LSA-PTM: A Propagation-Based Topic Model Using Latent Semantic Analysis on Heterogeneous Information Networks

Qian Wang[1,2], Zhaohui Peng[1,2,*], Fei Jiang[1,2], and Qingzhong Li[1,2]

[1] School of Computer Science and Technology, Shandong University, Jinan, China
[2] Shandong Provincial Key Laboratory of Software Engineering
{wangqian8636,jf29762}@gmail.com,
{pzh,lqz}@sdu.edu.cn

Abstract. Topic modeling on information networks is important for data analysis. Although there are many advanced techniques for this task, few methods either consider it into heterogeneous information networks or the readability of discovered topics. In this paper, we study the problem of topic modeling on heterogeneous information networks by putting forward LSA-PTM. LSA-PTM first extracts meaningful frequent phrases from documents captured from heterogeneous information network. Subsequently, latent semantic analysis is conducted on these phrases, which can obtain the inherent topics of the documents. Then we introduce a topic propagation method that propagates the topics obtained by LSA on the heterogeneous information network via the links between different objects, which can optimize the topics and identify clusters of multi-typed objects simultaneously. To make the topics more understandable, a topic description is calculated for each discovered topic. We apply LSA-PTM on real data, and experimental results prove its effectiveness.

Keywords: topic modeling, latent semantic analysis, topic propagation, heterogeneous information network.

1 Introduction

Information networks have been popularly used to represent the networked systems, such as social network, bibliographic network and so on. Data mining and knowledge discovery can be conducted on these networks [1]. Recently, with the explosion of textual documents in the heterogeneous information networks, such as papers, product reviews and other textual content, text-rich heterogeneous information networks come into being. Taking bibliographic data for example, one author can write several papers and one paper may be written by several authors. Likewise, each paper is commonly published in a venue (e.g., conferences, journals, etc.) and a venue usually publishes a number of papers. Thus we not only obtain the textual content of documents, but also the interactions among multi-typed objects as illustrated in Figure 1.

* Corresponding author.

J. Wang et al. (Eds.): WAIM 2013, LNCS 7923, pp. 13–24, 2013.

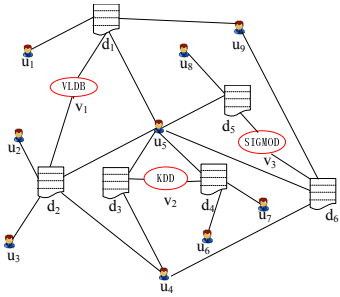

Fig. 1. Example of heterogeneous information network

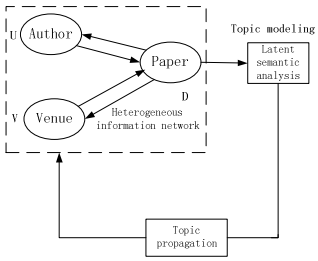

Fig. 2. Topic modeling on heterogeneous Information network

Topic modeling on information networks has been proved to be useful for document analysis. However, many topic models, such as LaplacianPLSI [2], NetPLSA [3] and iTopicmodel [4], merely deal with homogeneous networks. Besides, most existing methods [3, 4, 5, 6, 7] represent the discovered topics by word distributions. Although the topic word distributions may be intuitively meaningful, it is still difficult for users to fully understand the meaning of each topic. To address these problems, in this paper, we investigate and develop a novel topic model, LSA-PTM, for heterogeneous information networks, which solves the problems of topic modeling on heterogeneous information networks and the readability of discovered topics. LSA-PTM first extracts meaningful frequent phrases from documents that are captured from text-rich heterogeneous information network. The meaningful frequent phrases are useful to the readability of the discovered topics. Subsequently latent semantic analysis is conducted on these phrases, which can model the inherent topics of documents. Furthermore, we introduce a topic propagation method based on the links between different objects, which enhances the inherent topics and identify clusters of multi-typed objects simultaneously. The basic idea of the topic propagation is to propagate the topics obtained by topics models on the heterogeneous information networks as illustrated in Figure 2. To better understand the meaning of each topic, a topic description is computed for each topic.

In short, the contributions of our paper can be summarized as follows: (1) the proposed LSA-PTM effectively incorporates heterogeneous information networks with topic modeling; (2) LSA-PTM solves the intelligibility of each discovered topic by modeling each topic as a set of meaningful frequent phrases along with a topic description; (3) experiments are conducted on real dataset, and the results prove the effectiveness of our proposed model.

The rest of this paper is organized as follows. In section 2, we introduce some basic concepts. We elaborate LSA-PTM in Section 3. Section 4 presents the extensive experiment results. Finally, we discuss the related work in section 5 and conclude our work in section 6.

2 Basic Concepts

In this section, we formally introduce several related concepts and notations.

Definition 1. (Heterogeneous Information Network): A heterogeneous information network is defined as an information network with multiple types of objects and/or multiple types of links.

Text-rich Heterogeneous Information Network: When a heterogeneous information network contains a set of textual documents, it turns into a text-rich heterogeneous information network.

DBLP Bibliographic Network Example: As shown in Figure 1, there are three types of objects (i.e., $N = 3$) in this bibliographic network, including authors U, venues V and papers D. The bibliographic network can be denoted as $G = (D \cup U \cup V, E)$, where E is the set of edges that describe the relationships between papers $D = \{d_1, \ldots, d_n\}$ and authors $U = \{u_1, \ldots, u_l\}$ as well as venues $V = \{v_1, \ldots, v_o\}$.

Definition 2. (Topic Description): A topic description is a phrase containing several sequential words which can largely cover the semantic meaning of the discovered topic. In our model, each discovered topic is represented by a set of meaningful frequent phrases, and the topic description is calculated from these phrases.

For example, if a discovered topic is represented by a set of meaningful frequent phrases, such as “database systems”, “database management”, “distributed databases”, “relational databases”, and so on. Through a serial of calculation, “database systems” could be the topic description of this topic.

3 LSA-PTM

In this section, we elaborate our proposed model, LSA-PTM, which is a propagation-based topic model using latent semantic analysis on heterogeneous information networks.

3.1 Meaningful Frequent Phrases Extraction

Phrases that consist of two to five sequential words that appear in more than ψ documents in the corpus are defined as frequent phrases. The threshold ψ is to filter the “noise” phrases that just appear in a handful of documents, and it is decided by the number of documents in document corpus. For example, frequent phrases can be "data mining", "query processing ", and “data analysis" in data mining area.

In a preprocessing step, we use a sliding window over the document content to generate frequent phrases and lossy-counting [8] to discover the phrases with frequency above ψ, forming a frequent phrase list for each document. Each frequent phrase list contains the frequency that the phrase appears in this document.

Meaningful frequent phrases are defined as the frequent phrases that can describe the document corpus concisely. Frequent phrases extracted in preprocessing step contain a large list of stop word combinations and other meaningless phrases that fail

to capture the specific information in the documents. In order to get the meaningful frequent phrases, we exploit Algorithm 1 to weight the meaningful frequent phrases from the meaningless phrases appropriately.

Algorithm 1 summarizes the process of discovering meaningful frequent phrases for each document. In the processing step, a frequent phrase list has been formed for each document that contains the frequency of the phrases. We integrate these lists and add the frequency of the same phrases together to form an aggregate frequent phrase list, AFP-list. The input of Algorithm 1 is the frequent phrases in AFP-list, the number N_{con} of documents containing the phrase, total number N_{docs} of documents in document corpus, and the number m of meaningful frequent phrases that we want to compute. Our algorithm processes all the frequent phrases in AFP-list, and computes a relevance score with the formula in line 4. The formula boosts the meaningful frequent phrases, and filters the meaningless frequent phrases, with the similar idea with TF· IDF. Through this algorithm, we get the top m meaningful frequent phrases, (mfp_1, …, mfp_m).

Algorithm 1. Meaningful Frequent Phrases Extraction

Input: $\mathbf{N_{con}}$: number of documents containing the phrase; $\mathbf{N_{docs}}$: total number of documents in the corpus; **m**: number of meaningful frequent phrases; AFP-list

Output : queue: priority queue of <phrase, relevance>

1 **for** each frequent phrase in AFP-list **do**
2 $\text{phrase}_f \leftarrow$ frequency of the frequent phrase in AFP-list
3 $\text{phrase}_{len} \leftarrow$ sum of the phrase frequencies in AFP-list
4 $\text{relevance} \leftarrow \frac{\text{phrase}_f}{\text{phrase}_{len}} * \log \frac{N_{docs}}{N_{con}}$
5 insert <phrase, relevance> in queue
6 **if** queue has more than m items **then**
7 delete item from queue with the smallest relevance
8 **end if**
9 **end for**
10 **Return** the queue with m meaningful frequent phrases

3.2 LSA Based on Document-Phrase Matrix

LSA model [9] projects documents as well as terms into a low-dimensional semantic space by carrying out Singular Value Decomposition (SVD) [10], producing a set of topics associated with the documents and terms.

In our model, documents are represented as a set of meaningful frequent phrases extracted by Algorithm 1. Consider the analysis of document-phrase matrix, if there are a total of n documents and m meaningful frequent phrases in a document collection, phrase-document matrix, $A = [a_{ij}]_{m \times n}$, is constructed, with each entry a_{ij} representing the TF·IDF score of the *i-th* meaningful frequent phrase in the *j-th* document. For the matrix A, where without loss of generality $m \geqslant n$ and rank(A) = r, the SVD is defined as:

$$A = U \sum V^T \tag{1}$$

where $U = [u_1, u_2, ..., u_r]$ is a m×r column-orthonormal matrix whose columns are called left singular vectors; $\sum = diag[\sigma_1, \sigma_2, ..., \sigma_r]$ is an r×r diagonal matrix, whose diagonal elements are non-negative singular values sorted in descending order; $V = [v_1, v_2, ..., v_r]$ is an n×r column-orthonormal matrix whose columns are called right singular vectors.

Given an integer k (k<<r), LSA uses the first k singular vectors to represent the documents and meaningful frequent phrases in a same k-dimensional semantic space in which similar objects are expected to be near to each other. More precisely, each document is represented by a row of $[\sigma_1 v_1, \ldots, \sigma_k v_k]$ and each meaningful frequent phrase is represented by a row of $[\sigma_1 u_1, \ldots, \sigma_k u_k]$.

From the k-dimensional latent semantic space, we can get the inherent topics associated with papers and meaningful frequent phrases. LSA provides a simplified solution to model topics of documents in a text-rich heterogeneous information network. However, this model ignores the associated heterogeneous information network as well as other interacted objects, so it cannot model and make use of associated objects simultaneously.

3.3 Topic Propagation

In this section, we introduce a novel and general topic propagation method that effectively incorporates the heterogeneous information network with the textual documents for topic modeling, so as to estimate the topics of documents as well as the associated objects and improve the topic modeling results simultaneously.

To obtain the topics of other objects, a simple way is to propagate the topics from documents to other objects through the heterogeneous information network as shown in the dashed rectangle of Figure 2.

Principle. In a heterogeneous information network, the topic of an object without text content (e.g. u_1, v_1 in Figure 1) is decided by the topics of its connected documents. On the other hand, the topic of a document is affected by the estimated topics of its connected objects.

For example, the research topic of an author could be decided by his/her published papers. Conversely, the topic of a paper is influenced by its authors to a certain degree.

Definition 3. In k-dimensional semantic space, let matrix $[p(d_1), \ldots, p(d_n)]^T$ represents the papers, where the values of k-dimensional vector $p(d_i)$, $p(d_i)_j$ $(j = 1 \ldots k)$, represent the weights of paper d_i in the k topics; let matrix $[p(u_1), \ldots, p(u_l)]^T$ represents the authors, where the values of k-dimensional vector $p(u_i)$, $p(u_i)_j$ $(j = 1 \ldots k)$, represent the weights of author u_i in the k topics; let matrix $[p(v_1), \ldots, p(v_o)]^T$ represents the venues, where the values of k-dimensional vector $p(v_i)$, $p(v_i)_j$ $(j = 1 \ldots k)$, represent the weights of venue v_i in the k topics; let matrix $[p(mfp_1), \ldots, p(mfp_m)]^T$ represents the meaningful frequent phrases, where the values of k-dimensional vector $p(mfp_i)$, $p(mfp_i)_j$ $(j = 1 \ldots k)$, represent the weights of phrase mfp_i in the k topics.

Take Figure 1 for example, let us see how the topics propagate from papers to their neighboring objects, e.g., authors and venues. Given the initial vector of papers $p(d_i)$ in a certain k-dimensional topic space, the vector of an author $p(u_i)$ can be calculated by:

$$p(u_i) = \sum_{d_i \in D_u} \frac{p(d_i)}{|D_u|} \tag{2}$$

where D_u is the set of documents that are connected with author u, and $|D_u|$ is the number of these documents. Similarly, the vector of a venue $p(v_i)$ can be expressed as:

$$p(v_i) = \sum_{d_i \in D_v} \frac{p(d_i)}{|D_v|} \tag{3}$$

where D_v is the set of documents that are published in venue v.

On the other hand, let us see how the topics propagate from these objects without text content, e.g., authors and venues, to the papers, so as to reinforce the topics of papers. Along with the vectors of authors and venues calculated above, the vectors of papers can be reinforced by:

$$p(d_i) = \zeta p(d_i) + \frac{1-\zeta}{2}\left(\sum_{u \in U_d} \frac{p(u_i)}{|U_d|} + \sum_{v \in V_d} \frac{p(v_i)}{|V_d|}\right) \tag{4}$$

where U_d is the set of authors of document d, and V_d is the venues connected with document d. Here, ζ is the harmonic parameter that controls the balance between the inherent topics and the propagated topics.

According to Eqs. (2), (3) and (4), we calculate the ultimate vectors of different objects, e.g., papers, authors and venues, in the k-dimensional latent semantic space through M iterations (M is a maximum iteration based on empirically study). The topic propagation process can be summarized as Algorithm 2. Note that, if ζ=1, the topics of documents remain the original ones, while the topics of other objects are determined by their associated documents in one step.

Algorithm 2. Topic propagation

Input: $\boldsymbol{p(d_i)}$: the vector of paper; $\boldsymbol{p(u_i)}$: the vector of author; $\boldsymbol{p(v_i)}$: the vector of venue; $\boldsymbol{A}$: the set of authors; $\boldsymbol{V}$: the set of venues; $\boldsymbol{D}$: the set of papers; $\mathbf{t}$: the number of iterations; $\mathbf{M}$: a maximum iteration; $\boldsymbol{D_u}$: the set of documents connecting with author u; $\boldsymbol{D_v}$: the set of documents published in venue v; $\boldsymbol{U_d}$: the set of authors of document d; $\boldsymbol{V_d}$: the venues connected with document d

Output: the ultimate vectors $p(d_i)$, $p(u_i)$, $p(v_i)$

1 The initial values of $p(d_i)^{(0)}$ is obtained by LSA; t ⟵ 1

2 **do**

3 **for** each author in A and each venue in V **do**

4 $$p(u_i)^{(t)} = \frac{1}{|D_u|}\sum_{d_i \in D_u} p(d_i)^{(t-1)}, \quad p(v_i)^{(t)} = \frac{1}{|D_v|}\sum_{d_i \in D_v} p(d_i)^{(t-1)}$$

5 end for

6 **for** each paper in D **do**

7 $$p(d_i)^{(t-1)'} = p(d_i)^{(t-1)}, \quad p(d_i)^{(t-1)} = \zeta p(d_i)^{(t-1)'} + \frac{1-\zeta}{2}\left(\sum_{u \in U_d} \frac{p(u_i)^{(t)}}{|U_d|} + \sum_{v \in V_d} \frac{p(v_i)^{(t)}}{|V_d|}\right)$$

8 end for

9 t ⟵ t+1

10 **while** t ≤ M

11 **End while**

12 **Return** $(p(d_i), p(u_i), p(v_i))$

According to the similarity calculation of the ultimate vectors of papers, we can get the paper clusters. Analogously, we also can obtain the author clusters and venue clusters.

In the previous section, each meaningful frequent phrase has been mapped to the latent semantic space with the vector $p(mfp_i)$ by LSA. Select the centroid of each paper cluster, the Euclidean distance of each meaningful frequent phrase to the centroid can be calculated by:

$$dist(p(mfp_i),\ p(c_j)) = \sqrt{\sum_{z=1}^{k} (p(mfp_i)_z - p(c_j)_z)^2} \tag{5}$$

where vector $p(c_j)$ represents the centroid of each paper cluster, $p(c_j)_z$ represents the dimensional value to the corresponding topic.

According to formula (5), we assign each meaningful frequent phrase to the paper cluster with the smallest Euclidean distance. Now each paper cluster can be treated as an optimized topic associated with both papers and meaningful frequent phrases. The topics are modeled as a set of meaningful frequent phrases.

3.4 Topic Description

Although the discovered topics are modeled as a set of meaningful frequent phrases, in order to make them more understandable, it is necessary to calculate a topic description for each topic. The topic description is expected to basically capture the meaning of the topic.

For each topic, we pick the meaningful frequent phrase with the biggest cosine similarity with the documents in that topic as a topic description. For the document $p(d_i)$ and meaningful frequent phrase $p(mfp_j)$ in each topic, cosine similarity between them can be calculated: $sim(p(mfp_j),\ p(d_i)) = \frac{\sum_{z=1}^{k} p(mfp_j)_z\, p(d_i)_z}{\sqrt{\sum_{z=1}^{k} p(mfp_j)_z^2}\sqrt{\sum_{z=1}^{k} p(d_i)_z^2}}$.

For each meaningful frequent phrase, we add together its cosine similarity with the documents that belong to the same topic. Obviously, topic description is the meaningful frequent phrase with the biggest cosine similarity summation in each topic.

4 Experiments

In this section, we apply LSA-PTM to the real data and show its effectiveness over the state-of-the-art models.

4.1 Dataset and Metric

We evaluate the effectiveness LSA-PTM on Digital Bibliography and Library Project (DBLP) dataset [11]. In our experiments, we use a DBLP subset that belongs to four

areas, i.e. database (DB), data mining (DM), information retrieval (IR) and artificial intelligence (AI), and contains 1200 documents (abstracts and titles), 1576 authors and 8 conferences. We extract 1660 unique meaningful frequent phrases from this data collection with threshold ψ =5. A heterogeneous information network can be built on this dataset in which there are three types of objects: papers with text content, authors and venues without text content, and two types of relationships: paper-author and paper-venue, which consist of a total number of 4339 links. Note that we set the number of topics (k) to be 4.

To quantitatively identify the effectiveness of LSA-PTM, we adopt *F measure* as our metric [12]. The *F-measure* combines the *Precision* and *Recall* together. In our experiments, there are four categories. For each topic cluster, we calculate the *Precision* and *Recall* with regard to each given category. Specifically, for the obtained cluster j and the given category i : $Precision\ (i, j) = \frac{n_{ij}}{n_i}$, $Recall\ (i, j) = \frac{n_{ij}}{n_j}$, $F\ (i, j) = \frac{2 \times Precision\ (i, j) \times Recall\ (i, j)}{Precision\ (i, j) + Recall\ (i, j)}$, 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, n_j is the number of members in cluster j and $F(i, j)$ is the F measure of cluster j and category i. The *F-measure* of the whole clustering results is defined as a weighted sum over all the categories as follows: $F = \sum_i \frac{n_i}{n} \max_j \{F\ (i, j)\}$, where the max is taken over all clusters.

4.2 Experiment Results

We first analyze the topic modeling results with case studies. Then we discuss how to set the harmonic parameter of LSA-PTM to achieve the best performance. Finally, experiments are conducted to compare the performance of object clustering with different models.

Topic Analysis and Case Study

In our model, we set the harmonic parameter ζ = 0.9 and the iterations M = 9, the topic modeling results are shown in Table 1. Each discovered topic is modeled as a set of meaningful frequent phrases with a topic description in bold. For comparison, we conduct PLSA [5] and TMBP-Regu [7] on DBLP dataset. The most representative terms generated by PLSA and TMBP-Regu are shown in Table 2 and Table 3 respectively. Contrast of Table 1, Table 2 and Table 3, it is more easily to understand the meanings of the four topics by the topic descriptions, i.e., “database systems”, “data mining”, “information retrieval” and “artificial intelligence”, derived from LSA-PTM. For the first three topics of PLSA and TMBP-Regu, all these terms can describe the topics to some extent. For Topic 4, the topic description, “artificial intelligence”, derived from LSA-PTM is obviously more telling than “problem, algorithm, paper” derived by PLSA and “learning, based, knowledge” by TMBP-Regu. Thus, from the view of readability of the topics, LSA-PTM is better than PLSA as well as TMBP-Regu.

Table 1. The topic representation generated by LSA-PTM

Topic 1	Topic 2	Topic 3	Topic 4
database systems	**data mining**	**information retrieval**	**artificial intelligence**
database system	clustering algorithms	language model	knowledge-based algorithms
database management	classification algorithms	web search	machine learning
distributed databases	data cube	retrieval performance	pattern recognition
relational databases	data analysis	search engine	knowledge engineering
database structure	knowledge discovery	retrieval models	user interface
relational data model	mining problems	search results	knowledge based systems
query processing	data warehouse	semantic search	expert systems

The bold phrase in each topic is the topic description.

Table 2. The representative terms generated by PLSA

Topic 1	Topic 2	Topic 3	Topic 4
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

Table 3. The representative terms generated by TMBP-Regu

Topic 1	Topic 2	Topic 3	Topic 4
data	data	information	learning
database	mining	web	based
query	algorithm	retrieval	knowledge
databases	clustering	search	model
systems	classification	based	problem
queries	based	text	reasoning
system	algorithms	language	system
processing	rules	user	logic

Parameter Analysis

In our method, there exists a harmonic parameter ζ, in this section, we will study and evaluate the effect of the parameter ζ.

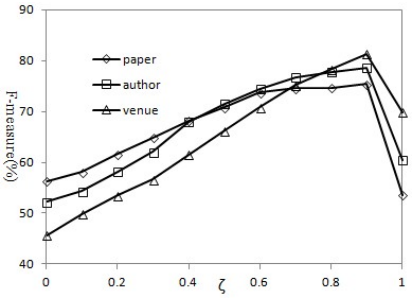

Fig. 3. The effect of varying parameter ζ in LSA-PTM

As mentioned in section 3.3, the harmonic parameter is used to control the balance between the inherent topics of documents from LSA and the propagated topics. When $\zeta = 1$, it is the LSA model. Figure 3 shows the performance of LSA-PTM with the varied harmonic parameter. We can see that the performance is improved over the LSA model when incorporating topic propagation on the heterogeneous information network with $\zeta < 1$. Note that with the decrease of ζ, the performance turns worse, and even worse than the LSA model, because the model relies more on the topic consistency while ignores the intrinsic topics of the documents. According to Figure 3, we set the harmonic parameter $\zeta = 0.9$ in other experiments.

Clustering Performance Comparison

We apply our model on the task of object clustering using DBLP dataset. The hidden topics extracted by the topic modeling approaches can be regarded as clusters. We calculate *F-measure* with the provided category to evaluate the clustering results. We compare the proposed LSA-PTM with the state-of-the-art methods: latent semantic analysis (LSA) [9], probabilistic latent semantic analysis (PLSA) [5], author-topic model (ATM) [13]. For each method, 15 test runs were conducted, and the final performance scores were obtained by averaging the scores from the 15 tests. The italic results of LSA in Table 3 are obtained by LSA-PTM with $\zeta = 1$.

Table 4. Clustering performance of different methods on DBLP

Metric (%)	F-measure			
Object	Paper	Author	Venue	Average
LSA	53.59	*60.40*	*69.9*	61.30
PLSA	58.45	-	-	58.45
ATM	64.00	61.13	-	62.56
LSA-PTM	75.35	78.61	81.36	78.44

As shown in Table 3, the proposed LSA-PTM achieves the best overall performance. This shows that integrating heterogeneous information network with topic modeling by topic propagation, LSA-PTM can have a better topic modeling power for clustering objects.

5 Related Work

Topic modeling has attracted much attention recently in multiple types of text mining tasks, such as information retrieval [14, 15], geographical topic discovery [16], and extract scientific research topics [17, 18].

Topic modeling on heterogeneous information networks was paid little attention in existing literatures. The latest study, TMBP-Regu [7], first directly incorporates heterogeneous information network and textual documents with topic modeling. But the topic modeling results of TMBP-Regu are presented by world distributions like most of existing topic models [3, 4, 5, 6]. Although the discovered topics interpreted

by the top keywords in the word distributions are intuitively meaningful, it is still difficult for a user to accurately comprehend the meaning of each topic. In our model, the discovered topics are modeled as a set of meaningful frequent phrases accompanied with topic descriptions that successfully solve the problem of intelligibility of discovered topics.

Many topic models, such as Latent Semantic Analysis (LSA) [9], Probabilistic Latent Semantic Analysis (PLSA) [5] and Latent Dirichlet Allocation (LDA) [6], have been successfully applied to or extended to many data analysis problems, including document clustering and classification [2], author-topic modeling [13]. However, most of these models merely consider the textual documents while ignoring the network structures. Recently, several proposed topic models, such as LaplacianPLSI [2], NetPLSA [3] and iTopicmodel [4], have combined topic modeling and network structures, but they only emphasize on the homogeneous networks, such as document network and co-authorship network, but not heterogeneous information networks. Our proposed model takes the heterogeneous network structures into topic modeling by topic propagation between textual objects and other types of objects. Experimental results are desirable.

6 Conclusion and Future Work

In this paper, LSA-PTM is proposed for topic modeling heterogeneous information networks and also solves the readability of the topic modeling results. First, LSA-PTM extracts meaningful frequent phrases from documents. Then latent semantic analysis is conducted on these phrases, so as to obtain the inherent topics of the documents. Moreover, we introduce a topic propagation method that optimizes the inherent topics using heterogeneous network structures between different objects and ultimately enhances the topic modeling. To make the topics more understandable, a topic description is calculated for each topic. Experimental results show the effectiveness of LSA-PTM. Our future work will apply LSA-PTM on large scale dataset to examine and verify its efficiency.

Acknowledge. This work was supported by the National Key Technologies R&D Program (Grant No.2012BAH54F04), the National Natural Science Foundation of China (Grant No.61003051), and the Natural Science Foundation of Shandong Province of China (Grant No.ZR2010FM033).

References

1. Sun, Y., Han, J., Yan, X., Yu, P.: Mining Knowledge from Interconnected Data: A Heterogeneous Information Network Analysis Approach. In: Proceedings of the VLDB Endowment, pp. 2022–2023. ACM Press (2012)
2. Cai, D., Mei, Q., Han, J., Zhai, C.: Modeling hidden topics on document manifold. In: Proceedings of the 17th ACM Conference on Information and Knowledge Management, pp. 911–920. ACM Press, New York (2008)

3. Mei, Q., Cai, D., Zhang, D., Zhai, C.: Topic modeling with network regularization. In: Proceedings of the 17th International Conference on World Wide Web, pp. 101–110. ACM Press, New York (2008)
4. Chen, Sun, Y., Han, J., Yu: itopicmodel: Information network-integrated topic modeling. In: Proceedings of Ninth IEEE International Conference on Data Mining, pp. 493–502. IEEE Computer Society, Miami (2009)
5. Hofmann, T.: Probabilistic latent semantic indexing. In: Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 50–57. ACM, New York (1999)
6. Blei, D.M., Ng, A.Y., Jordan, M.I.: Latent dirichlet allocation. The Journal of Machine Learning Research, 993–1022 (2003)
7. Deng, H., Han, J., Zhao, B., Yu, Y.: Probabilistic topic models with biased propagation on heterogeneous information networks. In: Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 1271–1279. ACM Press, New York (2011)
8. Simitsis, A., Baid, A., Sismanis, Y., Reinwald, B.: Multidimensional content exploration. The VLDB Endowment, 660–671 (2008)
9. 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)
10. Golub, G.H., Reinsch, C.: Singular value decomposition and least squares solutions. Numerische Mathematik, 403–420 (1970)
11. DBLP, `http://www.informatik.uni-trier.de/~ley/db/`
12. Larsen, B., Aone, C.: Fast and Effective Text Mining Using Linear-time Document Clustering. In: Proceedings of the Fifth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 16–22. ACM Press, New York (1999)
13. Steyvers, M., Smyth, P., Rosen-Zvi, M., Griffiths, T.L.: Probabilistic author-topic models for information discovery. In: Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 306–315. ACM Press, New York (2004)
14. Wei, X., Croft, W.B.: LDA-based document models for ad-hoc retrieval. In: Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, pp. 178–185. ACM Press, New York (2006)
15. Zhai, C., Lafferty, J.: Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of the Tenth International Conference on Information and Knowledge Management, pp. 403–410. ACM Press, New York (2001)
16. Yin, Z., Cao, L., Han, J., Zhai, C.: Geographical topic discovery and comparison. In: Proceedings of the 20th International Conference on World Wide Web, pp. 247–256. ACM Press, New York (2011)
17. He, D., Parker, D.S.: Topic dynamics: an alternative model of bursts in streams of topics. In: Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 443–452. ACM Press, New York (2010)
18. He, Q., Chen, B., Pei, J., Qiu, B., Mitra, P.: Detecting topic evolution in scientic literature: how can citations help? In: Proceedings of the 18th ACM Conference on Information and Knowledge Management, pp. 957–966. ACM Press, New York (2009)