Co-topic Network Method and Applications Postprint

Authors: Niu Liang

Date: 2017-11-08T00:00:00+00:00

Abstract

[Objective] By constructing a co-topic network, analyzing the relationships between topics, and optimizing the terms contained within topics. [Method] A co-topic network is generated from the "document-topic" bipartite graph according to weighted projection rules, key topics in the co-topic network are measured using a method that combines betweenness centrality and topic probability, community segmentation of the topic network is performed through the GN algorithm, and topic terms are optimized using a correlation method. [Results] Comparison between the co-topic network and the JSD-based K-means method reveals that under tests with three topic numbers (optimal topic number 28 and random topic numbers 20, 30), both produce identical numbers of clusters, with consistency degrees of cluster content reaching 100%, 95%, and 87%, respectively. [Limitations] Other community segmentation methods for co-topic networks have not been comprehensively covered. [Conclusion] The co-topic network accommodates the needs of high-dimensional data and can detect which topics in documents are important and which topics are closely connected.

Full Text

Preamble

ChinaXiv Collaborative Journal, Issues 272/273, 2016, Issues 7/8
Co-topic Network Method and Application*
Niu Liang (School of Economics & Management, China Jiliang University, Hangzhou 310018, China)

Abstract

[Objective] This study constructs a co-topic network to analyze relationships between topics and optimize the terms contained within each topic. **[Methods]**

We transform a "document-topic" bipartite graph into a co-topic network using weighted projection rules, measure key topics in the co-topic network by combining betweenness centrality with topic probability, apply the GN algorithm for community segmentation of the topic network, and optimize topic terms using a relevance method. **[Results]** Comparing co-topic networks with JSD-based K-means methods across three topic numbers (optimal 28 and subjective selections 20, 30) reveals identical clustering numbers in all cases, with clustering content consistency reaching 100%, 95%, and 87% respectively. **[Limitations]** Other community segmentation methods for co-topic networks were not comprehensively examined. **[Conclusions]** The co-topic network accommodates high-dimensional data requirements and can detect which topics are important in documents and which topics are closely connected.

Keywords: Co-topics network; LDA; Community partition; K-means

Classification Number: G250

Introduction

The utilization of scientific literature resources has long been a focus of academic attention. Previous research typically employed co-word analysis methods for bibliometric analysis of scientific literature, concentrating on improvements to analytical objects, metrics, and visualization techniques [1]. However, co-word analysis struggles to discover latent semantic connections within documents, failing to meet users' deep-seated needs for scientific information. The LDA topic model proposed in natural language processing [2] quickly gained traction in bibliometric analysis due to its semantic-focused term allocation [3-4], leading to classic extensions such as the AT model [5], TOT model [6], and CTM model [7].

Despite LDA' s achievements in scientific literature mining, traditional LDA models suffer from two significant problems. First, trained LDA models produce topics that lack interconnections. Traditional LDA often explains documents by selecting the single topic with the highest probability distribution, yet a document may embody multiple topics simultaneously. Consequently, traditional LDA fails to explain relationships among co-occurring topics or identify which topics are more important for document interpretation. Some studies have introduced principal component analysis and clustering methods to express topic relationships, compressing multi-dimensional topics into two dimensions for visualization via multidimensional scaling [8-9]. These clustering approaches employ KL divergence [5,10], Jensen-Shannon Divergence (JSD) [11-13], or cosine similarity [14] for distance measurement. However, principal component analysis oversimplifies high-dimensional data by compressing it into two dimensions, ignoring complexity. Moreover, selecting appropriate distance metrics for clustering remains problematic. Complex networks effectively accommodate high-dimensional data and can address clustering through community detection methods. Therefore, when exploring topic relationships, we construct a

document-topic bipartite network based on topic co-occurrence in documents and project it into a topic network for relationship detection. Literature combining topic models with complex networks is scarce: references [10] and [15] treat each user in collaboration networks as a document and all collaborators as terms, using topic models for user clustering—essentially preparing network data for LDA. Reference [16] uses community detection modularity as a hidden variable to enhance LDA performance, similar to AT and TOT models, without further discussing generated topic relationships.

Second, traditional LDA topic term composition frequently includes unimportant or weakly associated terms. Model outputs require domain experts to identify and modify meaningful terms [17-18], preventing automatic term selection [19-20]. Some models introduce external latent variables to improve term precision, such as the AT model incorporating author information or the TOT model treating time as a continuous observable variable. However, since these extended models share LDA's generative mechanism, they still cannot avoid producing irrelevant terms. Addressing these limitations, this paper accomplishes two objectives: (1) constructing a co-topic network through document-topic bipartite graph projection to detect topic relationships and important topics via community segmentation and centrality measurement; (2) optimizing topic term selection to identify terms most relevant to each topic.

2.1 LDA Topic Model

LDA is an unsupervised machine learning method employing the Bag-of-Words representation, where each document is treated as a term frequency vector, transforming textual information into easily modeled numerical data. LDA modeling assumes each document represents a probability distribution over topics, while each topic represents a probability distribution over terms. With T topics, the probability of term w_i in a given document is:

$$P(w_i) = \sum_{j=1}^{T} P(w_i|z = j)P(z = j)$$

where z is a latent variable indicating that term w_i originates from topic j. $P(w_i|z = j)$ represents the probability that term w_i belongs to topic j, while $P(z = j)$ gives the probability that topic j belongs to the current document. Intuitively, $P(w|z)$ reveals which terms are important for a topic, whereas $P(z)$ represents the topic distribution within a document. Topic content is reflected in $P(w|z)$, and document composition depends on the topic distribution $P(z)$. For instance, if a journal publishes papers in "statistical learning" and "machine learning" sections, the term probability distribution centers on these topics, with content reflected in $P(w|z)$ and sections reflected in $P(z)$. This composition of documents from topics and topics from terms constitutes a standard Bayesian classification problem.

Given D documents containing T topics (predetermined through iterative testing) composed of W independent terms from a vocabulary, $P(w|z)$ corresponds to a multinomial distribution for each topic over the W terms, denoted as Φ, i.e., $P(w|z) = \Phi$. The D documents in the corpus correspond to a multinomial distribution over T topics, denoted as θ, where for a given document d, $P(z) = \theta$. With term set $w = \{w_1, w_2, \dots, w_n\}$ where each w_i belongs to specific document d_i, LDA generates document d by repeatedly sampling a topic z from the document' s multinomial distribution θ, then sampling a term w from topic z' s multinomial distribution Φ, repeating this process N_d times where N_d is the total number of terms in document d. Two parameters require inference: the "document-topic" distribution θ and the "topic-term" distribution Φ, primarily through EM algorithms or Gibbs sampling.

2.2 Co-topic Network Construction

(1) Document-Topic Bipartite Graph and Projection

Most networks are single-mode networks composed of one node type. However, bipartite networks exist where nodes belong to different sets and edges connect nodes from different types. For document-topic bipartite networks, one node type represents documents and the other represents topics. Formalizing the document-topic bipartite network: let $G = \langle V, E \rangle$ with $X \cup Y = V$ and $X \cap Y = \emptyset$, such that each edge in G connects one endpoint from X and one from Y, denoted as $\langle X, Y, E \rangle$. Here, X represents documents and Y represents topics selected for each document, with the selection rule choosing topics whose probability exceeds the average in document X to build the document-topic bipartite network.

Bipartite networks are rarely analyzable without transformation because most network measures are designed for single-mode graphs. Therefore, bipartite graphs must be projected into single-mode graphs. The projection rule for node set X states: if any two nodes in X both connect to some node in Y, those two nodes in X are linked. The projection rule for node set Y works analogously. Figure 1 [Figure 1: see original paper] illustrates a bipartite graph and its projections, where (a) shows the bipartite graph, (b) shows the projection of X nodes, and (c) shows the projection of Y nodes.

Single-mode projection methods are practical and widely used but lose much structural information from the original bipartite network. The mapping only connects similar vertices without considering how many groups they share. Assigning weights to projections can preserve such information, typically by setting edge weights between two vertices in the projected network to the number of shared groups from the other set [21]. Newman argued that in scientist-paper bipartite graphs converted to scientist networks, group counts ignore author contribution levels, and author weights should vary with the number of coauthors per paper [22]. Zhou et al. noted that Newman' s method [22] overlooks the importance of single authors in projections and designed weights based on

resource allocation impacts [23].

Our projection rule treats topics as nodes, establishing an edge between two topics if they appear in the same document, indicating a relationship. If a document contains n topics, it generates $n(n-1)/2$ pairwise relationships. When two specific topics T_1 and T_2 co-occur in multiple documents, we avoid duplicate connections between T_1 and T_2, but this ignores document nodes in the bipartite graph. To reflect document nodes while showing topic connection strength, we define the weight between topics T_1 and T_2 as:

$$w_{T_1T^2} = \sum_{k=1}^{g} \frac{\delta_{kT_1} \cdot \delta_{kT_2}}{n_k - 1}$$

where g represents the number of documents, δ_{kT_1} equals 1 if topic T_1 appears in document k (otherwise 0), and n_k represents the number of topics in document k.

Figure 2 [Figure 2: see original paper] provides an example: topics T_1 and T_2 appear together in three documents containing 4, 2, and 3 topics respectively. Their relationship strengths are 1/3, 1, and 1/2, yielding a total relationship strength of $1/3 + 1 + 1/2 = 11/6$, which becomes the edge weight between T_1 and T_2. Following this projection rule, we generate a weighted co-topic network through single-mode projection of topics.

(2) Node Importance Calculation

Co-topic network node importance detection comprises two aspects: node centrality based on network topology and topic probability based on term distributions from the LDA model. Centrality measures include degree centrality, betweenness centrality, closeness centrality, eigenvector centrality, and k-shell/k-core [24]. This study combines betweenness centrality with topic term probability distribution to measure topic importance. Betweenness centrality identifies nodes that lie on the shortest paths between other nodes, making them crucial for establishing relationships. However, betweenness alone cannot reflect the intrinsic importance of non-bridge nodes, necessitating incorporation of non-topological properties. For topic networks, this non-topological attribute is the term probability distribution constituting the topic content. Therefore, we combine betweenness centrality with term probability distribution when calculating topic node importance.

The betweenness centrality of node V_i is the number of shortest paths passing through that node across the network:

$$C_B(i) = \sum_{u \neq i \neq w} \frac{\sigma_{uw}(i)}{\sigma_{uw}}$$

where σ_{uw} is the number of shortest paths between nodes V_u and V_w, and $\sigma_{uw}(i)$ is the number of those paths passing through node V_i. Higher betweenness indicates more central positioning, with nodes not appearing on any shortest path receiving a centrality of 0.

The topic term probability distribution for node V_i is:

$$P(i) = \sum_{d \in D} \theta_{di}$$

where θ_{di} is the multinomial distribution of topic node V_i for document d, and D is the document collection.

The final strength of topic nodes in the co-topic network combines betweenness centrality and topic probability distribution. Due to differing scales, we apply Min-Max normalization to create dimensionless values:

$$V_i = c \times \left[\frac{B(i) - \min(B(i))}{\max(B(i)) - \min(B(i))} + \frac{P(i) - \min(P(i))}{\max(P(i)) - \min(P(i))} \right]$$

where c is a tuning coefficient controlling visualization display size. Larger V_i values indicate greater topic importance and correspond to larger node areas in the topic network visualization.

(3) Co-topic Network Clustering

While co-topic networks reveal pairwise topic relationships and highlight important topics through edge weights, identifying whether multiple topics belong to the same category requires clustering. For co-topic networks, we employ community detection techniques. Common community partitioning methods include graph algorithms (spectral bisection, random walk, clique percolation) and hierarchical clustering (agglomerative and divisive). The former includes the CNM algorithm based on greedy agglomerative principles [25], while the latter includes the GN algorithm based on edge betweenness [26]. CNM suits large-scale networks, whereas GN works for small-to-medium networks. Since co-topic networks have few nodes and edges, we adopt the GN algorithm.

GN is a divisive community detection algorithm that progressively removes inter-community edges based on the principle of high intra-community cohesion and low inter-community cohesion, yielding relatively cohesive community structures. The algorithm uses edge betweenness to detect edge positions, calculating betweenness centrality for all edges, removing the edge with maximum betweenness, and iteratively recalculating until edge betweenness falls below threshold μ. The pseudocode is:

```
Input: A weighted or unweighted graph G = (V, E), Threshold
Output: A list of clusters
while |E(G)| > 0 do
```

```
    Calculate C_{u,v} for all (u, v)   E(G)
    maxBetweennessEdge = (x, y): C_{x,y} is maximum over all (x, y) in E(G)
    maxBetweennessValue = C_{x,y}
    if maxBetweennessValue $\geq$   then
        E(G) = E(G) - {maxBetweennessEdge}
    else
        Break out of loop
return Connected components of modified G
```

Since GN iteratively removes edges with maximum betweenness, it cannot automatically determine termination points, suffers from high time complexity due to repeated shortest path calculations, and cannot theoretically predict the final number of communities. Community count determination requires threshold μ tuning. Therefore, a metric is needed to evaluate results across different μ values. Newman introduced modularity Q to assess community partition quality [27-28]. However, modularity only measures relative quality because computing truly optimal modularity is computationally hard [29]. Thus, we consider the partition with maximum Q value as the ideal network division. Q ranges from 0 to 1, with higher values indicating more accurate community structure. The modularity Q formula is:

$$Q = \frac{1}{2m} \sum_{i,j} \left[A_{ij} - \frac{k_i k_j}{2m} \right] \delta(c_i, c_j)$$

where A_{ij} is the element in row i and column j of adjacency matrix A, k_i and k_j are degrees of nodes i and j, m is the total number of edges in the network, and $\delta(c_i, c_j)$ equals 1 if nodes i and j are in the same community (otherwise 0). The GN algorithm optimizes modularity Q directly, merging communities stepwise to maximize Q increase or minimize decrease. The peak Q_{max} during this process corresponds to the optimal community structure.

2.3 Topic Term Optimization

LDA topic models infer two parameters: the "document-topic" distribution θ and the "topic-term" distribution Φ. While co-topic networks address θ relationships, network node topics depend on Φ, making Φ optimization critical for topic interpretability. Traditional LDA outputs often contain unimportant or weakly associated terms, requiring domain expert intervention [17-18] and preventing automatic term selection [19-20]. Some models enhance term precision through external latent variables (e.g., AT introducing authors, TOT incorporating time), but these extensions, sharing LDA' s generative mechanism, still produce irrelevant terms.

To improve automatic term selection and exclude incoherent terms, reference [30] uses an intrinsic metric called "Lift" to rank topic terms, defined as the ratio of a term' s probability within a topic to its marginal probability across

the corpus. Reference [31] ranks terms by frequency and exclusivity. Reference [9] combines lift and exclusivity into a relevance method, where term-topic relevance is regulated by parameter λ. When λ approaches 1, frequently occurring terms are more relevant (as in [30]); when λ approaches 0, exclusive terms are more relevant (as in [31]). Given reference [9]' s precision in term generation, we construct topic term network nodes using this relevance method:

$$r(w,k|\lambda) = \lambda \log(\phi_{kw}) + (1-\lambda)\log\left(\frac{\phi_{kw}}{p_w}\right)$$

where λ determines the weight of lift in term w' s relevance to topic node k in the co-topic network. If $\lambda = 0$, term ranking depends entirely on lift (the ratio of term probability in topic ϕ_{kw} to corpus-wide probability p_w). If $\lambda = 1$, ranking depends solely on topic term probability ϕ_{kw}. Reference [9] recommends $\lambda = 0.6$.

3.1 Bibliographic Data Acquisition and Preprocessing

We retrieved data from CNKI' s "China Academic Journal Network Publishing Database," searching for the journal *Journal of Academic Libraries* from 1989 to 2015. For computational convenience, we merged downloaded bibliographic files by year, creating 27 annual documents numbered by year. Each document contains records including title, author, affiliation, keywords, and abstract. To analyze core research topics from 1989 to 2015, we focused on titles, keywords, and abstracts. The preprocessing steps were: (1) loop-reading the 27 files; (2) segmenting titles, keywords, and abstracts using the Rjieba package, cleaning irrelevant characters (connectors, null values, numbers, etc.) with regular expressions; (3) constructing a document-term frequency matrix using the tm package and optimizing feature terms through TF-IDF. Subsequent research builds upon this term frequency matrix.

3.2 Co-topic Network Analysis

Building a topic network requires determining the optimal number of topics for LDA. While typically set empirically, too few topics inadequately represent documents while too many cause redundancy. Several methods automatically discover optimal topic numbers, including Bayesian statistical approaches [3], KL divergence [32], cosine similarity [33], and JSD [34]. Due to its simplicity and computational efficiency, we adopt the Bayesian statistical method.

The calculation method appears in formulas (8) and (9), with results shown in Figure 3 [Figure 3: see original paper]. $P(w|z_k)_n$ is the frequency of word w assigned to topic k in random topic z, and n_k is the total number of words assigned to topic k. $P(w|T)$ can be approximated as the harmonic mean of $P(w|z)$ series, calculated as:

$$P(w|T) \approx \left(\frac{1}{S} \sum_{s=1}^{S} \frac{1}{P(w|z^{(s)})} \right)^{-1}$$

where $\Gamma(\cdot)$ is the standard Gamma function.

Selecting $k = 28$ as the maximum topic number, we construct a 27$\times$28"*document*−*topic*"*matrix*(D $\times$T$), named θ in LDA. Each document d comprises k topics, but based on probability distribution, only several important topics represent the document. We select representative topics for each document using the rule of choosing topics with probabilities exceeding the document' s average topic probability. Table 1 shows partial document-topic selection results.

Table 1 Selected Document Topics (Partial)

Document (D)	Topic (T)
D1	T4, T23, T26, T27
D2	T12, T23, T26, T27
D3	T13, T14
D4	T13, T28

Using this data to construct a bipartite graph yields the document-topic bipartite network shown in Figure 4 [Figure 4: see original paper].

To better illustrate topic network clustering, we apply the GN algorithm for community segmentation, using modularity Q to evaluate optimality across different thresholds μ. Modularity peaks at 0.4904142 during the second-to-last iteration, automatically dividing the network into two communities as shown in Figure 5 [Figure 5: see original paper].

Important nodes in the co-topic network are T13, T3, T9, and T23. Larger edge weights indicate more frequent co-occurrence when interpreting documents, reflecting stronger connections. Larger node areas signify greater importance in document interpretation. For example, T13 and T9 share substantial edge weight, indicating frequent co-occurrence, and their large node areas confirm they are important representative topics. To clarify topic importance, we examine the top 30 topic terms as word clouds for T13, T3, T9, and T23 in Figure 6 [Figure 6: see original paper], revealing that digitization, service quality, and library automation have become key focus areas of *Journal of Academic Libraries*.

3.3 Comparison Between Co-topic Network and JSD-based K-means

Kim et al. compared various distance metrics (cosine similarity, Jaccard coefficient, Kendall' s tau, DCG, KL divergence, JSD), finding JSD performs best

for topic distance measurement [35]. To demonstrate the co-topic network' s capability in detecting topic relationships and important nodes, we compare it with the best-performing JSD-based K-means clustering, examining both visualization and multidimensional scaling representations.

JSD is a KL divergence-based metric that improves upon KL' s asymmetry, becoming a common measure for probabilistic topics. The JSD between any two topics T_p and T_q is:

$$JSD(T_p||T_q) = \frac{1}{2}D_{KL}(T_p||M) + \frac{1}{2}D_{KL}(T_q||M)$$

where $M = \frac{1}{2}(T_p + T_q)$ and D_{KL} is KL divergence.

To observe consistency between JSD and co-topic network measurements, we test three topic numbers: the optimal 28 and subjective selections 30 and 20. Co-topic network analysis produces community segmentation Q-value fusion shown in Figure 7 [Figure 7: see original paper]. The black line represents the optimal 28 topics, while red and blue lines represent 30 and 20 topics respectively. All three reach maximum Q-values at the second-to-last iteration (red circles), indicating each co-topic network clusters into two communities. The visualizations appear in the upper half of Figure 8 [Figure 8: see original paper].

To verify whether JSD-based K-means clustering numbers match modularity-based partitions, we determine optimal K-means cluster counts using the Silhouette Coefficient, which ranges from -1 to +1 with higher values indicating better clustering [36]. We enumerate cluster counts k from 2 to 8, running K-means 25 times per k to avoid local optima and calculating average silhouette coefficients. The k with maximum coefficient is selected. Results in Figure 9 [Figure 9: see original paper] show optimal $k = 2$ for all three topic numbers (30, 28, 20), marked by red circles. The silhouette-determined cluster count perfectly matches the modularity-based partition.

Applying K-means with $k = 2$ to 30, 28, and 20 topics (node sizes proportional to topic probability distribution) yields results in the lower half of Figure 8 [Figure 8: see original paper]. Comparing co-topic network communities with K-means clusters reveals high similarity: 100% for 28 topics, 95% for 20 topics, and 87% for 30 topics. Table 2 details the comparison by topic number.

Table 2 Comparison of Clustering Results

Co-topic Network Community Division	JSD-based K-means Clustering
28 topics: 3,4,7,10,11,12,20,23,25,26,27 / 16,17,18,19,21,22,24,28	3,4,7,10,11,12,20,23,25,26,27 / 16,17,18,19,21,22,24,28
20 topics: 1,3,5,6,7,8,11,12,14,15,20,22,23,24,25,26,29 / 2,9,16,19,4,10,13,28,30,17,18,21,27	1,2,3,5,6,7,8,9,11,12,14,15,16,19,20,22,23,24,25,26,29 / 4,10,13,28,30,17,18,27

Co-topic Network Community Division	JSD-based K-means Clustering
30 topics: 1,2,4,5,7,8,10,11,15,16,17,20 / 3,6,9,12,13,14,18,19	1,4,5,7,8,10,11,15,16,17,20 / 2,3,6,9,12,13,14,18,19

The optimal 28-topic network produces identical results between co-topic network community division and JSD-based K-means clustering because optimal topics are non-redundant. Subjective selections of 30 and 20 topics show minor deviations but remain largely consistent. For instance, in 20-topic clustering, topic 2 appears in different clusters, but its terms discuss library digitization issues similar to its clustered neighbors. Moreover, co-topic networks establish weighted edge relationships showing which topics co-occur to interpret documents—something JSD-based clustering cannot explain. Additionally, betweenness centrality enables detection of topics used across most documents, a capability absent in JSD clustering.

4 Results and Discussion

This paper achieves two goals through co-topic network analysis of scientific literature: (1) establishing network relationships among topics to address the lack of connections in traditional LDA-generated topics; (2) optimizing topic term selection and identifying relevant terms visualized as word clouds.

Co-topic network analysis differs from principal component-based topic relationship exploration by accommodating high-dimensional data needs, eliminating distance metric selection dilemmas, and detecting which topics are closely linked and co-occur in document interpretation.

Study limitations include: (1) Journal articles contain not only textual information but also author data. How can we combine topic and author information to analyze topic evolution? While the AT model [5] addresses single authors studying multiple topics, future research should explore how collaboration networks affect studied topics to identify knowledge communities. (2) Co-word networks [37] are commonly used for scientific literature structural analysis. Future work should examine the theoretical differences, connections, and distinctions between co-topic networks and co-word networks in scientific literature analysis.

References

[1] Tang Guoyuan, Zhang Wei. Development and Analysis of Co-word Analysis Method at Home and Abroad [J]. Library and Information Service, 2014, 58(22): 138-145.

[2] Blei D M, Ng A Y, Jordan M I, et al. Latent Dirichlet Allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

[3] Griffiths T L, Steyvers M. Finding Scientific Topics [J]. Proceedings of the

National Academy of Sciences of the United States of America, 2004, 101(1): 5228-5235.

[4] Sugimoto C R, Li D, Russell T G, et al. The Shifting Sands of Disciplinary Development: Analyzing North American Library and Information Science Dissertations Using Latent Dirichlet Allocation [J]. Journal of the American Society for Information Science and Technology, 2011, 62(1): 85-204.

[5] Rosen-Zvi M, Griffths T, Steyvers M, et al. The Author-topic Model for Authors and Documents [C]. In: Proceedings of the 20th Conference on Uncertainty in Artificial Intelligence.

[6] Wang X, McCallum A. Topics Over Time: A Non-Markov Continuous-Time Model of Topical Trends [C]. In: Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2006: 424-433.

[7] Blei D M, Lafferty J D. A Correlated Topic Model of Science [J]. The Annals of Applied Statistics, 2007, 1(1): 17-35.

[8] Mimno D. Computational Historiography: Data Mining in a Century of Classics Journals [J]. Journal on Computing and Cultural Heritage, 2012, 5 (1): 1-19.

[9] Sievert C, Shirley K E. LDAvis: A Method for Visualizing and Interpreting Topics [C]. In: Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces.

[10] Zhang H, Qiu B, Giles C L, et al. An LDA-based Community Structure Discovery Approach for Large-scale Social Networks [C]. In: Proceedings of the 2007 IEEE International Conference on Intelligence and Security Informatics. 2007.

[11] Wang X, Zhang K, Jin X, et al. Mining Common Topics from Multiple Asynchronous Text Streams [C]. In: Proceedings of the 2nd ACM International Conference on Web Search and Data Mining. 2009.

[12] Newman D, Asuncion A, Smyth P, et al. Distributed Algorithms for Topic Models [J]. Journal of Machine Learning Research, 2009, 10(12): 1801-1828.

[13] Gretarsson B, O'Donovan J, Bostandjiev S, et al. TopicNets: Visual Analysis of Large Text Corpora with Topic Modeling [J]. Transactions on Intelligent Systems & Technology, 2012, 3(2): 565-582.

[14] He Q, Chen B, Pei J, et al. Detecting Topic Evolution in Scientific Literature: How Can Citations Help? [C]. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management. 2009.

[15] Cha Y, Cho J. Social-network Analysis Using Topic Models [C]. In: Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2012.

[16] Li D, He B, Ding Y, et al. Community-based Topic Modeling for Social Tagging [C]. In: Proceedings of the 19th ACM International Conference on Information and Knowledge Management. 2010.

[17] Chuang J, Ramage D, Manning C D, et al. Interpretation and Trust: Designing Model-Driven Visualizations for Text Analysis [C]. In: Proceedings of the 2012 SIGCHI Conference on Human Factors in Computing Systems. 2012.

[18] Hall D, Jurafsky D, Manning C D. Studying the History of Ideas Using Topic Models [C]. In: Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing. 2008.

[19] Chang J, Boyd-Graber J, Wang C, et al. Reading Tea Leaves: How Humans Interpret Topic Models [R]. Advances in Neural Information Processing Systems 22 (NIPS 2009).

[20] Mimno D, Wallach H M, Talley M, et al. Optimizing Semantic Coherence in Topic Models [C]. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing. 2011.

[21] Latapy M, Magnien C, Del Vecchio N. Basic Notions for the Analysis of Large Two-mode Networks [J]. Social Networks, 2008, 30(1): 31-48.

[22] Newman M E J. Scientific Collaboration Networks. I. Network Construction and Fundamental Results [J]. Physical Review E, 2001, 64(1): 016131.

[23] Zhou T, Ren J, Medo M, et al. Bipartite Network Projection and Personal Recommendation [J]. Physical Review E, 2007, 76(4): 046115.

[24] Ren Xiaolong, Lv Linyuan. Review of Ranking Nodes in Complex Networks [J]. Chinese Science Bulletin, 2014, 59(13): 1175-1197.

[25] Newman M E J. Fast Algorithm for Detecting Community Structure in Networks [J]. Physical Review E, 2004, 69(6): 066111.

[26] Girvan M, Newman M. Community Structure in Social and Biological Networks [J]. Proceedings of the National Academy of Sciences, 2002, 99(12): 7821-7826.

[27] Clauset A, Newman M E J, Moore C. Finding Community Structure in Very Large Networks [J]. Physical Review E, 2004, 70(6): 066111.

[28] Newman M E J. Modularity and Community Structure in Networks [OL]. ArXiv: physics/0602124v1.

[29] Brandes U, Delling D, Gaertler M, et al. Maximizing Modularity is Hard [OL]. arXiv: Physics/0608255.

[30] Taddy M A. On Estimation and Selection for Topic Models [C]. In: Proceedings of the 18th International Conference on Artificial Intelligence and Statistics. 2015.

[31] Bischof J M, Airoldi E M. Summarizing Topical Content with Word Frequency and Exclusivity [C]. In: Proceedings of the 29th International Conference on Machine Learning. Omnipress.

[32] Arun R, Suresh V, Madhavan V C E, et al. On Finding the Natural Number of Topics with Latent Dirichlet Allocation: Some Observations [A]. // Advances in Knowledge Discovery and Data Mining [M]. Springer Berlin Heidelberg, 2010.

[33] Cao J, Xia T, Li J, et al. A Density-based Method for Adaptive LDA Model Selection [J]. Neurocomputing, 2008, 72(7-9): 1775-1781.

[34] Deveaud R, SanJuan E, Bellot P. Accurate and Effective Latent Concept Modeling for Ad Hoc Information Retrieval [J]. Document Numérique, 2014, 17(1): 61-84.

[35] Kim D, Oh A. Topic Chains for Understanding a News Corpus [C]. In: Proceedings of the 12th International Conference on Computational Linguistics and Intelligent Text Processing. 2011.

[36] Zhu Lianjiang, Ma Bingxian, Zhao Xuequan. Clustering Validity Analysis Based on Silhouette Coefficient [J]. Journal of Computer Application, 2010, 32(S2): 139-141.

[37] Wang Xiaoguang. Formation and Evolution of Science Knowledge Network (I): A New Research Method Based on Co-word Network [J]. Journal of the China Society for Scientific and Technical Information, 2009, 28(4): 599-605.

Support Data

Support data [1-3] are self-archived by the author, E-mail: niutyut@126.com; **support data [4-7]** are available in the journal' s online version at http://www.infotech.ac.cn.

[1] Niu Liang. data_{preprocessing}.R. Segmentation processing of titles, keywords, and abstracts from *Journal of Academic Libraries* bibliographic files, construction of document-term frequency matrix, and calculation of optimal topic numbers.

[2] Niu Liang. CoTopic_{network}.R. Construction of document-topic bipartite graphs, generation of weighted co-topic networks, measurement of key topics, modularity calculation, and community segmentation for three topic numbers (28, 30, 20).

[3] Niu Liang. JSD_K-means.R. JSD-based K-means silhouette coefficient calculation and topic clustering visualization for three topic numbers (28, 30, 20).

[4] Niu Liang. JAL.rar. Bibliographic literature of *Journal of Academic Libraries* (1989-2015) downloaded from CNKI.

[5] Niu Liang. ldamodel.rar. Topic modeling data and document term frequency data under three topic numbers (28, 30, 20) for co-topic network and K-means

clustering.

[6] Niu Liang. modularity.rar. Modularity data under three topic numbers (28, 30, 20) for determining optimal community segmentation numbers.

[7] Niu Liang. silhouette-coefficient.rar. Silhouette coefficient data under three topic numbers (28, 30, 20) for determining optimal clustering numbers.

Conflict of Interest Statement: The author declares no conflict of interest.

Received Date: 2016-03-09
Revised Date: 2016-05-09

Note: Figure translations are in progress. See original paper for figures.