	0.7294	**0.7719**	0.7683
***CV*-Score**					
Citeulike-a	0.4532	0.4867	0.4921	**0.5028**	0.5016
Cora	0.4456	0.4739	0.4704	**0.4833**	0.4821

Table 8: Performance of NRTM w.r.t. different number of negative samples with 50 topics.

Datasets	$N_{neg} = 1$	$N_{neg} = 3$	$N_{neg} = 5$	$N_{neg} = 7$	$N_{neg} = 9$
HR@10					
Citeulike-a	0.9626	0.9731	0.9793	**0.9801**	0.9798
Cora	0.9539	0.9571	0.9679	**0.9762**	0.9726
NDCG@10					
Citeulike-a	0.7110	0.7512	0.7871	**0.8175**	0.7975
Cora	0.6987	0.7097	0.7598	0.7689	**0.7714**
***CV*-Score**					
Citeulike-a	0.4793	**0.5081**	0.5028	0.5074	0.4936
Cora	0.4591	0.4684	**0.4850**	0.4732	0.4833

coherence scores, indicating the necessity to bring more negative feedback of citations for training.

5 RELATED WORK

5.1 Topic Modeling with Relational Data

An amount of Bayesian latent variable models are proposed to learn the complementary information from latent topics and citation networks. Aside from the Relational Topic Model [6] described in Section 2.1, Discriminate RTM (gRTM) [8] is a further extension of RTM that aims at increasing the model capacity by learning a full interaction matrix as well as the inference accuracy by Gibbs sampling. Topic-Link LDA [20] collectively considers the topic similarity and the community closeness to the generative model. LTAI [14] further incorporates the information of authors to the generation of citations, leading to a more integrative model.

Despite the success of these models in joint modeling of topics and relational data, they either suffer from limited model expressiveness and tricky derivations of variational inference, both of which prevent them from being widely applied in practical problems. Comparing to these methods, our proposed NRTM could learn more informative representation for topics via the stacked variational auto-encoder, and can be efficiently inferred in a black-box way via the SGVB [15] method.

5.2 Neural Topic Modelling

Our NRTM is also closely related to neural topic models [4, 5, 9, 24, 25, 30, 31, 41]. Among these models, an interesting approach is to adopt neural variational inference [15] to formulate topic modelling as Bayesian deep generative models [5, 24, 25, 30].

In particular, Neural Variational Document Model (NVDM) [22] is a bag-of-words generative model that leverages Variational Auto-Encoders [15] to learn latent representations for documents, through which the topic-word assignment can be inferred by the weights of decoders. The recent proposed ProdLDA [30] uses product of experts to generate the words in replacement of the mixture assumption at word-level in LDA, and leads to a significant improvement on the coherence scores of topics. More recent works [24] extend neural topic models to Bayesian non-parametric settings, and also achieve competitive coherence scores compared to ProdLDA.

Inheriting advantages of VAE, these models can be inferred in a black-box way and is capable of learning deeper representations for topics. Nevertheless, relational data, as an informative source to complement the document content, has not been considered in previous neural topic models. Therefore, in NRTM we consider the joint learning of article topics and citation networks.

5.3 Collaborative Filtering with Auxiliary Information

Citation recommendation is also a traditional task in recommender systems, in which the document contents act as auxiliary information. Due to the sparsity of network data, the latent representations of auxiliary information can provide suggestive insights for valid recommendations. Matrix factorization based models [1, 13, 21, 37, 38, 40, 42] can be useful tools to learn representations from network data. In order to further improve the expressiveness of these models, we resort to deep neural networks. Auto-encoders, owing to their intrinsic structures to compress data, have been widely explored to learn representations for the auxiliary knowledge [17, 18, 34–36]. Deep Collaborative Filtering [17] combines marginalized denoising auto-encoder (mDA) [7] with probabilistic matrix factorization (PMF) to learn the representations of side information. Collaborative Deep Learning (CDL) [36] is proposed as a hierarchical Bayesian neural network to jointly learn the document features as well as the implicit feedback. Based on CDL, RDL [35] inherits a similar idea but targets at link prediction problems.

The above approaches are closely related to our NRTM since auto-encoders are both adopted to learn representations. Nevertheless, we further deploy MLPs instead of MF to learn the complex interaction of latent topics, which has not been explored to the best of our knowledge.

Other approaches [11, 19, 39] seek to extend collaborative filtering via neural networks. For instance, Neural Collaborative Filtering [11] passes the latent factors of users/items through multilayer perceptions and achieves significant better performance than MF based methods. However, these methods seldom consider capturing pairwise interactions of latent factors enriched by the auxiliary information as is done in NRTM.

6 CONCLUSION

In this paper, we propose the Neural Relational Topic Model (NRTM), a Bayesian deep generative model to mutually leverage the correlation between latent topics and citation networks. We design the stacked variational auto-encoders, which enjoys two advantages: First, the stacked architecture brings more representative latent topics; second, SVAE is equipped with neural variational inference, and the posterior approximation of latent topics can be done in a black box manner. To further improve the expressiveness of our model, we adopt multilayer perceptions to mutually leverage the correlation between latent topics and citation networks. Experimental results on two real world datasets demonstrate that our model can effectively take advantage of the underlying correlation, and outperform the state-of-the-art methods on both tasks.

One promising direction for the future work is to adopt other neural architectures to further increase the model capacity. Another direction is to explore different approximated distributions over the latent topics so that the model enjoys better probabilistic explanations as a Bayesian deep generative model.

ACKNOWLEDGMENTS

The work described in this paper was partially supported by the Research Grants Council of the Hong Kong Special Administrative Region, China (No. CUHK 14208815 and No. CUHK 14234416 of the General Research Fund), Microsoft Research Asia (2018 Microsoft Research Asia Collaborative Research Award), and Natural Science Foundation of China (No. 61572111).

REFERENCES

[1] Haoli Bai, Zenglin Xu, Bin Liu, and Yingming Li. 2016. Hierarchical Probabilistic Matrix Factorization with Network Topology for Multi-relational Social Network. In *ACML*. 270–285.

[2] Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. 2007. Greedy layer-wise training of deep networks. In *NIPS*. 153–160.

[3] David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3, Jan (2003), 993–1022.

[4] Ziqiang Cao, Sujian Li, Yang Liu, Wenjie Li, and Heng Ji. 2015. A Novel Neural Topic Model and Its Supervised Extension.. In *AAAI*. 2210–2216.

[5] Dallas Card, Chenhao Tan, and Noah A Smith. 2017. A Neural Framework for Generalized Topic Models. *arXiv preprint arXiv:1705.09296* (2017).

[6] Jonathan Chang and David Blei. 2009. Relational topic models for document networks. In *AISTATS*. 81–88.

[7] Minmin Chen, Zhixiang Xu, Kilian Weinberger, and Fei Sha. 2012. Marginalized denoising autoencoders for domain adaptation. *arXiv preprint arXiv:1206.4683* (2012).

[8] Ning Chen, Jun Zhu, Fei Xia, and Bo Zhang. 2015. Discriminative relational topic models. *IEEE TPAMI* 37, 5 (2015), 973–986.

[9] Adji B Dieng, Chong Wang, Jianfeng Gao, and John Paisley. 2016. Topicrnn: A recurrent neural network with long-range semantic dependency. *arXiv preprint arXiv:1611.01702* (2016).

[10] Xiangnan He, Xiaoyu Du, Xiang Wang, Feng Tian, Jinhui Tang, and Tat-Seng Chua. 2018. Outer Product-based Neural Collaborative Filtering. In *IJCAI*. 2227–2233.

[11] Xiangnan He, Lizi Liao, Hanwang Zhang, Liqiang Nie, Xia Hu, and Tat-Seng Chua. 2017. Neural collaborative filtering. In *WWW*. 173–182.

[12] Xiangnan He, Hanwang Zhang, Min-Yen Kan, and Tat-Seng Chua. 2016. Fast matrix factorization for online recommendation with implicit feedback. In *CIKM*. ACM, 549–558.

[13] Zhao Kang, Chong Peng, Ming Yang, and Qiang Cheng. 2016. Top-n recommendation on graphs. In *CIKM*. ACM, 2101–2106.

[14] Jooyeon Kim, Dongwoo Kim, and Alice Oh. 2017. Joint Modeling of Topics, Citations, and Topical Authority in Academic Corpora. *arXiv preprint arXiv:1706.00593* (2017).

[15] Diederik P. Kingma and Max Welling. 2013. Auto-encoding variational bayes. *CoRR* abs/1312.6114 (2013).

[16] Jey Han Lau, David Newman, and Timothy Baldwin. 2014. Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In *EACL*. 530–539.

[17] Sheng Li, Jaya Kawale, and Yun Fu. 2015. Deep collaborative filtering via marginalized denoising auto-encoder. In *CIKM*. ACM, 811–820.

[18] Xiaopeng Li and James She. 2017. Collaborative variational autoencoder for recommender systems. In *KDD*. ACM, 305–314.

[19] Bin Liu, Lirong He, Yingming Li, Shandian Zhe, and Zenglin Xu. 2018. NeuralCP: Bayesian Multiway Data Analysis with Neural Tensor Decomposition. *Cognitive Computation* (2018). DOI:http://dx.doi.org/10.1007/s12559-018-9587-4

[20] Yan Liu, Alexandru Niculescu-Mizil, and Wojciech Gryc. 2009. Topic-link LDA: joint models of topic and author community. In *ICML*. ACM, 665–672.

[21] Hao Ma, Haixuan Yang, Michael R Lyu, and Irwin King. 2008. Sorec: social recommendation using probabilistic matrix factorization. In *CIKM*. ACM, 931–940.

[22] Chris J Maddison, Andriy Mnih, and Yee Whye Teh. 2016. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712* (2016).

[23] Andrew McCallum, Andres Corrada-Emmanuel, and Xuerui Wang. 2005. Topic and Role Discovery in Social Networks.. In *IJCAI*, Vol. 5. Citeseer, 786–791.

[24] Yishu Miao, Edward Grefenstette, and Phil Blunsom. 2017. Discovering Discrete Latent Topics with Neural Variational Inference. In *ICML*.

[25] Yishu Miao, Lei Yu, and Phil Blunsom. 2016. Neural variational inference for text processing. In *ICML*. 1727–1736.

[26] Willie Neiswanger, Chong Wang, Qirong Ho, and Eric P. Xing. 2014. Modeling Citation Networks Using Latent Random Offsets. In *UAI*. 633–642.

[27] Rajesh Ranganath, Sean Gerrish, and David Blei. 2014. Black box variational inference. In *AISTATS*. 814–822.

[28] Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. 2014. Stochastic backpropagation and approximate inference in deep generative models. *arXiv preprint arXiv:1401.4082* (2014).

[29] Michael Röder, Andreas Both, and Alexander Hinneburg. 2015. Exploring the space of topic coherence measures. In *WSDM*. ACM, 399–408.

[30] Akash Srivastava and Charles A. Sutton. Autoencoding variational inference for topic models. In *ICLR*.

[31] Fei Tian, Bin Gao, Di He, and Tie-Yan Liu. 2016. Sentence Level Recurrent Topic Model: Letting Topics Speak for Themselves. *arXiv preprint arXiv:1604.02038* (2016).

[32] Martin J Wainwright, Michael I Jordan, and others. 2008. Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning* 1, 1–2 (2008), 1–305.

[33] Chong Wang and David M Blei. 2011. Collaborative topic modeling for recommending scientific articles. In *KDD*. ACM, 448–456.

[34] Hao Wang, Xingjian Shi, and Dit-Yan Yeung. 2015. Relational Stacked Denoising Autoencoder for Tag Recommendation.. In *AAAI*. 3052–3058.

[35] Hao Wang, Xingjian Shi, and Dit-Yan Yeung. 2017. Relational Deep Learning: A Deep Latent Variable Model for Link Prediction.. In *AAAI*. 2688–2694.

[36] Hao Wang, Naiyan Wang, and Dit-Yan Yeung. 2015. Collaborative deep learning for recommender systems. In *KDD*. ACM, 1235–1244.

[37] Zenglin Xu, Bin Liu, Shandian Zhe, Haoli Bai, Zihan Wang, and Jeniffer Neville. 2018. Variational Random Function Model for Network Modeling. *IEEE TNNLS* (2018), 1–7.

[38] Zenglin Xu, Feng Yan, and Yuan (Alan) Qi. 2012. Infinite Tucker Decomposition: Nonparametric Bayesian Models for Multiway Data Analysis. In *ICML*.

[39] Hong-Jian Xue, Xin-Yu Dai, Jianbing Zhang, Shujian Huang, and Jiajun Chen. 2017. Deep matrix factorization models for recommender systems. In *IJCAI*. 233–240.

[40] Feng Yan, Zenglin Xu, and Yuan (Alan) Qi. 2011. Sparse matrix-variate Gaussian process blockmodels for network modeling. In *UAI*. 745–752.

[41] Jichuan Zeng, Jing Li, Yan Song, Cuiyun Gao, Michael R. Lyu, and Irwin King. 2018. Topic Memory Networks for Short Text Classification. In *EMNLP*.

[42] Tong Zhao, Julian McAuley, and Irwin King. 2014. Leveraging social connections to improve personalized ranking for collaborative filtering. In *CIKM*. ACM, 261–270.