# Collective Topic Modeling for Heterogeneous Networks

Hongbo Deng
Dept. of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
hbdeng@uiuc.edu

Bo Zhao
Dept. of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
bozhao3@uiuc.edu

Jiawei Han
Dept. of Computer Science
University of Illinois at Urbana-Champaign
Urbana, IL 61801
hanj@cs.uiuc.edu

## ABSTRACT

In this paper, we propose a joint probabilistic topic model for simultaneously modeling the contents of multi-typed objects of a heterogeneous information network. The intuition behind our model is that different objects of the heterogeneous network share a common set of latent topics so as to adjust the multinomial distributions over topics for different objects collectively. Experimental results demonstrate the effectiveness of our approach for the tasks of topic modeling and object clustering.

**Categories and Subject Descriptors:**
H.3.3 [**Information Storage and Retrieval**]: Information Search and Retrieval—*text mining, clustering*

**General Terms:** Algorithms, Experimentation

**Keywords:** Topic modeling, heterogeneous network

## 1. INTRODUCTION

In the age of Web 2.0, various kinds of textual documents, blogs, papers, and other user-generated content are published online and connected with users and other objects, leading to a heterogeneous information network with multi-typed objects. Many topic models, such as PLSA [2] and LDA [1], have been proposed and shown to be useful for document analysis. However, very little research has been conducted on modeling the topics of documents as well as their associated objects simultaneously in heterogeneous networks. Although there are several extensions of the topic model proposed to consider the relationships between objects, including the Author-Topic model [3] and Author-Conference-Topic model [5], these models are designed specifically for academic networks, which cannot deal with many more general cases. Generally, the interactions among multi-typed objects play a key role at disclosing the rich semantics of the network. Taking bibliographic data as an example, papers are highly related to their authors and associated venues, because an author can be characterized based on his published papers while a venue consists of various papers in a specific research area. Therefore, it is reasonable to assume there is a common set of latent topics for different objects in a heterogeneous network. Inspired by this intuition, we propose a collective topic model by considering both the textual information of documents and the relations between different objects, which could improve the performance of both topic modeling and object clustering.


*SIGIR'11,* July 24–28, 2011, Beijing, China.
ACM 978-1-4503-0757-4/11/07.

## 2. MODEL FORMULATION

Let $G = (V, E)$ denote a bibliographic heterogeneous network consisting of three types of object sets: a document set $\mathcal{D} = \{d_1, d_2, ..., d_{|D|}\}$, an author set $\mathcal{A} = \{a_1, a_2, ..., a_{|A|}\}$ and a venue (conference) set $\mathcal{C} = \{c_1, c_2, ..., c_{|C|}\}$. We represent a document $d$ with a bag of words $\{w_1, w_2, ..., w_{|d|}\}$, and use $N$ to denote the term-document matrix where $N_{ij}$ is the co-occurrences of word $w_i$ in $d_j$. In such a heterogeneous network, documents are linked with authors and venues based on the authorship and publish relationship, so it is reasonable to build a 'virtual document' for each author and venue by aggregating their associated documents. Then we obtain the term-author matrix $A$ and the term-venue matrix $C$. In this way, the associations between different objects in the heterogeneous network are indirectly modeled through the content. The basic idea of topic models is to model documents with a finite mixture model of $K$ latent topics and estimate the model parameters by fitting the data with the model. Documents, authors and venues are generally composed of words, so each of them can be decomposed by topic models, such as PLSA [2], respectively.

Rather than applying each separately, it is reasonable to merge them into a joint probabilistic model with a common set of underlying topics as shown in Fig. 1. Based on PLSA, one can define the following joint model for predicting terms in different objects: $P(w_i|d_j) = \sum_k P(w_i|z_k)P(z_k|d_j)$, $P(w_i|a_l) = \sum_k P(w_i|z_k)P(z_k|a_l)$ and $P(w_i|v_m) = \sum_k P(w_i|z_k)P(z_k|v_m)$. Notice that all the decompositions share the same latent topics $P(w_i|z_k)$. Thus the learned topics must be consistent across multi-typed objects. In this way, the relationships between different objects of the heterogeneous network are indirectly modeled in the proposed collective topic modeling. In general, we propose maximizing the following joint log-likelihood function with relative weights $\alpha$, $\beta$ and $\gamma$

$$\begin{aligned}\mathcal{L} = & \sum_i \Big(\alpha \sum_j N_{ij} \log \sum_{k=1}^K P(w_i|z_k)P(z_k|d_j) \\ & +\beta \sum_l A_{il} \log \sum_{k=1}^K P(w_i|z_k)P(z_k|a_l) \\ & +\gamma \sum_m C_{im} \log \sum_{k=1}^K P(w_i|z_k)P(z_k|v_m)\Big).\end{aligned} \quad (1)$$

This model can be easily extended to handle general cases, e.g., news articles and their associated entities.

Following the EM approach it is straightforward to derive a set of re-estimation equations. For the E-step, the posterior probabilities of the latent variables associated with each observation are formulated as follows

$$\begin{aligned}P(z_k|w_i, d_j) &= \frac{P(w_i|z_k)P(z_k|d_j)}{P(w_i|d_j)}, \\ P(z_k|w_i, a_l) &= \frac{P(w_i|z_k)P(z_k|a_l)}{P(w_i|a_l)}, \\ P(z_k|w_i, v_m) &= \frac{P(w_i|z_k)P(z_k|v_m)}{P(w_i|v_m)}.\end{aligned} \quad (2)$$

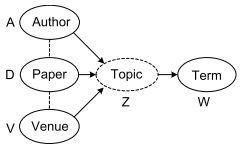


**Figure 1: An illustration of collective topic model.**

The multinomial distributions over topics are recomputed in the M-step according to

$$\begin{aligned} P(z_k|d_j) &= \frac{\sum_i N_{ij} P(z_k|w_i,d_j)}{\sum_{i'} N_{i'j}}, \\ P(z_k|a_l) &= \frac{\sum_i A_{il} P(z_k|w_i,a_l)}{\sum_{i'} A_{i'l}}, \\ P(z_k|v_m) &= \frac{\sum_i C_{im} P(z_k|w_i,v_m)}{\sum_{i'} C_{i'm}}, \end{aligned} \tag{3}$$

along with the mixing latent topics

$$\begin{aligned} P(w_i|z_k) \propto \alpha \frac{\sum_j N_{ij} P(z_k|w_i,d_j)}{\sum_i \sum_j N_{ij} P(z_k|w_i,d_j)} \\ + \beta \frac{\sum_l A_{il} P(z_k|w_i,a_l)}{\sum_i \sum_l A_{il} P(z_k|w_i,a_l)} + \gamma \frac{\sum_m C_{im} P(z_k|w_i,v_m)}{\sum_i \sum_m C_{im} P(z_k|w_i,v_m)}. \end{aligned} \tag{4}$$

For simplicity, we set $\alpha = \beta = \gamma = 1$. With an initial random guess of $\{P(w_i|z_k), P(z_k|d_j), P(z_k|a_l),\ P(z_k|v_m)\}$, the collective topic model (CTM) applies the E-step and M-step equations until a termination condition is met.

# 3. EXPERIMENTS

In this experiment, we use a subset of the DBLP records that belongs to four areas: database, data mining, information retrieval and artificial intelligence, and contains 28,569 documents, 28,702 authors and 20 conferences. The abstract is collected for representing each document, and this data collection has 11,771 unique terms. Moreover, we use a labeled data set [4] with 4057 authors, 100 papers and all 20 conferences for quantitative accuracy evaluation. For more details about the labeled data set, please refer to [4].

In order to visualize the hidden topics and compare different approaches, we extract topics from the data using both PLSA and CTM. Since the testing data is a mixture of four areas, it is interesting to see whether the extracted topics could automatically reveal this mixture. Therefore, in both PLSA and CTM, we predefine the number of topics to be 4. To make the comparison fair, we use the same starting points for PLSA and CTM. The most representative terms generated by CTM and PLSA are shown in Table 1. For the first three topics, although different algorithms select slightly different terms, all these terms can describe the corresponding topic to some extent. For Topic 4 (AI), the top keywords like "learning, based, knowledge" derived from CTM is obviously more telling than "problem, algorithm, paper" derived by PLSA. Similar subtle differences can be observed for Topic 3 (IR) as well. Intuitively, CTM selects more related terms for each topic than PLSA, which shows the better performance of CTM.

Now we give a quantitative evaluation of these models on object clustering. The hidden topics extracted by the topic modeling approaches can be regarded as clusters. The estimated conditional probability, e.g., $P(z_k|d_i)$ and $P(z_k|a_l)$, can be used to infer the cluster label for each object. In this experiment, we investigate the use of topic modeling approach for object clustering. To demonstrate how the multi-typed object clustering performance can be improved by collective topic modeling, we compared with the following state-of-the-art clustering algorithms.

- PLSA [2]: Performing on each object separatively.
- Author-Topic Model (ATM) [3].
- NetClus [4]: We implemented a topic based NetClus which utilizes the topic distribution instead of the word distribution for each document.

**Table 1: The most representative terms generated by our CTM model and the PLSA model. The terms are selected according to the probability $P(w|z)$.**

| Topic 1 (DB) | Topic 2 (DM) | Topic 3 (IR) | Topic 4 (AI) |
|---|---|---|---|
| Collective Topic Model (CTM) | | | |
| data | data | web | learning |
| database | mining | information | based |
| query | learning | retrieval | knowledge |
| databases | based | search | system |
| queries | clustering | based | reasoning |
| systems | algorithm | text | systems |
| system | classification | document | problem |
| management | analysis | user | logic |
| xml | approach | query | model |
| PLSA | | | |
| data | data | information | problem |
| database | mining | retrieval | algorithm |
| systems | learning | web | paper |
| query | based | based | reasoning |
| system | clustering | *learning* | logic |
| databases | classification | knowledge | based |
| management | algorithm | text | time |
| distributed | *image* | search | algorithms |
| queries | analysis | system | *search* |

**Table 2: Clustering results of different methods.**

| | Accuracy(%) | | | NMI (%) | | |
|---|---|---|---|---|---|---|
| Method | conf | paper | author | conf | paper | author |
| PLSA | 81.00 | 57.80 | 80.29 | 77.84 | 30.69 | 54.39 |
| ATM | - | **77.00** | 74.13 | - | 52.21 | 40.67 |
| NetClus | 79.75 | 65.00 | 70.82 | 76.69 | 40.96 | 47.43 |
| CTM | **85.25** | 76.65 | **83.55** | **80.01** | **52.83** | **59.98** |

Table 2 reports the evaluation results of different methods using two metrics, the accuracy and the normalized mutual information (NMI). The final performance scores were obtained by averaging the scores from 20 tests. As we can see, our CTM approach gets the best performance. Moreover, the improvement of CTM over PLSA and NetClus is more significant on the results of papers than other two objects. Although ATM obtains comparable performance to CTM in terms of papers, our CTM approach can obtain significant improvements in terms of authors. This shows that CTM model could mutually enhance topic modeling across different objects by considering both the content and relations of multi-typed objects collectively.

As future work, instead of modeling the associations between different objects through the content indirectly, we aim at adopting topic propagation techniques to directly integrate the content with relations for collective topic modeling of heterogeneous networks.

## *Acknowledgements*

The work was supported in part by the NSF IIS-09-05215, U.S. Air Force Office of Scientific Research MURI award FA9550-08-1-0265, and by the U.S. Army Research Laboratory under Cooperative Agreement No. W911NF-09-2-0053 (NS-CTA).

# 4. REFERENCES

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *JMLR*, 3:993–1022, 2003.

[2] T. Hofmann. Probabilistic latent semantic indexing. In *SIGIR*, pages 50–57, 1999.

[3] M. Steyvers, P. Smyth, M. Rosen-Zvi, and T. L. Griffiths. Probabilistic author-topic models for information discovery. In *KDD*, pages 306–315, 2004.

[4] Y. Sun, Y. Yu, and J. Han. Ranking-based clustering of heterogeneous information networks with star network schema. In *KDD*, pages 797–806, 2009.

[5] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su. Arnetminer: extraction and mining of academic social networks. In *KDD*, pages 990–998, 2008.