

Text Network Exploration via Heterogeneous Web of Topics

Junxian He*, Ying Huang[†], Changfeng Liu*, Jiaming Shen[‡], Yuting Jia[‡], Xinbing Wang*

*Department of Electronic Engineering, Shanghai Jiao Tong University, Shanghai, China

[†]Department of Automation, Shanghai Jiao Tong University, Shanghai, China

[‡]Department of Computer Science and Engineering, Shanghai Jiao Tong University, Shanghai, China

{junxian831, hy941001, stevenevets, sjm940622, hnxxjyt, xwang8}@sjtu.edu.cn

Abstract—A text network refers to a data type that each vertex is associated with a text document and the relationship between documents is represented by edges. The proliferation of text networks such as hyperlinked webpages and academic citation networks has led to an increasing demand for quickly developing a general sense of a new text network, namely *text network exploration*. In this paper, we address the problem of text network exploration through constructing a heterogeneous web of topics, which allows people to investigate a text network associating word level with document level. To achieve this, a probabilistic generative model for text and links is proposed, where three different relationships in the heterogeneous topic web are quantified. We also develop a prototype demo system named *TopicAtlas* to exhibit such heterogeneous topic web, and demonstrate how this system can facilitate the task of text network exploration. Extensive qualitative analyses are included to verify the effectiveness of this heterogeneous topic web. Besides, we validate our model on real-life text networks, showing that it preserves good performance on objective evaluation metrics.

I. INTRODUCTION

The information age has witnessed an increasing amount of unstructured data, most of which are in the form of text and possess high degrees of connectivity among themselves. We refer to this type of data as *text network* as shown in Figure 1a. Such text networks are ubiquitous in the real world. Typical representatives include hyperlinked webpages, online social network with user profiles, and academic citation network.

With the rapid increase of available text networks, the demand for exploring them quickly has continued to grow. When faced with a new or unfamiliar text network, people may first ask a basic question: “What is there?”. To answer this question, we resort to the notion of *exploratory search* [1] which is proposed to help people develop a general sense of the properties of new text network before embarking on more specific inquiries [2].

Due to its importance, exploratory search has been investigated intensively. For example, Sinclair et al. use word frequency lists, frequency distribution plots and keyword-in-context models to enhance computer-assisted reading [3]. More recently, a computational technique named “topic modeling” achieves great success through providing insight into a corpus’ contents [2], [4], [5]. Nevertheless, existing topic models are still far from adequate for text network exploration since the significance of topics represented only by words is

limited for exploration task without an insight on document level.

To address this problem, we view each document as a “bag of links” [6], [7]. For example, in the academic paper network, a paper with k references is viewed as a document with k “link tokens”. Then, we can model these documents within a topic model framework where a new type of “topics” characterized by distributions over documents arises and important documents are assigned with high probabilities. By combining “word token” and “document token”, each document is composed of two parts as shown in Figure 1b, and two different types of topics are included as illustrated in Figure 1c. To distinguish the two categories of topics, we call them *WordTopic* and *DocTopic* respectively.

However, it is still inconvenient to explore a text network since users can only inspect the individual topic in isolation. Therefore, we expect to uncover the relations between topics to enable users to examine not only a topic itself but also the related fields and important documents. With that in mind, a complete heterogeneous topic web which displays three different types of relationships as described in Figure 1d is indispensable. Although the relationship between WordTopics (*Word-Word relation*) has been investigated previously [7]–[14], the connections between DocTopic and DocTopic (*Doc-Doc relation*) and WordTopic and DocTopic (*Word-Doc relation*) have not been studied before.

To construct such heterogeneous topic web, we propose a probabilistic generative model called **MHT** (Model for Heterogeneous Topic web), where all three relationships are quantified. Our experiments on two academic citation networks demonstrate that MHT not only produces reliable heterogeneous topic web with high-quality topics but also possesses strong generalizability and predictive power.

Furthermore, we build **TopicAtlas**, a prototype demo system for convenient navigation in heterogeneous topic web. TopicAtlas displays Word-Word relation, Doc-Doc relation, and Word-Doc relation in a unified framework. With TopicAtlas, users are able to freely wander around the text network via WordTopics and DocTopics.

To summarize, our contributions are three folds:

- To the best of our knowledge, we are the first to present the idea of heterogeneous web of topics and construct it successfully.

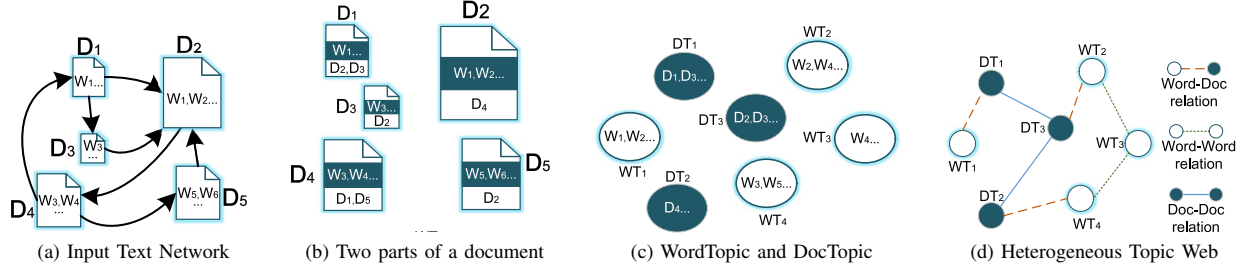


Fig. 1. Illustration of some concepts. (a) Input text network. (b) Two parts of a document. W represents the “word token” part, and D below W represents “document token” part. (c) WordTopic (WT) and DocTopic (DT). (d) Heterogeneous topic web with two types of topics and three types of relationships.

- We propose MHT, a probabilistic generative model that helps extract two types of topics along with their heterogeneous relationships.
- We develop TopicAtlas, a prototype system for text network exploration. TopicAtlas allows users to investigate the heterogeneous topic web with details and explore text network easily.

The rest of paper is organized as follows. In section II, we discuss some related works. We introduce MHT and its inference in section III and IV. In section V, we conduct the experiment and evaluate our model. Finally, we summarize this paper and discuss some future works in section VI.

II. RELATED WORK

In terms of exploratory search. When dealing with large collections of digitized historical documents, very often only little is known about the quantity, coverage and relations of its content. In order to get an overview, exploring the data beyond simple “lookup” approaches is needed. The notion of exploratory search has been introduced to cover such cases [1].

Chaney and Blei [15] make an early effort in exploratory search via visualizing traditional topic models, where a navigator of documents is created and allows users to explore the hidden structure. Gretarsson et al. build a relatively mature system called TopicNets [4], which enables users to visualize individual document sections and their relations within the global topic document. Maiya et al. [16] build the topic similarity network for exploration and recognize how topics form large themes. Recently, Jahnichen et al. develop a complete framework in this field [17], they depict probability distributions as tag clouds and permit the identification of related topic groups or outliers.

While the works mentioned above convey some information visually, these approaches consider the data as isolated-document corpus rather than linked text networks. With only text they cannot conduct a serious analysis for a text network on a document level. Specifically, although some of them are able to retrieve *topic-related* documents, there is no possibility for them to identify *topic-significant* documents, which are more crucial in exploratory search. Therefore, we introduce DocTopic and propose the idea of heterogeneous topic web to enable users to keep track of related topic groups, relevant documents and significant documents.

In terms of topic modeling. Topic models are proposed to address the problem of topic identification in large document collections. In topic models, each document is associated with a topic distribution and each topic is associated with a word distribution. Two popular models in this field are Probabilistic Latent Semantic Analysis (PLSA) [18] and Latent Dirichlet Allocation (LDA) [19]. They are both generative and unsupervised models, introducing latent topics into the generative process.

However, traditional topic models only consider text and ignore the significant link information. Recently, some variants of topic models are proposed for jointly analyzing text and links. A major part of them models the link information as evidence of content similarity between two documents [10]–[12], [20]–[24], but this kind of approach is not able to detect important documents with respect to a specific topic. Another categories of methods which generate the links from DocTopics can recognize significant documents [6], [7], [9], [25], [26]. But these works fail to construct a complete heterogeneous topic web composed of WordTopic, DocTopic and three different types of relations among them. Although the connection between WordTopics has been investigated before [7]–[14], to the best of our knowledge, we are the first to model two types of topics and three types of relations jointly and build the heterogeneous topic web successfully.

III. MODEL FOR HETEROGENEOUS TOPIC WEB

In this part we describe the framework and generative process of MHT (Model for Heterogeneous Topic web), whose graphical representation is illustrated in Figure 2.

A. Framework

We consider the input text network as a graph $G(V, E)$, where V is the set of document vertices and E is the set of directed edges or links. $v_i \in V$ represents the i^{th} document and $e_{ij} \in E$ connects two vertices v_i and v_j . Each document is associated with a bag of words and a bag of links. We denote w_{in} as the n^{th} word token in document v_i , and y_{il} expresses the l^{th} link token (document token) in v_i .

In classical topic models each document is associated with a document specific topic distribution, which is used to draw a topic for each word in the generative process. Note that the “topic” here actually represents WordTopic in our notation

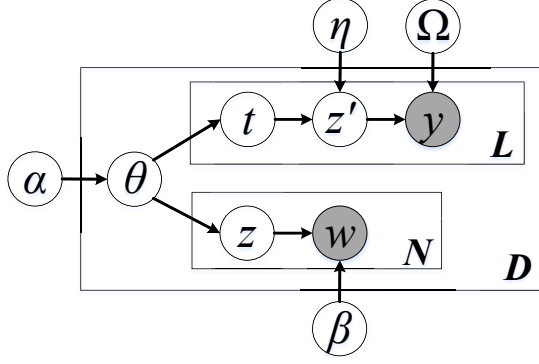


Fig. 2. Graphical Representation of MHT

framework. Similarly, adopt the assumption of “bag of links” and each document is associated with a DocTopic distribution, which can generate linked documents. Since these two distributions are totally different, some transition procedure between them is required to jointly model text and links.

Based on the discussion above, we employ a transition distribution over DocTopics η to depict the relation between the two types of topics.

B. Generative Process

Details for full generative process of our proposed model MHT are demonstrated below.

For each document v_i , where $i = 1, \dots, D$:

- 1) Generate WordTopic distribution: $\theta_i \sim \text{Dir}(\cdot|\alpha)$
- 2) For each word w_{in} , where $n = 1, \dots, N_i$:
 - a) Draw a WordTopic: $z_{in} \sim \text{Mult}(\cdot|\theta_i)$
 - b) Draw a word: $w_{in} \sim \text{Mult}(\cdot|\beta_{z_{in}})$
- 3) For each link y_{il} , where $l = 1, \dots, L_i$:
 - a) Draw a transition topic: $t_{il} \sim \text{Mult}(\cdot|\theta_i)$
 - b) Draw a DocTopic: $z'_{il} \sim \text{Mult}(\cdot|\eta_{t_{il}})$
 - c) Draw a linked document: $y_{il} \sim \text{Mult}(\cdot|\Omega_{z'_{il}})$

Step 1 and Step 2 are the same as classical topic model to generate words. A major distinction of MHT from other models is Step 3, where we employ a transition latent variable t as an “intermediary” from WordTopic domain to DocTopic domain. In this transition stage, we introduce a transition parameter η to express the relation between WordTopic and DocTopic so that the generation of DocTopic is equivalent to drawing it from $\theta\eta$. Thus η serves as a transition matrix from θ to a “spurious” underlying mixed DocTopic distribution θ' . More specifically, for a given WordTopic k , the value of $\eta_{kk'}$ indicates the probability for generating DocTopic k' , i.e. $p(z' = k'|z = k) = \eta_{kk'}$. With that in mind, we can see how η works on transforming WordTopic domain into DocTopic domain.

IV. MODEL LEARNING

To learn MHT, we resort to the variational EM inference method. For each document v_i , we use a fully factorized variational distribution to approximate the posterior distribution:

$$q(\theta_i, z_i, t_i, z'_i) = q(\theta_i|\gamma_i) \prod_{n=1}^{N_i} q(z_{in}|\phi_{in}) \times \prod_{l=1}^{L_i} q(t_{il}|\lambda_{il}) \prod_{l=1}^{L_i} q(z'_{il}|\sigma_{il}), \quad (1)$$

where $q(\theta_i|\gamma_i)$ is Dirichlet distribution and $q(z_{in}|\phi_{in})$, $q(t_{il}|\lambda_{il})$ and $q(z'_{il}|\sigma_{il})$ are all multinomial distributions. Then we will try to maximize the evidence lower bound defined by:

$$\text{ELBO} = \sum_{i=1}^D (\mathbb{E}_q[\log p(\theta_i, z_i, t_i, z'_i, \mathbf{w}_i, \mathbf{y}_i|\alpha, \eta, \beta, \Omega)] - \mathbb{E}_q[\log q(\theta_i, z_i, t_i, z'_i)]), \quad (2)$$

In the E-step, we update γ, ϕ, λ and σ iteratively to approximate the posterior distribution. Then, in the M-step, α, β, η and Ω are renewed to maximize ELBO. Due to the limitation of space, we only provide crucial equations here.

$$\phi_{ink} \propto \beta_{kx} \exp(\Psi(\gamma_{ik})). \quad (3)$$

$$\gamma_{ik} = \alpha_k + \sum_{n=1}^{N_i} \phi_{ink} + \sum_{l=1}^{L_i} \lambda_{ilk}. \quad (4)$$

$$\lambda_{ilk} \propto \exp(\Psi(\gamma_{ik}) + \sum_{k'=1}^{K_\eta} \sigma_{ilk'} \log \eta_{kk'}). \quad (5)$$

$$\sigma_{ilk'} \propto \Omega_{k'd} \exp(\sum_{k=1}^{K_w} \lambda_{ilk} \log \eta_{kk'}). \quad (6)$$

$$\beta_{kx} \propto \sum_{i=1}^D \sum_{n=1}^{N_i} w_{in}^x \phi_{ink}. \quad (7)$$

$$\eta_{kk'} \propto \sum_{i=1}^D \sum_{l=1}^{L_i} \sigma_{ilk'} \lambda_{ilk}. \quad (8)$$

$$\Omega_{k'd} \propto \sum_{i=1}^D \sum_{l=1}^{L_i} y_{il}^d \sigma_{ilk'}. \quad (9)$$

Here, $\Psi(\cdot)$ is the digamma function, $w_{in}^x = 1$ if $w_{in} = x$, and 0 otherwise. Likewise, $y_{il}^d = 1$ if $y_{il} = d$, and 0 otherwise. α is updated by Newton-Raphson algorithm, the interested readers may refer to [19].

First, for each document, we execute step (3) to (6) iteratively until convergence. And then we update α, β, η and Ω . The whole process is in an outer loop until the lower bound ELBO converges.

V. EXPERIMENTS

In this section, we demonstrate how our proposed system – TopicAtlas effectively explores text networks. For repeatability, the codes, datasets, results and the demo TopicAtlas are available to the public¹.

¹<https://River459.github.io/research/>

A. Datasets

We use the following two datasets in our experiments:

ACL Anthology Network (AAN). AAN [27] is a public scientific literature dataset in the Natural Language Processing (NLP) field with 20,989 abstracts of papers and 125,934 citations.

CiteseerX. CiteseerX² is a well-known scientific literature digital library that primarily focuses on the literature in computer and information science. We collect a subset of CiteseerX dataset, which includes the abstracts of 716,800 documents and 1,760,574 links.

B. Parameter Setting

On the task of exploring heterogeneous topic web, we first need to select a reasonable topic number, which is a non-trivial task in topic models. To achieve this, we first preprocess the data using classical LDA model with varying topic numbers and evaluate the topic interpretability in terms of the topic coherence score [28]. Among the candidate topic numbers 50, 70, 90, 110, 130, and 150, topic number 70 leads to the highest topic coherence score for both AAN and CiteseerX. For simplicity, we set the topic number of WordTopic and DocTopic equal. Therefore, we implement MHT with 70 WordTopics and 70 DocTopics to explore the text networks in the two datasets. In addition, we follow the convention of [29] and initialize $\alpha = 0.01$. The parameters η , β and Ω are randomly initialized since we do not have any prior knowledge.

Furthermore, as discussed above, we use variational EM inference to learn the parameters in MHT. In our experiments, for both datasets the inner variational inference loop terminates when the fractional increase of ELBO is less than 10^{-9} in two successive iterations, or the number of iterations exceeds 100. For the outer EM loop, we stop it when the relative increment ratio is less than 10^{-4} , or the number of iterations exceeds 50.

C. Heterogeneous Topic Web Construction

We use co-occurrence probability to quantify the strength of the three types of relations in heterogeneous topic web.

Word-Word Relation Strength. Since we assume the generation of WordTopics is independent with each other when the document v is given, the Word-Word relation strength can be calculated as follows:

$$p(z_1 = k_1, z_2 = k_2 | D) = \sum_{z'} \sum_i p(z' | D) p(v_i | z'; D) \times p(z_1 = k_1 | v_i; D) \times p(z_2 = k_2 | v_i; D), \quad (10)$$

where $p(z | v; D)$ and $p(v | z'; D)$ can be obtained from θ and Ω respectively. Posterior expectation of θ is given by:

$$\theta_{ik} = \frac{\#(v = i, z = k) + \alpha_k}{\sum_{k^*=1}^{K_w} (\#(v = i, z = k^*) + \alpha_{k^*})}, \quad (11)$$

²<http://citeseer.ist.psu.edu/oai.html>

where $\#(v = i, z = k)$ represents the number of words assigned with WordTopic k in document v_i and the assignment can be obtained from ϕ . K_w is the number of WordTopics.

In addition, the empirical posterior distribution over DocTopics can be computed as:

$$p(z' = k' | D) = \frac{\#(z' = k')}{\sum_{k^*} \#(z' = k^*)}, \quad (12)$$

where $\#(z' = k')$ represents the number of documents assigned with DocTopic k' and can be obtained from σ .

Doc-Doc Relation Strength. Based on the assumption that DocTopics are generated independently given a WordTopic, we can compute Doc-Doc relation strength as:

$$p(z'_1 = k'_1, z'_2 = k'_2 | D) = \sum_z p(z | D) p(z'_1 = k'_1 | z; D) \times p(z'_2 = k'_2 | z; D). \quad (13)$$

η represents $p(z' | z; D)$ and similarly the empirical posterior distribution over WordTopics is given by:

$$p(z = k | D) = \frac{\#(z = k)}{\sum_{k^*} \#(z = k^*)}. \quad (14)$$

Word-Doc Relation Strength. Word-Doc relation strength can be easily computed by Bayes' theorem:

$$p(z = k, z' = k' | D) = p(z' = k' | z = k; D) p(z = k | D). \quad (15)$$

Summarizing DocTopic. While top words are able to represent WordTopic explicitly, on the document side there are only distributions over documents to express DocTopics. However, generally it would be preferable to summarize topics with a few words [30]. With that in mind, we leverage the words in abstracts to summarize DocTopics. Specifically, for a given DocTopic k' , we compute the expectancy of word w as:

$$\mathbb{E}(w | z' = k') = \sum_{d=1}^D \Omega_{k'd} \cdot \#(w, d). \quad (16)$$

Then the words with high expectancy are selected as *indicative words* of this DocTopic, which will be displayed in our demo system TopicAtlas.

D. TopicAtlas

We design TopicAtlas based on the constructed heterogeneous topic web. An overview of TopicAtlas is displayed in Figure 3. Aiming to help users navigate in an unfamiliar text network, TopicAtlas has the following features:

Topic Landscape Exhibition. We display top 10 keywords for each WordTopic, and top 5 representative documents and top 10 indicative words for each DocTopic. The diameters of topic vertices express their corresponding *topic dominance* or *topic importance*, which is indicated by $p(z | D)$ for each WordTopic and $p(z' | D)$ for each DocTopic.

Accurate Relationship. The three types of relations correspond to three types of edges in the graph. The weights of these edges are the ratio of the co-occurrence probability we calculate to the prior probability of a random edge (0.0002). The thickness of the edges is proportionate to these values and we remove those whose weights are negligible.

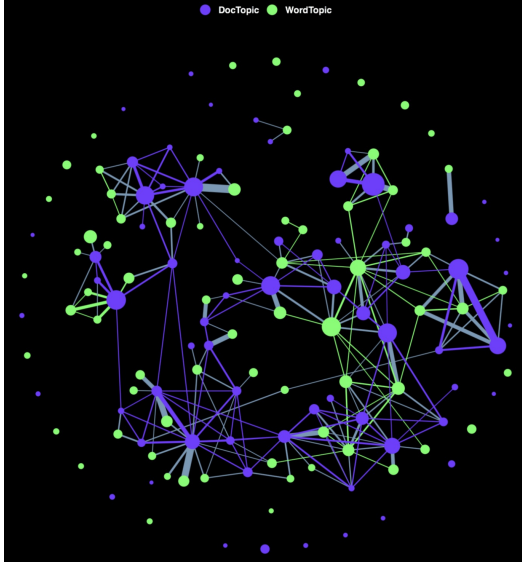


Fig. 3. An overview of TopicAtlas. Different colors indicate different types of topics, and the node size expresses the dominance of corresponding topic. Thickness of edges is proportionate to relation strength (best seen in color).

E. Text Network Exploration via Heterogeneous Topic Web

In this part, we engage in an in-depth exploration of the heterogeneous topic web. To facilitate the analytic reasoning, three auxiliary subgraphs of TopicAtlas are presented here: *Word-Word* subgraph, *Doc-Doc* subgraph and *Word-Doc* subgraph. As the name suggests, Word-Word subgraph only includes the edges between WordTopics, Doc-Doc subgraph contains merely the edges between DocTopics, and Word-Doc subgraph displays edges between WordTopics and DocTopics. Due to the limitation of the space, we only give analysis for CiteseerX here and interested readers can refer to the public demo for the AAN TopicAtlas.

1) **Word-Word Relation:** As shown in Figure 4a, 62.87% of WordTopic nodes have no connection with other WordTopic nodes, which implies that one paper mainly focuses on one WordTopic. This phenomenon agrees with our intuition: most of high quality scientific papers show clear themes.

Though the connection between WordTopics is not strong, there are still a few nodes which link to multiple WordTopics worth investigating. On the basis of previous recognition that the content of documents is generally “pure”, we believe that those WordTopics which enjoy high co-occurrence probability with various other WordTopics are foundation of certain scientific fields. In Figure 4a, WordTopic w45 (degree: 9), w44 (degree: 6), w16 (degree: 5), and w25 (degree: 5) have the highest degrees. The corresponding WordTopics are “distributed system”, “programming language”, “software design”, and “semantic reasoning”. Obviously they are all general and basic. Take “distributed system” as an example, distributed system achieves efficiency improvement of solving computational problems and therefore has broad applications

in different fields such as telephone networks, routing algorithms, network file system, etc. As a case study, we show WordTopic w45 “distributed system” and its related WordTopics in Figure 5.

2) **Doc-Doc Relation:** The DocTopics are closely connected as shown in the Figure 4b, which indicates that authors tend to cover multiple DocTopics in the reference list. It is intuitive because a comprehensive reference section is desired for most authors. Furthermore, since ubiquitous techniques are likely to be cited in a variety of distinct domains, we expect nodes with high degrees in the Doc-Doc subgraph represent DocTopics about universal principle and method. In Figure 4b, the top four highest-degree nodes are DocTopic d63 (degree: 11), d28 (degree: 7), d21 (degree: 7), d17 (degree: 7) and they represent “linear system method”, “logic programming”, “model checking” and “conservation law” respectively. Unsurprisingly, these DocTopics are basic techniques and laws.

In addition to examining DocTopics from a global perspective, inspecting details of specific DocTopic provides insight into a text network on the document level. The DocTopic allows us to assess topic-aware impact of papers since the top documents in one given DocTopic are generally the most popular and representative ones. In Figure 6 we list top 5 documents in the most dominant DocTopic d35 and its neighbours d41, d56, d61.

3) **Word-Doc Relation:** We summarize the contributions of Word-Doc relation from three perspectives. These examples are illustrated in Figure 7.

Connect WordTopic and DocTopic reasonably. As Figure 7 suggests, the DocTopic d17 is about “conservation law”, and its neighbouring WordTopics are w54 “particle phase energy”, w1 “quantum theory” and w55 “equations and solutions”. These topics cover some basic components of quantum mechanics. In addition, WordTopic w36 is about “shared memory processor”, and it has a strong link with DocTopic d44 “shared memory system” and d67 “cache performance”. Also, it connects with DocTopic d20 “power analysis of design” through a edge weighting about 15 since energy reduction plays an important role in shared memory processor. Besides, WordTopic w57 “mobile robot navigation” is connected with DocTopic d49 “mobile robot localization” and d26 “motion planning”. These connections expose the main structure of “mobile navigation”. There are a lot of other examples in our heterogeneous topic web, readers can check them in our demo TopicAtlas.

Link WordTopics indirectly. The missing co-occurrence phenomenon between WordTopics results in difficulty in spotting relevant WordTopics. However, DocTopics can serve as intermediaries between WordTopics and uncover the hidden relationship. More specifically, if two WordTopics co-occur frequently with the same DocTopic, then we can say the two WordTopics are related. For example, WordTopic w43 “image wavelet filter” is connected with WordTopics w13 “dimensional curve reconstruction”, w20 “volume rendering” and w31 “visual motion tracking” through DocTopic d11

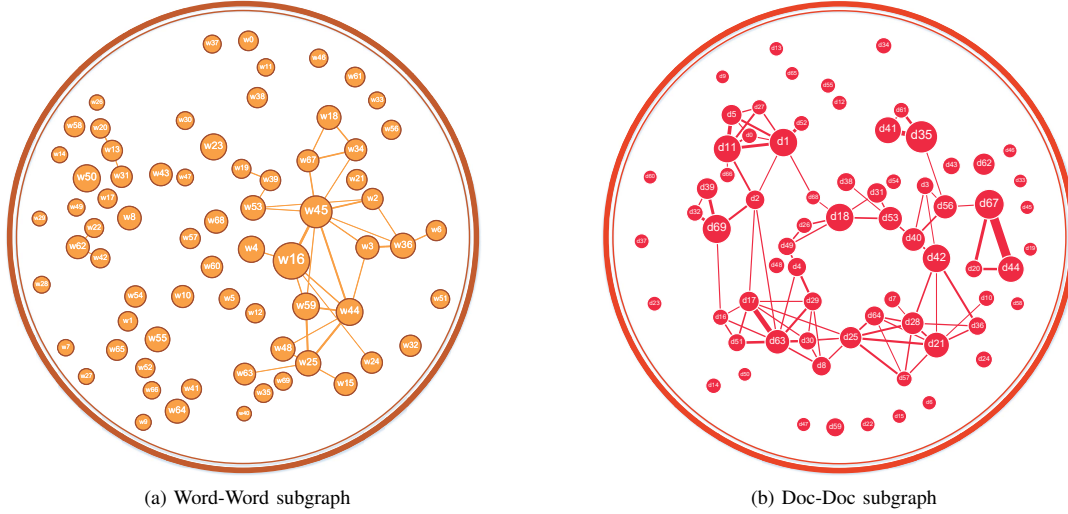


Fig. 4. Subgraphs of heterogeneous topic web: (a) Word-Word Subgraph and (b) Doc-Doc Subgraph (best seen in color). The yellow nodes in the Word-Word subgraph represent WordTopics and the red nodes in the Doc-Doc subgraph represent DocTopics

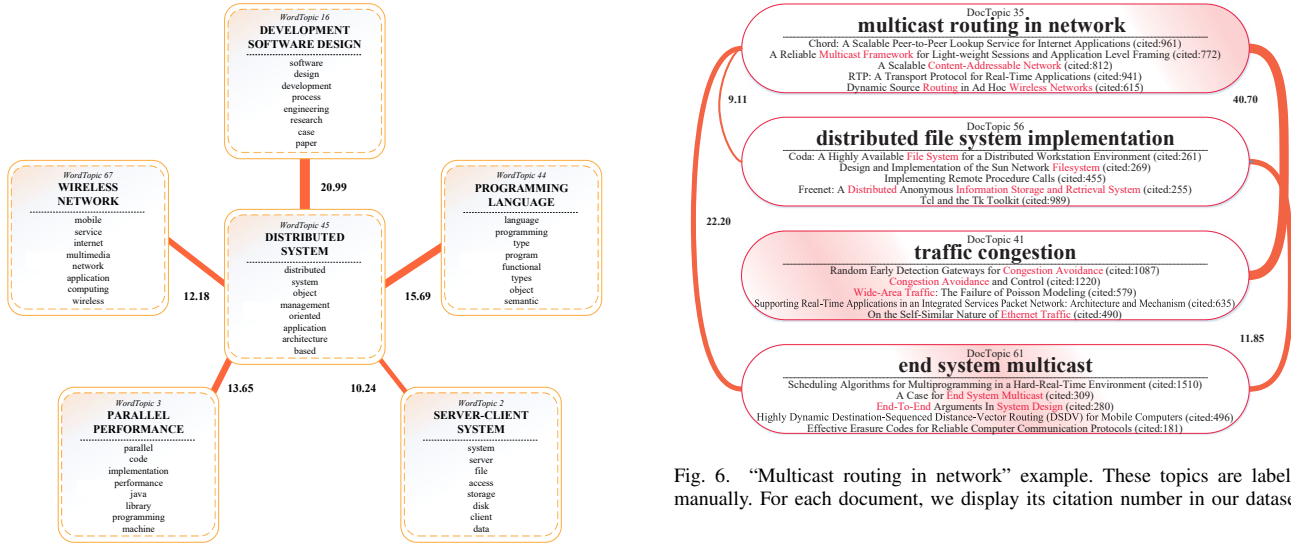


Fig. 5. “Distributed system” example. These topics are labeled manually.

“image based algorithm”, which agrees with the fact that many volume rendering and visual motion tracking models are wavelet-based.

Locate Relevant Documents. Through establishing connection between DocTopics and WordTopics, users can investigate relevant documents for WordTopics. Note that instead of simply recognizing all related documents for WordTopics, TopicAtlas organizes the relevant documents according to DocTopics and allows for inspecting them in different aspects. If a researcher aims to find relevant documents for WordTopic w45 “distributed system”, he can locate papers about the implementation of distributed file or network system in d56,

Fig. 6. “Multicast routing in network” example. These topics are labeled manually. For each document, we display its citation number in our dataset.

examine distributed system architecture stuff in d40, get to know some data management or toolkit documents in distributed system from d54, or explore papers about distribution application in real-time system from d3. With the relevant documents sorted, the researcher is less prone to be swamped by the flood of information.

F. Topic Modeling

Since we aim to obtain effective heterogeneous topic web, it is important to ensure that the introduction of the transition parameter has not come at the expense of the semantic quality of topics and the generalizability of the topic model.

1) Comparative Methods: We compare our method MHT with mixed-membership model (MM) [6], Link-PLSA-LDA [22] and RTM [10], all of which are joint models for both

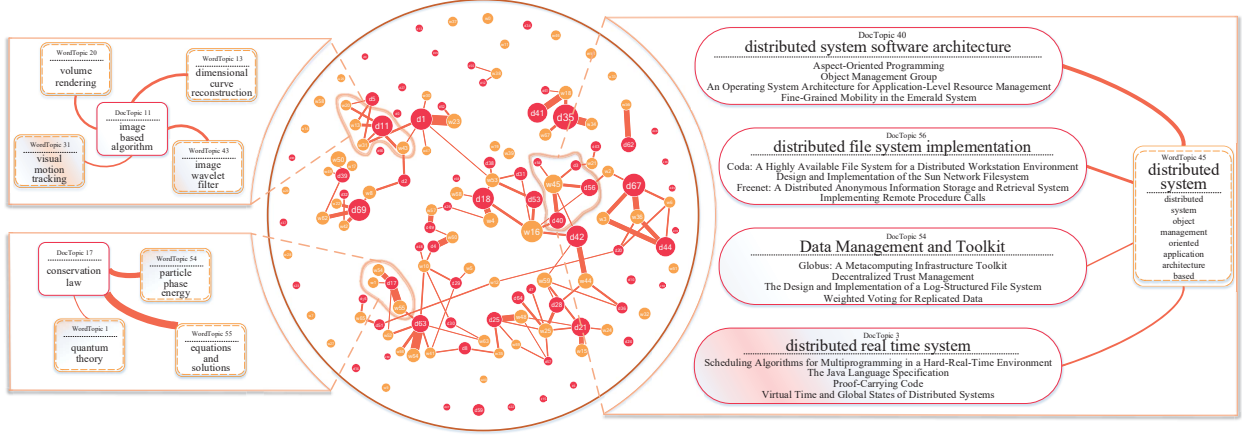


Fig. 7. Word-Doc Subgraph and some instances. The red nodes represent DocTopics and the orange nodes indicate WordTopics. Only the edges between WordTopics and DocTopics are displayed. Doctopic 11 and Doctopic 17 are expressed by indicative words.

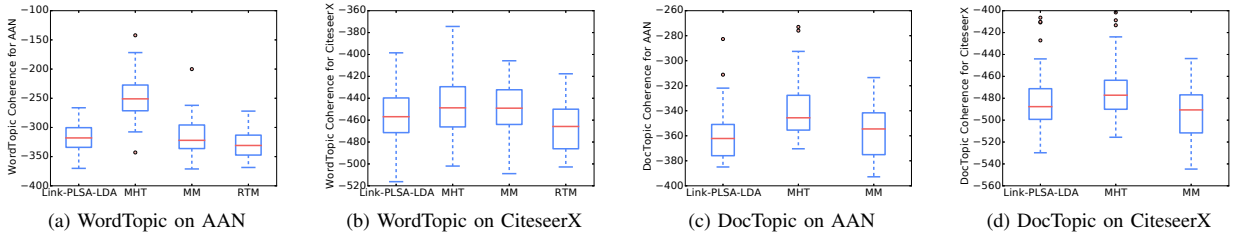


Fig. 8. Topic coherence for WordTopic and DocTopic in two datasets (higher is better).

text and links. Mixed membership model is proposed by Erosheva et al. to classify documents [6]. Nallapati et al. [22] propose two well-known joint topic models Pairwise-Link-LDA and Link-PLSA-LDA. Pairwise-Link-LDA models the presence and absence of links in a pairwise manner while Link-PLSA-LDA views links as “link tokens”. Since Link-PLSA-LDA outperforms Pairwise-Link-LDA with respect to heldout likelihood and recall, we only include Link-PLSA-LDA in our baseline methods. The core idea of RTM is that topic relations directly account for the presence of links. To guarantee the justness, all these models are inferred through variational EM algorithm and parameters are initialized with the same way as MHT.

2) *Topic Interpretability*: There are some metrics for evaluating topic interpretability such as *PMI* [31], *word intrusion* [30], and *topic coherence* [28]. We adopt *topic coherence* in our experiment. For one thing, while word intrusion needs expert annotations, topic coherence is an automated evaluation metric and does not rely on human annotators. For another, topic coherence does not reference collections outside the training data as PMI dose. Also, topic coherence is proven more closely associated with the expert annotations than PMI [28]. Although it is originally designed for WordTopics, by using the indicative words as keywords, we can also

calculate the topic coherence for DocTopics. To distinguish the two different topic coherence score, we denote them as *WordTopic coherence* and *DocTopic coherence*.

We compare the topic coherence score of different methods for all topics, and the results are illustrated in Figure 8. As RTM does not produce DocTopics, it is not included in the comparison. Obviously, our model preserves comparable topic qualities to the baseline methods.

3) *Held-Out Log Likelihood*: Held-out Log Likelihood is a well-accepted metric to measure the generalizability and predictive power of topic models. To ease the favor for text and obtain a convincing result, we filter out the documents with less than 3 links and 8 links for AAN and CiteseerX respectively, and get a collection of AAN with 16,350 documents and CiteseerX with 61,901 documents.

Our experimental set-up is as follows. We randomly split data into five folds and repeat the experiment for five times, for each time we use one fold for test, four folds for training, and we report the average values in Figure 9. The performance of MHT is better than the baseline methods. Note that we exclude RTM in this part since held-out log likelihood favors RTM significantly due to its pairwise manner.

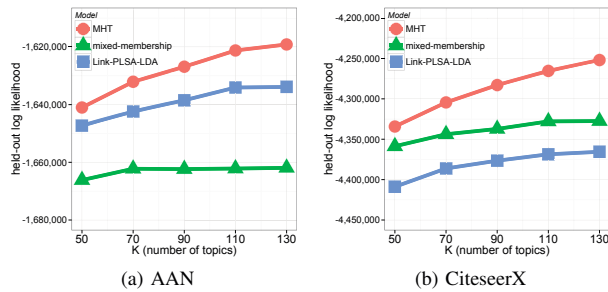


Fig. 9. Held-out log likelihood for both text and links on two datasets. (higher is better)

VI. CONCLUSION

In this paper, we present *MHT*, short for *Model for Heterogeneous Topic Web*, a unified generative model involving two types of topics, namely WordTopic and DocTopic. The relationships between the two types of topics, Word-Word relation, Doc-Doc relation and Word-Doc relation, are quantified, based on which we construct the heterogeneous web of topics. In the experiment, we construct the heterogeneous topic web of AAN and CiteseerX collection and build a prototype demo system, called *TopicAtlas* to exhibit the heterogeneous topic web and assist users' exploration. Qualitative analyses are presented to demonstrate the effectiveness of *TopicAtlas*. Besides, *MHT* shows good performance as a topic model with respect to topic interpretability and held-out log likelihood.

REFERENCES

- [1] G. Marchionini, "Exploratory search: From finding to understanding," *Commun. ACM*, vol. 49, no. 4, pp. 41–46, Apr. 2006. [Online]. Available: <http://doi.acm.org/10.1145/1121949.1121979>
- [2] L. F. Klein, J. Eisenstein, and I. Sun, "Exploratory thematic analysis for digitized archival collections," *Digital Scholarship in the Humanities*, p. fqv052, 2015.
- [3] S. Sinclair, "Computer-assisted reading: Reconceiving text analysis," *Literary and Linguistic Computing*, vol. 18, no. 2, pp. 175–184, 2003.
- [4] B. Gretarsson, J. Odonovan, S. Bostandjiev, T. Höllerer, A. Asuncion, D. Newman, and P. Smyth, "Topicnets: Visual analysis of large text corpora with topic modeling," *ACM Transactions on Intelligent Systems and Technology (TIST)*, vol. 3, no. 2, p. 23, 2012.
- [5] E. Alexander, J. Kohlmann, R. Valenza, M. Witmore, and M. Gleicher, "Serendip: Topic model-driven visual exploration of text corpora," in *Visual Analytics Science and Technology (VAST), 2014 IEEE Conference on*. IEEE, 2014, pp. 173–182.
- [6] E. Erosheva, S. Fienberg, and J. Lafferty, "Mixed-membership models of scientific publications," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5220–5227, 2004.
- [7] X. Wang, C. Zhai, and D. Roth, "Understanding evolution of research themes: a probabilistic generative model for citations," in *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2013, pp. 1115–1123.
- [8] D. Blei and J. Lafferty, "Correlated topic models," *Advances in neural information processing systems*, vol. 18, p. 147, 2006.
- [9] R. M. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM, 2008, pp. 542–550.
- [10] J. Chang and D. M. Blei, "Relational topic models for document networks," in *International conference on artificial intelligence and statistics*, 2009, pp. 81–88.
- [11] Q. He, B. Chen, J. Pei, B. Qiu, P. Mitra, and L. Giles, "Detecting topic evolution in scientific literature: how can citations help?" in *Proceedings of the 18th ACM conference on Information and knowledge management*. ACM, 2009, pp. 957–966.
- [12] R. Nallapati, D. A. McFarland, and C. D. Manning, "Topicflow model: Unsupervised learning of topic-specific influences of hyperlinked documents," in *AISTATS*, 2011, pp. 543–551.
- [13] L. Weng and T. M. Lento, "Topic-based clusters in egocentric networks on facebook," in *ICWSM*, 2014.
- [14] C. Wang, J. Liu, N. Desai, M. Danilevsky, and J. Han, "Constructing topical hierarchies in heterogeneous information networks," *Knowledge and Information Systems*, vol. 44, no. 3, pp. 529–558, 2015.
- [15] A. J.-B. Chaney and D. M. Blei, "Visualizing topic models," in *ICWSM*, 2012.
- [16] A. S. Maiya and R. M. Rolfe, "Topic similarity networks: visual analytics for large document sets," in *Big Data (Big Data), 2014 IEEE International Conference on*. IEEE, 2014, pp. 364–372.
- [17] P. Jahnichen, P. Oesterling, G. Heyer, T. Liebmann, G. Scheuermann, and C. Kuras, "Exploratory search through visual analysis of topic models," *Digital Humanities Quarterly (special issue)*, 2015.
- [18] T. Hofmann, "Probabilistic latent semantic indexing," in *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*. ACM, 1999, pp. 50–57.
- [19] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003.
- [20] L. Dietz, S. Bickel, and T. Scheffer, "Unsupervised prediction of citation influences," in *Proceedings of the 24th international conference on Machine learning*. ACM, 2007, pp. 233–240.
- [21] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *Proceedings of the 17th international conference on World Wide Web*. ACM, 2008, pp. 101–110.
- [22] R. Nallapati and W. W. Cohen, "Link-plsa-lda: A new unsupervised model for topics and influence of blogs," in *ICWSM*, 2008.
- [23] Y. Liu, A. Niculescu-Mizil, and W. Gryc, "Topic-link lda: joint models of topic and author community," in *proceedings of the 26th annual international conference on machine learning*. ACM, 2009, pp. 665–672.
- [24] T. Le and H. W. Lauw, "Probabilistic latent document network embedding," in *Data Mining (ICDM), 2014 IEEE International Conference on*. IEEE, 2014, pp. 270–279.
- [25] D. Cohn and H. Chang, "Learning to probabilistically identify authoritative documents," in *ICML*. Citeseer, 2000, pp. 167–174.
- [26] D. Cohn and T. Hofmann, "The missing link-a probabilistic model of document content and hypertext connectivity," *Advances in neural information processing systems*, pp. 430–436, 2001.
- [27] D. R. Radev, P. Muthukrishnan, and V. Qazvinian, "The acl anthology network corpus," in *Proceedings of the 2009 Workshop on Text and Citation Analysis for Scholarly Digital Libraries*. Association for Computational Linguistics, 2009, pp. 54–61.
- [28] D. Mimno, H. M. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 2011, pp. 262–272.
- [29] T. L. Griffiths and M. Steyvers, "Finding scientific topics," *Proceedings of the National Academy of Sciences*, vol. 101, no. suppl 1, pp. 5228–5235, 2004.
- [30] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Advances in neural information processing systems*, 2009, pp. 288–296.
- [31] D. Newman, J. H. Lau, K. Grieser, and T. Baldwin, "Automatic evaluation of topic coherence," in *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 2010, pp. 100–108.