cluTM: Content and Link Integrated Topic Model on Heterogeneous Information Networks

Qian Wang[1], Zhaohui Peng[1](✉), Senzhang Wang[2], Philip S. Yu[3,4],
Qingzhong Li[1], and Xiaoguang Hong[1]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
wangqian8636@gmail.com, {pzh,lqz,hxg}@sdu.edu.cn

[2] School of Computer Science and Engineering, Beihang University, Beijing, China
szwang@buaa.edu.cn

[3] Department of Computer Science, University of Illinois at Chicago, Chicago, USA
psyu@uic.edu

[4] Institute for Data Science, Tsinghua University, Beijing, China

Abstract. Topic model is extensively studied to automatically discover the main themes that pervade a large and unstructured collection of documents. Traditional topic models assume the documents are independent and there are no correlations among them. However, in many real scenarios, a document may be interconnected with other documents and objects, and thus form a text related heterogeneous network, such as the DBLP bibliographic network. It is challenging for traditional topic models to capture the link information associated to diverse types of objects in such a network. To this end, we propose a **u**nified **T**opic **M**odel cluTM by incorporating both the document **c**ontent and various **l**inks in the text related heterogeneous network. cluTM combines the textual documents and the link structures by the proposed joint matrix factorization on both the text matrix and link matrices. Joint matrix factorization can derive a common latent semantic space shared by multi-typed objects. With the multi-typed objects represented by the common latent features, the semantic information can be therefore largely enhanced simultaneously. Experimental results on DBLP datasets demonstrate the effectiveness of cluTM in both topic mining and multiple objects clustering in text related heterogeneous networks by comparing against state-of-the-art baselines.

1 Introduction

As a powerful way to discover the hidden semantics of the document collection, topic models are extensively studied and successfully applied to many text mining tasks, such as information retrieval [1] and document clustering [2]. Traditional topic models assume that the documents are independent and do not consider the correlation among them. With the flourish of Web application, textual documents such as papers, blogs and product reviews, are not only getting richer, but also interconnecting with other objects like users in various ways; and

J. Li and Y. Sun (Eds.): WAIM 2015, LNCS 9098, pp. 207–218, 2015.
DOI: 10.1007/978-3-319-21042-1_17

therefore form text related heterogeneous information networks [3]. Take the bibliographic data shown in Figure 1 as an example. There are three types of objects, papers, authors, and venues in such a heterogeneous network. These objects form two types of relationships: the *author-write-paper* relationship between authors and papers, and the *venue-publish-paper* relationship between venues and papers. It is challenging for traditional topic models to capture the rich information, especially the link information in such a text related heterogeneous information network.

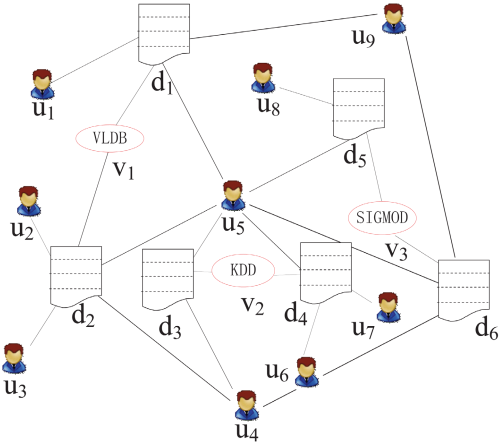

Fig. 1. The bibliographic heterogeneous network with three types of objects: papers, authors, venues and two types of links: *author-write-paper*, *venue-publish-paper*

Traditional topic models, such as latent semantic analysis (LSA) [4], probabilistic latent semantic analysis (PLSA) [5], and latent dirichlet allocation (LDA) [6] focus on purely utilizing the textual information to discover topics with the assumption that the documents are *i.i.d.* (independent and identical distributed). With the explosive of interconnected textual contents with rich link information, the assumption may not hold and traditional models become less effective. Although some attempts, such as LaplacianPLSI [7], NetPLSA [8], and iTopicmodel [9] have been conducted to combine topic modeling with link information in a homogeneous network, how to integrate various types of links associated to different types of objects into a unified topic model is still less studied.

Although the links among documents as well as other types of objects might be helpful for analyzing text, it is non-trivial to handle the rich heterogeneous information in a unified framework. First, it is challenging to model the semantic information of links such as the *author-write-paper* and *venue-publish-paper* relationships. Different from text, link structure is a totally different type of information and can not be easily added to traditional topic models in a straightforward manner. Second, there are usually several types of different objects in a heterogeneous information network. Different types of objects may have their own inherent information and should be treated differently. How to use the different types of objects and integrate them in a unified way with the textual information also makes the studied problem challenging.

In this paper, we propose a unified topic model named cluTM by incorporating the heterogeneous link information into topic modeling. cluTM learns a latent semantic space by jointly factorizing the document-phrase matrix and the link matrices with latent semantic analysis. The basic idea is that the textual documents and link information in the heterogeneous information networks have similar latent semantic features. For example, in the bibliographic data, a paper contains several topics. Likewise, the researchers and venues also have their preferred research topics associated to related papers. The inherent connections between contents and links can be therefore constructed by assuming that the text matrix and link matrices share the same latent semantic features. With such an assumption, all the objects in the heterogeneous information network are projected into a unified latent semantic space based on the common latent semantic features. In the unified latent semantic space, each object is represented as a vector. Topics of documents and clusters of other types of objects can be easily obtained by calculating the similarity of the vectors.

We summarize the main contributions of this paper as follows:

- study the novel problem of topic mining and multi-objects clustering simultaneously in a text related heterogeneous information network;
- propose a unified topic model to seamlessly integrate the content of textual documents and links by joint matrix factorization;
- extensive experiments on DBLP dataset show the effectiveness of the proposed model on topic modeling and object clustering by comparing against traditional topic models.

The rest of this paper is organized as follows. Section 2 introduces some basic concepts. We elaborate cluTM in Section 3. Section 4 presents the extensive experiment results. We discuss the related work in section 5 and finally conclude our work in section 6.

2 Preliminaries

In this section, we formally introduce several related concepts and notations to help us state the problem.

Definition 1. ***Information Network*** *[3]. Given a set of objects from K types $\mathcal{X} = \{X_k\}_{k=1}^{K}$, where X_k is a set of objects belonging to the k_{th} type, a graph $G =< V, E >$ is called an information network on objects $\mathcal{X}$, if $V = \mathcal{X}$, and E is a binary relation on V. Specifically, we call such an information network* ***heterogeneous information network*** *when $K \geq 2$.*

Definition 2. ***Text Information Network.*** *An information network $G =< V, E >$ with K types of objects is called a text information network if there* *exists at least one type of text object in the network, i.e.* $\exists X_k \in \mathcal{X}$ *that the type of* X_k *is text. Specifically, we call a text information network* ***heterogeneous text information network*** *when* $K \geq 2$.

DBLP Bibliographic Network Example. We use the DBLP bibliographic network as an example to illustrate the heterogeneous text information network. As shown in Figure 1, there are three types of objects, i.e., authors A, venues VE and papers D, and two types of links among papers, authors, and venues. The type of paper object is text. The bibliographic network can be denoted as $G = (D \cup A \cup VE, E)$, where E is a set of edges that describe the relationships between papers $D = \{d_1, ..., d_n\}$, authors $A = \{a_1, ..., a_l\}$ as well as venues $VE = \{ve_1, ..., ve_o\}$.

In our model, each topic can be represented as a set of meaningful frequent phrases [10], definited as follows.

Definition 3. ***Meaningful Frequent Phrases*** *Meaningful frequent phrases are defined as the phrases that capture the main themes of the document collection. Meaningful frequent phrases lay a foundation for the readability of the discovered topics. They can be represented as* $MFP = \{mfp_1, mfp_2, ..., mfp_M\}$, *where* mfp_m *denoting the* m_{th} *meaningful frequent phrase.*

3 cluTM: Incorporating Text and Links in a Unified Framework

We will first revisit the classic LSA model that is widely used to discover topics of document by matrix factorization. Motivated by LSA model, we next introduce how to conduct the matrices factorization on the *document-author* matrix and *document-venue* matrix. Finally, we elaborate how to combine the content and link information by joint matrix factorization with a assumption that these matrices share the same latent semantic space.

3.1 LSA on Document-Phrase Matrix

We use the classic LSA model to discover the latent topics of documents. The key idea of LSA model is to project documents as well as terms into a relatively low dimensional vector space, namely the latent semantic space, and produce a set of topics associated with documents [4].

In our model, documents are represented as a bag of meaningful frequent phrases. Consider the analysis of document-phrase matrix $M_{D-MFP} \in R^{n \times m}$, and it is a sparse matrix whose rows represent documents, and columns represent phrases, where n is the number of documents and m is the number of meaningful frequent phrases. Singular vector decomposition [12] is performed on matrix M_{D-MFP} as follows:

$$M_{D-MFP} = U_{D-MFP} \Sigma_{D-MFP} V^T_{D-MFP} \tag{1}$$

where U_{D-MFP} and V_{D-MFP} are orthogonal singular matrices U^T_{D-MFP} $U_{D-MFP} = V^T_{D-MFP}V_{D-MFP} = I$ (I is the identity matrix) and Σ_{D-MFP} is a diagonal matrix containing the singular values of M_{D-MFP}.

Given an integer k ($k << rank(M_{D-MFP})$), LSA only remains the first k singular vectors, $U_{D-MFP} \in R^{n\times k}$, $V_{D-MFP} \in R^{m\times k}$, and sets all but the largest k singular values to zero, $\Sigma_{D-MFP} \in R^{k\times k}$. The matrix $Y = U_{D-MFP}\Sigma_{D-MFP}$ ($Y \in R^{n\times k}$) defines a new representation of documents that each column corresponds to a topic and each row is a $k-$dimensional vector representing the weights of a document in the k topics. Therefore, the LSA approximation of M_{D-MFP} can be obtained by YV^T_{D-MFP}.

This can be transformed into an optimization problem that aims to approximate matrix M_{D-MFP} with YV^T_{D-MFP} as follows,

$$min\, ||M_{D-MFP} - YV^T_{D-MFP}||^2_F + \gamma_1||V_{D-MFP}||^2_F \qquad (2)$$

where $||\cdot||_F$ is the Frobenius norm, γ_1 is the parameter, $\gamma_1||V_{D-MFP}||^2_F$ is a regularization term to improve the robustness. The $i-th$ row vector of Y can be considered as the latent semantic feature vector of document d_i.

3.2 Link Matrices Factorization

Taking the bibliographic network in Figure 1 as an example again, the relationships between papers and authors as well as papers and venues can be represented by link matrices $M_{D-A} \in R^{n\times l}$ and $M_{D-VE} \in R^{n\times o}$, respectively. l is the number of authors and o is the number of venues. In LSA model, a document contains several topics with each topic associated with a set of frequently used terms. Likewise, an author also has several preferred research topics with each research topic associated with a set of related papers. If we consider the authors and papers as documents and words respectively, we can use the similar idea to LSA to analyze the latent semantic of the $author - paper$ link. Motivated by above idea, the link matrices M_{D-A} can also be factorized by SVD as follows,

$$M_{D-A} = U_{D-A}\Sigma_{D-A}V^T_{D-A} \qquad (3)$$

where U_{D-A} and V_{D-A} are orthogonal matrices $U^T_{D-A}U_{D-A} = V^T_{D-A}V_{D-A} = I$ and the diagonal matrix Σ_{D-A} contains the singular values of M_{D-A}.

Likewise, each venue prefers to accept papers of some particular research topics. The latent topics of a venue preferring can be obtained by factorizing the document-venue matrix using SVD as follows,

$$M_{D-VE} = U_{D-VE}\Sigma_{D-VE}V^T_{D-VE} \qquad (4)$$

where U_{D-VE} and V_{D-VE} are orthogonal matrices $U^T_{D-VE}U_{D-VE} = V^T_{D-VE}V_{D-VE} = I$ and the diagonal matrix Σ_{D-VE} contains the singular values of M_{D-VE}.

Similar to LSA model, we also only keep the first k singular vectors and set the other singular values to zero. For the matrix M_{D-A} and M_{D-VE}, we use the matrix $Y_{D-A} = U_{D-A}\Sigma_{D-A}$ and $Y_{D-VE} = U_{D-VE}\Sigma_{D-VE}$ to represent the document respectively. Thus the matrices M_{D-A} and M_{D-VE} can be represented as follows

$$M_{D-A} \approx Y_{D-A}V_{D-A}^T \tag{5}$$

$$M_{D-VE} \approx Y_{D-VE}V_{D-VE}^T \tag{6}$$

where V_{D-A} is a $l \times k$ matrix, V_{D-VE} is a $o \times k$ matrix, and Y_{D-A}, Y_{D-VE} are the latent semantic feature matrices. Each column of Y_{D-A} and Y_{D-VE} represents a topic and each row is $k-$dimensional vector representing the weights of a document in the k topics. Therefore, Y_{D-A} and Y_{D-VE} are very similar to the matrix Y. In our model, to combine the content of textual documents and link information in the heterogeneous text information network, we assume that they share the latent semantic feature Y, i.e. $Y_{D-A} = Y_{D-VE} = Y$.

3.3 Combing Content and Link by Joint Matrix Factorization

Based on the assumption discussed above, the document-phrase matrix M_{D-MFP} and link matrices M_{D-A}, M_{D-VE} are connected by the latent semantic feature Y, that is, the latent feature for content is tied to the latent feature for links. Our model aims to find a latent semantic feature Y that best explains the semantic captured by M_{D-MFP} and M_{D-A}, M_{D-VE} simultaneously. Furthermore, different types of objects and links reflect distinctive semantics of a heterogeneous text information network, so they should be treated differently. To achieve these goals, we propose a joint matrix factorization framework to fuse them into such an unified optimization problem,

$$\begin{aligned} & minJ(Y, V_{D-MFP}, V_{D-A}, V_{D-VE}) \\ =& min\{\lambda(||M_{D-MFP} - YV_{D-MFP}^T||_F^2 + \gamma_1||V_{D-MFP}||_F^2) \\ & + \alpha(||M_{D-A} - YV_{D-A}^T||_F^2 + \gamma_2||V_{D-A}||_F^2) \\ & + \beta(||M_{D-VE} - YV_{D-VE}^T||_F^2 + \gamma_3||V_{D-VE}||_F^2)\} \end{aligned} \tag{7}$$

where λ, α and β ($\lambda > 0$, $\alpha > 0$, $\beta > 0$) are parameters to balance the relative importance of document-phrase matrix M_{D-MFP} and link matrices M_{D-A}, M_{D-VE}. We set a constraint $\lambda + \alpha + \beta = 1$. γ_1, γ_2 and γ_3 are regularization parameters that improve the robustness. V_{D-MFP}, V_{D-A} and V_{D-VE} are $m \times k$, $l \times k$, and $o \times k$ matrix respectively. Y is a $n \times k$ matrix. Note that if $\alpha = 0$, $\beta = 0$, thus $\lambda = 1$, the unified topic model boils down to the LSA model on document-phrase matrix.

The optimization problem aims to simultaneously approximate M_{D-MFP}, M_{D-A}, M_{D-VE} by YV_{D-MFP}^T, YV_{D-A}^T, YV_{D-VE}^T respectively, a product of two low-dimensional matrices with regularizations. The joint optimization illustrated in Eq.7 can be solved by using the standard Conjugate Gradient (CG) method. The gradients for the object function J are computed as follows:

$$\frac{\partial J}{\partial V_{D-MFP}} = \lambda(V_{D-MFP}Y^TY - M_{D-MFP}^TY) + \lambda\gamma_1 V_{D-MFP} \tag{8}$$

$$\frac{\partial J}{\partial V_{D-A}} = \alpha(V_{D-A}Y^TY - M_{D-A}^TY) + \alpha\gamma_2 V_{D-A} \tag{9}$$

$$\frac{\partial J}{\partial V_{D-VE}} = \beta(V_{D-VE}Y^TY - M_{D-VE}^TY) + \beta\gamma_3 V_{D-VE} \tag{10}$$

$$\begin{aligned}\frac{\partial J}{\partial Y} =&\lambda(YV_{D-MFP}^TV_{D-MFP} - M_{D-MFP}V_{D-MFP})\\ &+ \alpha(YV_{D-A}^TV_{D-A} - M_{D-A}V_{D-A})\\ &+ \beta(YV_{D-VE}^TV_{D-VE} - M_{D-VE}V_{D-VE})\end{aligned} \tag{11}$$

The new optimal latent semantic feature Y is to capture both the document-phrase matrix M_{D-MFP} and the link matrices M_{D-A}, M_{D-VE} in the heterogeneous text information network.

A unified latent semantic space can be constructed based on the obtained optimal latent semantic feature Y. All the objects in the heterogeneous information network are projected into the unified latent semantic space in which each paper, meaningful frequent phrase, author and venue is represented by a k-dimensional vector. According to the similarity calculation of vectors, we can get the topics. Analogously, the author clusters and venue clusters also can be obtained by similarity calculation.

4 Evaluations

In this section, we evaluate cluTM on the real dataset. First, we introduce the experiment setup, including the dataset and evaluation metric. Then we show the experimental results from the following three aspects: case study, parameters analysis, and quantitive comparison with baselines.

4.1 Dataset and Metric

We evaluate cluTM on the Digital Bibliography and Library Project (DBLP) dataset. In our experiments, we select papers from DBLP of four research areas, i.e. database (DB), data mining (DM), information retrieval (IR) and artificial intelligence (AI). The selected dataset contains 1200 papers, 1576 authors and 8 conferences. We extract 1660 meaningful frequent phrases from these papers. The heterogeneous text information network of this dataset contains three types of objects: papers, authors and venues, and two types of links: *paper-author* link and *paper-venue* link. There are 3139 *paper-author* links and 1200 *paper-venue* links in total. Link matrices M_{D-A}, M_{D-VE} are constructed from the heterogeneous text information network, and the element value in matrix M_{D-MFP} is obtained by using the $tf - idf$ weight of the phrases. As we select the papers from four research areas, we set the number of topics k to be 4.

For a quantitative evaluation, we use F1-measure as the metric. In our experiments, there are four topic clusters. For each topic cluster, we calculate the Precision and Recall with regard to each given category. Specifically, for the obtained

Table 1. The topic representation generated by cluTM

Topic 1	Topic 2	Topic 3	Topic 4
database systems	data mining	information retrieval	artificial intelligence
database management	data analysis	language models	machine learning
relational databases	data clustering	web search	knowledge-based algorithms
data integration	classification algorithms	learn to rank	knowledge based systems
distributed databases	knowledge discovery	search engine	expert systems
query processing	mining problems	keyword search	pattern recognition
distributed computing	rule learning	document retrieval	knowledge engineering
query optimization	pattern matching	semantic search	user interface

cluster label j and the true cluster label i, the precision can be calculated by $Precision(i, j) = \frac{n_{ij}}{n_j}$, and the recall can be calculated by $Recall(i, j) = \frac{n_{ij}}{n_i}$, where n_{ij} is the number of members of category i in cluster j, n_i is the number of members in the given category i, and n_j is the number of members in cluster j. Based on precision and recall, the F1-measure of cluster j and i can be calculated by

$$F1(i, j) = \frac{2 \times Precision(i, j) \times Recall(i, j)}{Precision(i, j) + Recall(i, j)}. \tag{12}$$

The F1-measure of the whole clustering results is defined as a weighted sum over all the categories as follows: $F1 = \sum_i \frac{n_i}{n} max_j F1(i, j)$.

4.2 Experimental Results

We first analyze the topic modeling results with case studies. Then we discuss the effect of parameters on performance. Finally, experiments are conducted to compare the performance of object clustering with different models.

Table 2. Topics discovered by PLSA and TMBP-Regu

Topics discovered by PLSA				Topics discovered by TMBP-Regu			
Topic 1	Topic 2	Topic 3	Topic 4	Topic 1	Topic 2	Topic 3	Topic 4
data	data	information	problem	data	database	information	learning
database	mining	retrieval	algorithm	database	mining	web	based
systems	learning	web	paper	query	algorithm	retrieval	knowledge
query	based	based	reasoning	databases	clustering	search	model
system	clustering	learning	logic	systems	classification	based	problem
databases	classification	knowledge	based	queries	based	text	reasoning
management	algorithm	text	time	system	algorithms	language	system
distributed	image	search	algorithm	processing	rules	user	logic

Topic Analysis with Case Study. In our model, we set parameters $\lambda = 0.6$, $\alpha = 0.3$, and $\beta = 0.1$ due to the better performance based on our empirical experiment results. The topic modeling results are shown in Table 1. Each discovered topic is represented as a set of meaningful frequent phrases.

PLSA [5] and TMBP-Regu [11] are selected as baselines. The most representative terms generated by PLSA and TMBP-Regu on the DBLP dataset are shown in Table 2. Compared with the results in Table 2, the results shown in Table 1 is easier to understand the meanings of the four topics by meaningful frequent phrases, i.e., “database systems”, “data mining”, “information retrieval”, and “artificial intelligence”. cluTM and TMBP-Regu achieve better performance than PLSA by considering the heterogeneous text information network.

For the first three topics discovered by PLSA and TMBP-Regu, although different algorithms select different terms, all these terms can reveal the topics to some extent. For Topic 4, the topics such as“artificial intelligence”, derived from cluTM is obviously better than the terms “problem, algorithm, paper” derived by PLSA and “learning, based, knowledge” derived by TMBP-Regu. Therefore, from the view of readability of the topics, cluTM is better than PLSA and TMBP-Regu by representing the topics by a set of meaningful frequent phrases.

Parameter Analysis. In our model, there are three essential parameters, λ, α and β in joint matrix factorization. In this section, we study the effect of these parameters on the performance of the proposed cluTM.

Table 3. The effect of parameters on paper, author, and venue

F1-measure		α (Paper)										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
λ	0	0.1213	0.1235	0.1268	0.1306	0.1339	0.1373	0.1410	0.1452	0.1436	0.1367	0.1305
	0.2	0.2555	0.2561	0.2627	0.2692	0.2747	0.2783	0.2906	0.2869	0.2724	-	-
	0.4	0.5212	0.5263	0.5408	0.5621	0.5881	0.5840	0.5516	-	-	-	-
	0.6	0.7098	0.7417	0.7544	**0.7857**	0.7336	-	-	-	-	-	-
	0.8	0.5406	0.5538	0.5511	-	-	-	-	-	-	-	-
F1-measure		α (Author)										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
λ	0	0.2317	0.2365	0.2472	0.2521	0.2598	0.2636	0.2684	0.2739	0.2691	0.2619	0.2508
	0.2	0.4256	0.4310	0.4539	0.4623	0.4807	0.4885	0.4982	0.4914	0.4749	-	-
	0.4	0.5815	0.5872	0.6138	0.6294	0.6581	0.6563	0.6226	-	-	-	-
	0.6	0.7501	0.7863	0.7949	**0.8332**	0.8058	-	-	-	-	-	-
	0.8	0.6274	0.6500	0.6388	-	-	-	-	-	-	-	-
F1-measure		α (Venue)										
		0	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
λ	0	0.3122	0.3243	0.3294	0.3361	0.3459	0.3536	0.3623	0.3789	0.3685	0.3544	0.3206
	0.2	0.5209	0.5262	0.5337	0.5462	0.5629	0.5813	0.5896	0.5735	0.5572	-	-
	0.4	0.6047	0.6231	0.6475	0.6602	0.6828	0.6893	0.6579	-	-	-	-
	0.6	0.7936	0.8315	0.8520	**0.8668**	0.8476	-	-	-	-	-	-
	0.8	0.6418	0.6596	0.6302	-	-	-	-	-	-	-	-

As mentioned in section 3.3, these parameters are used to balance the relative importance of document-phrase matrix M_{D-MFP}, link matrices M_{D-A} and

M_{D-VE}. When $\alpha = 0$, $\beta = 0$, the joint regularization framework boils down to the LSA model. Since $\lambda + \alpha + \beta = 1$, we vary λ from 0 to 1 by step 0.2 and α from 0 to 1 by step 0.1 respectively. Tables 3 report the results with the varied parameter values.

Tables 3 show that the best performance is obtained with $\lambda = 0.6$, $\alpha = 0.3$, thus $\beta = 0.1$. When $\lambda < 1$, the joint regularization framework takes into account both textual documents and links in the heterogeneous text information network. We observe that the performance is improved over the LSA model ($\lambda = 1$) when incorporating link information. One can also observe that the $document-author$ matrix M_{D-A} is more important than the $document - venue$ matrix M_{D-VE} in the joint regularization framework by the different values of α, β. Note that with the decrease of λ, the performance becomes worse and even worse than the standard LSA. This is mainly because cluTM relies more on the topic consistency between the content of textual documents and links while ignores the intrinsic topic of the textual documents. Due to the superior performance, we empirically set $\lambda = 0.6$, $\alpha = 0.3$, $\beta = 0.1$ in the following experiments.

Clustering Performance Comparison of Objects. We apply cluTM on the task of object clustering. The discovered topics can also be regarded as clusters. We can obtain the clustering results of other objects similarly.

The proposed cluTM is compared with the following two state-of-the-art baselines: latent semantic analysis (LSA), and LSA-PTM [10]. Table 4 reports the clustering performance comparison on different methods.

Table 4. Clustering performance comparasion

Metric	F1-measure			
Object	Paper	Author	Venue	Average
LSA	0.5359	0.6040	0.6990	0.6130
LSA-PTM	0.7535	0.7861	0.8136	0.7844
cluTM	0.7857	0.8332	0.8668	0.8286

For the DBLP data, cluTM and LSA-PTM cluster all types of objects in different groups by considering both the textual documents and the link information. As one can see, both cluTM and LSA-PTM achieve better performance than LSA. This shows that integrating the heterogeneous network structures into topic modeling does help us better cluster the objects. Meanwhile, compared with LSA-PTM, cluTM is consistently better on all the three types of objects. This is mainly because LSA-PTM combines the textual content and heterogeneous network structures as two independent stages, while cluTM combines the textual documents and the heterogeneous network structures into a joint regularization framework such that they can mutually enhance each other.

5 Related Work

Topic modeling is an unsupervised approach to automatically discover the latent semantic of document collections. It has attracted a lot of attention in multiple types of text mining tasks, such as information retrieval [1], geographical topic discovery [13], topic level information diffusion modeling in social media [18].

Many topic models, such as latent semantic analysis (LSA) [4], probabilistic latent semantic analysis (PLSA) [5] and Latent Dirichlet Allocation (LDA) [6] have been successfully applied or extended to many data analysis problems, including document clustering and classification [7,14], author-topic modeling [15,16]. However, most of these models merely consider the textual documents while ignore the network structures. Several proposed topic models, such as LaplacianPLSI [7], NetPLSA [8] and iTopicmodel [9] have combined topic modeling and network structures, but they only emphasize on the homogeneous networks, such as document network and co-authorship network. Recent study [10] integrates the heterogeneous network structures into topic modeling, however, it combines the textual documents and the heterogeneous network structures as two independent stages. Our model combines the textual documents and heterogeneous network structures into a joint regularization framework in which the textual content analysis and heterogeneous network analysis can mutually enhance each other. Experimental results prove the effectiveness of our model.

Link analysis has been a hot topic for a few years since the advent of Pagerank and HITS. Many techniques have been proposed to analyze the heterogeneous networks. For example, [17] proposed a Co-HITS algorithm for bipartite graph analysis. Graph-based methods have been widely and successfully applied in data mining and information retrieval, such as text classification [14], and document re-ranking [19]. However, most of existing work treats different objects uniformly. Our work is different from them, as we focus on heterogeneous information networks and propose a joint regularization framework, in which different types of objects are treated in a different way.

6 Conclusion

In this paper, we proposed a unified topic model cluTM to effectively discover topics of documents and cluster objects of various types simultaneously on heterogeneous text information networks. cluTM first conducted latent semantic analysis on the content of textual documents and factorized the link matrices of objects by SVD separately; then fused all the matrices into a single, compact feature representation by joint matrix factorization to find the common latent feature. By projecting all the objects in the heterogeneous text information networks into the unified latent semantic space, topics of documents and clusters of other objects could be finally obtained by calculating their similarity. We evaluated cluTM on DBLP bibliographic dataset against several state-of-the-art baselines. Experimental results showed the effectiveness of cluTM.

Acknowledgments. This work was supported in part by US NSF (Grant No.CNS-1115234), NSFC (Grant No.61170052, No.61100167), and the Natural Science Foundation of Shandong Province of China (Grant No.ZR2013FQ009).

References

1. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: SIGIR, pp. 178–185 (2006)
2. Xie, P.T., Xing, E.P.: Integrating document clustering and topic modeling. In: UAI, pp. 694–703 (2013)
3. Sun, Y.Z., Han, J.W., Yan, X., Yu, P.S.: Mining knowledge from interconnected data: a heterogeneous information network analysis approach. In: VLDB Endowment, pp. 2022–2023 (2012)
4. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. Journal of the American Society for Information Science and Technology, 391–407 (1990)
5. Hofmann, T.: Probabilistic latent semantic indexing. In: SIGIR, pp. 50–57 (1999)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. Journal of Machine Learning Research, 993–1022 (2003). ACM Press, New York
7. Cai, D., Mei, Q., Han, J., Zhai, C.: Modeling hidden topics on document manifold. In: CIKM, pp. 911–920 (2008)
8. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. In: WWW, pp. 101–110 (2008)
9. Sun, Y.Z., Han, J.W., Gao, J., Yu, Y.T.: Itopicmodel: information network-integrated topic modeling. In: ICDM, pp. 493–502 (2009)
10. Wang, Q., Peng, Z., Jiang, F., Li, Q.: LSA-PTM: a propagation-based topic model using latent semantic analysis on heterogeneous information networks. In: Wang, J., Xiong, H., Ishikawa, Y., Xu, J., Zhou, J. (eds.) WAIM 2013. LNCS, vol. 7923, pp. 13–24. Springer, Heidelberg (2013)
11. Deng, H., Han, J., Zhao, B., Yu, Y.: Probabilistic topic models with biased propagation on heterogeneous information networks. In: KDD, pp. 1271–1279 (2011)
12. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. Numerische Mathematic, 403–420 (1970). Springer, Heidelberg
13. Yin, Z., Cao, L., Han, J., Zhai, C.: Geographical topic discovery and comparison. In: WWW, pp. 247–256 (2011)
14. Zhu, S., Yu, K., Chi, Y., Gong, Y.: Combining content and link for classification using matrix factorization. In: SIGIR, pp. 487–494 (2007)
15. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.L.: Probabilistic author-topic models for information discovery. In: KDD, pp. 306–315 (2004)
16. Tang, J., Zhang, R.M., Zhang, J.: A topic modeling approach and its integration into the random walk framework for academic search. In: ICDM, pp. 1055–1060 (2008)
17. Deng, H., Lyu, M.R., King, I.: A generalized Co-HITS algorithm and its application to bipartite graphs. In: KDD, pp. 239–248 (2009)
18. Wang, S.Z., Hu, X., Yu, P.S., Li, Z.J.: MMRate: inferring multi-aspect diffusion networks with multi-pattern cascades. In: KDD, pp. 1246–1255 (2014)
19. Deng, H., Lyu, M.R., King, I.: Effective latent space graph-based re-ranking model with global consistency. In: WSDM, pp. 212–221 (2009)