LIMTopic: A Framework of Incorporating Link based Importance into Topic Modeling

Dongsheng Duan, Yuhua Li*, Ruixuan Li, *Member, IEEE,* Rui Zhang, Xiwu Gu, and Kunmei Wen

Abstract—Topic modeling has become a widely used tool for document management. However, there are few topic models distinguishing the importance of documents on different topics. In this paper, we propose a framework $LIMTopic$ to incorporate link based importance into topic modeling. To instantiate the framework, RankTopic and HITSTopic are proposed by incorporating topical pagerank and topical HITS into topic modeling respectively. Specifically, ranking methods are first used to compute the topical importance of documents. Then, a generalized relation is built between link importance and topic modeling. We empirically show that LIMTopic converges after a small number of iterations in most experimental settings. The necessity of incorporating link importance into topic modeling is justified based on KL-Divergences between topic distributions converted from topical link importance and those computed by basic topic models. To investigate the document network summarization performance of topic models, we propose a novel measure called log-likelihood of ranking-integrated document-word matrix. Extensive experimental results show that LIMTopic performs better than baseline models in generalization performance, document clustering and classification, topic interpretability and document network summarization performance. Moreover, RankTopic has comparable performance with relational topic model (RTM) and HITSTopic performs much better than baseline models in document clustering and classification.

Index Terms—Link Importance; Topic Modeling; Model Framework, Document Network; Log-likelihood of Ranking-integrated Document-word Matrix

1 INTRODUCTION

Due to its sound theoretical foundation and promising application performance, topic modeling has become a well known text mining method and is widely used in document navigation, clustering, classification [1] and information retrieval. Given a set of documents, the goal of topic modeling is to discover semantically coherent clusters of correlated words known as topics, which can be further used to represent and summarize the content of documents. The most well known topic models include PLSA (Probabilistic Latent Semantic Analysis) [2] and LDA (Latent Dirichlet Allocation) [3]. By using topic modeling, documents can be modeled as multinomial distributions over topics instead of those over words. Topics can serve as better features of documents than words because of its low dimension and good semantic interpretability.

With the widespread use of online systems, such as academic search engines [4], documents are often hyper-linked together to form a document network. A document network is formally defined as a collection of documents that are connected by links. In general, documents can have various kinds of textual contents, such as research papers, web pages and tweets. Documents can also be connected via a variety of links. For example, papers can be connected together via citations, web pages can be linked by hyper-links, and tweets can link to one another according to the retweet relationship.

To take advantage of the link structure of a document network, some link combined topic models, such as iTopic [5], have been proposed. However, most existing topic models do not explicitly distinguish the importance of documents on different topics, while in practical situations documents have different degrees of importance on different topics, thus treating them as equally important may inherently hurt the performance of topic modeling. To quantify the importance of documents on different topics, topical ranking methods [6] can be used, which is extensions of basic ranking algorithms, such as pagerank [7] and HITS (Hyperlink-Induced Topic Search) [8]. Although these ranking methods are initially proposed for the purpose of ranking web pages, it can be also used to rank other kind of documents, such as research publications cited by each other, since concepts and entities in those domains are similar [9]. In this work, we propose to incorporate link based importance into topic modeling.

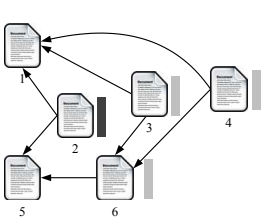
Specifically, topical ranking methods are employed to compute the importance scores of documents over topics, which are then leveraged to guide the topic modeling process. The proposed framework is called *Link Importance Based Topic Model*, denoted as *LIMTopic* for short. Compared to existing topic models, LIMTopic distinguishes the importance of documents while per-

- *D. Duan, Y. Li, R. Li, X. Gu and K. Wen are with School of Computer Science and Technology, Huazhong University of Science and Technology, China. D. Duan is also with National Computer network Emergency Response technical Team/Coordination Center of China.*
  E-mail: duandongsheng@gmail.com, {idcliyuhua, rxli, guxiwu, kmwen}@hust.edu.cn
- *R. Zhang is with Department of Computing and Information Systems, University of Melbourne, Australia*
  Email: rui.zhang@unimelb.edu.au

Corresponding Author

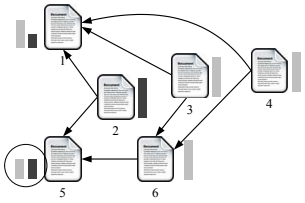
word \ doc	1	2	3	4	5	6
normalized	0	0	1	1	0	1
cuts	0	0	1	1	0	1
image	0	0	1	1	0	1
segmentation	0	0	1	1	0	1
graph	0	0	1	1	0	1
spectral	0	1	0	0	0	0
clustering	0	1	0	0	0	0
community	0	1	0	0	0	0
detection	0	1	0	0	0	0
network	0	1	0	0	0	0

(a) Documents

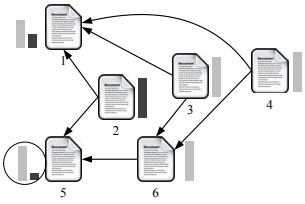


(b) Link structure

Fig. 1. An artificial document network. There are two topics in these documents, which are represented by gray and dark bars respectively. Documents are labeled by corresponding bars beside them.



(a) iTopic



(b) RankTopic

Fig. 2. The topic proportions of documents output by iTopic and RankTopic respectively. Higher bars indicate more proportions.

forming topic modeling. The philosophy behind the methodology is that the more important documents are given more weights than the less important ones. To instantiate the framework, RankTopic and HITSTopic are proposed based on document's link importance computed by topical pagerank and topical HITS (Hyperlink-Induced Topic Search) respectively.

As a motivating example, let's see a small artificial network with six documents as Figure 1 shows. The left side of the figure is the *word-document matrix*, and the right side is a fictional link structure among those imaginary documents. Traditional topic model (i.e. PLSA or LDA) discovers two topics, which are represented by gray and dark bars respectively. The two topics can be interpreted as "image segmentation" and "community detection" from corresponding words in them. The height of the bar beside each document indicate the document's topic proportion. Since documents 1 and 5 have no words, both of them are not labeled by any topics.

However, from the link structure, we have reason to believe that documents 1 and 5 should have been labeled by some topics because they are cited by documents with the two topics. As a link combined topic model, iTopic [5] can alleviate this issue to some degree. Figure 2(a) illustrates the topic detection result of iTopic, from which we can see that documents 1 and 5 are labeled by the two topics but with different proportions. Document 1 has more proportions on gray topic than on dark one while document 5 has the same proportion on them. Notice that document 1 is cited by two gray topics (documents 3 and 4) and one dark (document 2), while document 5 is cited by one gray (document 6) and one dark (document 2). iTopic treats neighboring documents as equally important such that the topic proportions of both documents 1 and 5 are computed as averages of topic proportions of their neighbors.

However, documents can have various importance on different topics, so treating them as equally important may obtain inaccurate topics. RankTopic incorporates the link based importance into topic modeling such that it can well distinguishes the importance of documents. Figure 2(b) shows the topic detection result of RankTopic, from which we can see that document 5 has much more proportions on gray topic than dark one. The underlying reason is that document 6 ranks high on gray topic as it is cited by two gray topics, while document 2 ranks low because it is not cited by any documents. There are more evidence showing that document 5 is more likely about gray topic than dark one. In the aspect of capturing such evidence, RankTopic performs reasonably better than iTopic and other network regularization based topic models, such as NetPLSA [10], which motivates our study on RankTopic.

In the above example, we clearly see that RankTopic can well incorporate the importance of documents into topic modeling and addresses the drawbacks of some existing topic models. The necessity of incorporating link based importance into topic modeling is empirically justify based on the KL-Divergence between topic distributions converted from topical ranking and those computed by basic topic model in the experimental section. We also experimentally demonstrate that RankTopic and HITSTopic can perform better than some baseline models in generalization performance, document clustering and classification performance by setting a proper balancing parameter. In addition, we find that RankTopic has comparable performance with one of the state-of-the-art link based relational topic model (RTM) in the above measures and HITSTopic performs much better than all the compared models in terms of document clustering and classification performance. Moreover, we find topics detected by LIMTopic are more interpretable than those detected by some baseline models and still comparable with RTM, and LIMTopic fits the whole document best among all the compared topic models in terms of the log-likelihood of ranking-integrated document-word matrix.

To summarize, compared with existing topic models LIMTopic has the following distinguished characteristics.

- Existing topic models assume that documents plays equally important role in topic modeling. In contrast, LIMTopic incorporates the importance of documents into topic modeling and benefit from such combination.
- Previous works treat topic modeling and link based importance computing as two independent issues while LIMTopic puts them together and makes them mutually enhanced in a unified framework.
- LIMTopic is flexible since ranking and topic modeling are orthogonal to each other such that different

ranking and topic modeling methods can be used according to specific application requirements.
- LIMTopic outperforms the state-of-the-art topic models in summarizing the whole document network in terms of a novel measure called the log-likelihood L_{rank} of ranking-integrated document-word matrix.

The rest of the paper is organized as follows. Section 2 reviews the related works. Section 3 presents the preliminaries about topic modeling and ranking. We propose LIMTopic framework and present parameter learning algorithm for the proposed model in Section 4. Experimental settings and results are demonstrated in Section 5 and we conclude this paper in Section 6.

2 Related Work

2.1 Topic Models

Topic models have been widely studied in the text mining community due to its solid theoretical foundation and promising application performance. PLSA [2] and LDA [3] are two well known basic topic models. Since they are proposed, various kinds of extensions have been proposed by incorporating more contextual information, such as time [11], [12], [13], [14], authorship [15], and links [5], [16], [10], [17], [18], [19]. Another kind of extension is to extract sharing common topics across multiply text streams [20]. Some others combine other generative models with topic model. For example, literature [21] combine community and topic into a unified model. Wang et al. [22] combine collaborative filtering and LDA for recommending scientific publications. In [23], sentiment and topic is combined in an fully generative model to detect both of them simultaneously from text. The present work also incorporates links into topic modeling but uses different way from previous works. Although most earlier link combined topic models can capture the topical correlations between linked documents, there are few works leveraging the topical ranking of documents to guide the topic modeling process. The most similar work to ours may be the TopicFlow model [24]. The distinguished features of present work from TopicFlow lie in the following folds. First, LIMTopic provides a more flexible combination between link importance and topic modeling while TopicFlow couples flow network and topic modeling tightly. This feature makes LIMTopic more extendable. Second, LIMTopic builds a generalized relation between link importance and topic modeling rather than a hard relation like TopicFlow. Third, the topic specific influence of documents computed by TopicFlow can actually serve as the topical ranking in LIMTopic.

2.2 Ranking

Our work is also tightly related to ranking technology. The most well known link based ranking algorithms are PageRank [7] and HITS [8]. Both algorithms are based on the phenomenon that rich gets richer. Considering that the ranking of documents are dependent on their contents, topic sensitive pagerank [25] is proposed by biasing the documents on a particular topic. Topical link analysis [6] extends the algorithms by calculating a vector of scores to distinguish the importance of documents on different topics. [26] proposes random walk with topic nodes and random walk at topical level to further rank documents over heterogenous network. [27] takes the dynamic feature of citation network into account and proposes FurtureRank to compute the expected future citations of papers and to rank their potential prestige accordingly. [28] proposes P-Rank to rank the prestige of papers, authors and journals in a heterogenous scholarly network. RankClus [29], [30] further extends the method to heterogenous information networks to rank one kind of node with respect to another. Compared to RankClus which performs ranking based on hard clustering, we incorporate link based importance into topic modeling which is a soft clustering. Another difference is that RankClus is a clustering algorithm based on only links while LIMTopic is a topic modeling framework based on both links and texts.

2.3 Community Detection

Link based community detection is also relevant to our study. Community detection is a fundamental link analysis problem that has been extensively studied [31]. Traditional community detection algorithms partition the network into groups of nodes such that links among group members are much denser than those cross different groups. The state-of-the-art community detection algorithms include spectral clustering [32] and modularity optimization approach [33]. However, such algorithms do not consider text information associated with each node to help clustering. We would like to mention PCL-DC [34], which is a community detection algorithm by combining links and textual contents. The node popularity introduced in PCL-DC can also be regarded as link based importance. However, PCL-DC introduces the popularity variable in the link based community detection model (PCL) but does not directly use it in the discriminative content (DC) model, while LIMTopic explicitly incorporates link importance into the generative model for textual contents.

3 Preliminaries

3.1 Topic Modeling

Topic modeling aims at extracting conceptually coherent topics shared by a set of documents. In the following, we describe topic model PLSA [2] upon which LIMTopic is built. We choose the most basic topic model PLSA rather than LDA, because the prior in LDA is noninformative while LIMTopic can be regarded as PLSA with informative prior.

Given a collection of N documents D, let V denote the total number of unique words in the vocabulary and K

represent the number of topics, the goal of PLSA is to maximize the likelihood of the collection of documents with respect to model parameters Θ and B.

$$P(D|\Theta, B) = \prod_{i=1}^{N}\prod_{w=1}^{V}\left(\sum_{z=1}^{K}\theta_{iz}\beta_{zw}\right)^{s_{iw}} \quad (1)$$

where $\Theta = \{\theta\}_{N\times K}$ is the topic distribution of documents, $B = \{\beta\}_{K\times V}$ is the word distribution of topics, and s_{iw} represents the times that word w occurs in document i.

After the inference of PLSA, each topic is represented as a distribution over words in which top probability words form a semantically coherent concept, and each document can be represented as a distribution over the discovered topics.

3.2 Topical Link Importance

Link importance is the documents' global importance computed based only on the link structure of the document network. However, ranking documents by a single global importance score may not make much sense because documents should be ranked sensitive to their contents. Based on this consideration, topical link analysis [6], i.e. topical pagerank and topical HITS, are proposed. In the following, we take topical pagerank as an example to briefly introduce topical link analysis.

As the input of topical pagerank, each document i is associated with a topic distribution θ_i, which can be obtained via topic modeling methods. Taking the topic distribution of documents into account, topical pagerank produces an importance vector for each document, in which each element represents the importance score of the document on each topic. Letting γ_{zi} denote the importance of document i on topic z, topical pagerank is formally expressed as

$$\gamma_{zi}^{(t)} = \lambda\sum_{j\in I_i}\frac{\alpha\gamma_{zj}^{(t-1)} + (1-\alpha)\theta_{jz}\gamma_{\cdot j}^{(t-1)}}{|O_j|} + (1-\lambda)\frac{\theta_{iz}}{M} \quad (2)$$

where α and λ are parameters that control the process of prorogating the ranking score, which are both empirically set to 0.85. $\gamma_{\cdot j} = \sum_{z=1}^{K}\gamma_{zj}$ denotes the global importance of document j, I_i is the set of in-link neighbors of document i, $|O_j|$ denotes the number of out-link neighbors of document j, and θ_{jz} is the topic proportion of document j on topic z and M is the total number of documents.

The process of topical link analysis is illustrated in Figure 3 excluding the thick line. It can be seen that topical link analysis first performs topic modeling to obtain documents' topic distribution and then performs topical link analysis to obtain the topicial link importance of documents, thus it regards link analysis and topic modeling separately. It is worthy to point out that the original topical link analysis [6] method uses supervised learning method based on predefined categories from Open Directory Project (ODP) other than topic modeling methods to obtain the topic distribution of documents.

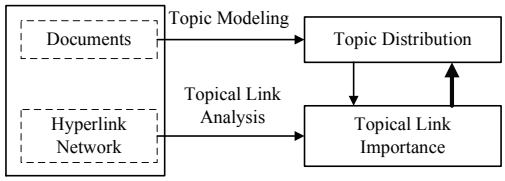


Fig. 3. Mutual enhancement framework between topic distribution and link based importance

4 Link Importance based Topic Modeling Framework

4.1 Relation between link importance and Topic Modeling

To incorporate the link importance into topic modeling as the thick line in Figure 3 shows, it is essential to build the relation between them. However, there is no closed solution for establishing this relation. Here, we present a natural way to achieve this end.

Notice that the link importance γ_{zi} can be interpreted as the probability $P(i|z)$ of the node i involved in the topic z by normalizing the importance vector such that $\sum_{i=1}^{M}P(i|z) = 1, \forall z$. By using the sum and product rules of the Bayesian theorem, the topic proportion $P(z|i)$ can be expressed in terms of γ_{zi}.

$$\theta_{iz} = P(z|i) = \frac{P(i|z)p(z)}{\sum_{i'=1}^{M}P(i|z)p(z)} = \frac{\gamma_{zi}\pi_z}{\sum_{z'=1}^{K}\gamma_{z'i}\pi_{z'}} \quad (3)$$

where $\pi_z = P(z)$ is the prior probability of topic z.

By using the above interpretation, the topic proportion of a document is decomposed into the multiplication of topical link importance and the prior distribution of topics. However, there is still a problem for the above equation. Topical link importance is computed based on the link structure of the document network, which inevitably has noise in practical situations. We observe some self-references in the ACM digital library, which is usually caused by some error editing behavior. Inappropriate and incomplete references may also exist. Therefore, equating between the topical link importance γ_{zi} and the conditional probability $P(i|z)$ also bring much noise into the topic modeling. One possible solution for this problem is to detect the noise links and remove them from the document network. However, spam detection itself is a challenging issue, which is out of the scope of this paper.

To reduce the effects of noise, we model the degree of our belief on the link importance instead of removing the noise links. Specifically, we transform Equation 3 to a more generalized one by introducing a parameter ξ ranging from 0 to 1 to indicate our belief on the link importance as follows.

$$\theta_{iz} = P(z|i) \propto [\xi\gamma_{zi} + (1-\xi)\phi_{zi}]\pi_z \quad (4)$$

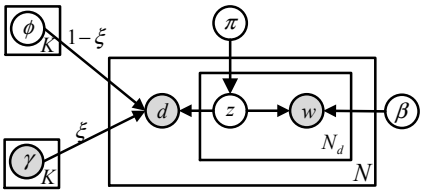


Fig. 4. Link importance based topic modeling framework

where $\phi_{zi} = p(i|z)$ has the same interpretation as γ_{zi}, but it is a hidden variable rather than an observed one.

In Equation 4, if $\xi = 0$, the topic proportions are the same as that in PLSA, and if $\xi = 1$, the topic proportions are completely dependent on the topical ranking. Intermediate values of ξ balance between the above two extreme cases. The larger the value of ξ, the more information of link importance is incorporated into the topic modeling. Therefore, Equation 3 is actually a special case of Equation 4 by setting ξ to 1.

4.2 LIMTopic Framework

Based on the generalized relation between link importance and topic proportion, we can replace θ in PLSA with the right side of Equation 4, which results in the link importance based topic modeling framework *LIMTopic*. Figure 4 shows the graphical representation of the LIMTopic framework. Different from the traditional topic models, the probability $p(i|z)$ of a document i involved in a topic z is governed by the weighted mixture of topical ranking γ_{zi}, and the hidden variable ϕ_{zi} in the LIMTopic model such that the effects of link importance on topic modeling is integrated.

To instantiate LIMTopic framework, we incorporate topical link importance computed by topical pagerank and topic HITS into topic modeling, results in RankTopic and HITSTopic respectively. For topical pagerank, the resulting ranking vector is simply taken as the topical link importance. However, topical HITS computes two ranking vectors for each document rather than one vector like topical pagerank. Since authority or hubness value of a document only reflects one kind of importance of the document in the network, we would like to combine them together to represent the overall importance of the document. Specifically, we first compute the sums of authority and hubness vectors, then normalize them in the dimension of each topic and regard this result as the topical link importance of documents.

In LIMPTopic, the topical link importance γ of documents is labeled as observational variable (shaded in Figure 4) since it can be obtained by the topical pagerank or topical HITS algorithm, although in an overall view topical link importance is in fact unknown. By incorporating topical link importance γ_{zi} into the topic modeling, the link information is naturally taken into account since the topical link analysis process is performed on the link structure.

In LIMTopic Framework, the likelihood of a collection of documents D with respect to the model parameters is

$$P(D|\gamma,\pi,\phi,\beta) = \prod_{i=1}^{M}\prod_{w=1}^{V}\left(\sum_{z=1}^{K}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z\beta_{zw}\right)^{s_{iw}} \quad (5)$$

where the definition of all the notations can be found in the previous parts of this paper. Next, the maximum likelihood estimation is adopted to derive the model parameters involved in LIMTopic.

4.3 Derivation of LIMTopic

To obtain the (local) maximum of the likelihood in Equation 5, the expectation maximization (EM) algorithm is employed. Detailed derivation of the EM updating rules is as follows.

The logarithm of the likelihood function is

$$\begin{aligned} L &= \log P(D|\gamma,\pi,\phi,\beta) \\ &= \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\log\sum_{z=1}^{K}\beta_{zw}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z \end{aligned} \quad (6)$$

In the E-step, the posterior distribution $P(z|i,w)$ of topics conditioned on each document-word pair (i,w) is computed by Equation 7.

$$\psi_{iwz}^{(t)} = P^{(t)}(z|i,w) \propto \beta_{zw}^{(t)}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}^{(t)}\right]\pi_z^{(t)} \quad (7)$$

Then, the lower bound of L can be derived by using Jensen inequality twice as following,

$$\begin{aligned} L &= \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\log\sum_{z=1}^{K}\psi_{iwz}^{(t)}\frac{\beta_{zw}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z}{\psi_{iwz}^{(t)}} \\ &\geq \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\psi_{iwz}^{(t)}\log\beta_{zw}\left[\xi\gamma_{zi} + (1-\xi)\phi_{zi}\right]\pi_z \\ &\quad - \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\psi_{iwz}^{(t)}\log\psi_{iwz}^{(t)} \\ &\geq \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}[\xi\psi_{iwz}^{(t)}\log\beta_{zw}\gamma_{zi}\pi_z \\ &\quad + (1-\xi)\psi_{iwz}^{(t)}\log\beta_{zw}\phi_{zi}\pi_z] - \sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\psi_{iwz}^{(t)}\log\psi_{iwz}^{(t)} \end{aligned}$$

In the M-step, the lower bound of L is maximized under the constraints $\sum_{w=1}^{V}\beta_{zw} = 1$, $\sum_{z=1}^{K}\pi_z = 1$ and $\sum_{i=1}^{M}\phi_{zi} = 1$. Through introducing Lagrange multipliers, the constrained maximization problem is converted to the following one.

$$\begin{aligned} &\max_{\theta,\pi}\sum_{i=1}^{M}\sum_{w=1}^{V} s_{iw}\sum_{z=1}^{K}\left[\xi\psi_{iwz}^{(t)}\log\beta_{zw}\gamma_{zi}\pi_z + (1-\xi)\psi_{iwz}^{(t)}\log\beta_{zw}\phi_{zi}\pi_z\right] \\ &+\sum_{z=1}^{K}\lambda_z\left(\sum_{w=1}^{V}\beta_{zw} - 1\right) + \lambda\left(\sum_{z=1}^{K}\pi_z - 1\right) + \sum_{z=1}^{K}\lambda'_z\left(\sum_{i=1}^{M}\phi_{zi} - 1\right) \end{aligned}$$

The above maximization problem has a closed form solution as follows, which gives out the update rules that monotonically increase L.

$$\beta_{zw}^{(t+1)} \propto \sum\nolimits_{i=1}^{M} s_{iw}\psi_{iwz}^{(t)} \quad (8)$$

$$\pi_z^{(t+1)} \propto \sum\nolimits_{i=1}^{M}\sum\nolimits_{w=1}^{V} s_{iw}\psi_{iwz}^{(t)} \quad (9)$$

$$\phi_{zi}^{(t+1)} \propto \sum\nolimits_{w=1}^{V} s_{iw}\psi_{iwz}^{(t)} \tag{10}$$

As the parameter updating process converges, the topic proportion θ can be computed by using Equation 4.

4.4 The Learning Algorithm of LIMTopic

With LIMTopic, we can build a mutual enhancement framework by organizing topic modeling and link importance into an alternative process illustrated in Figure 3. By introducing LIMTopic as the thick line shows, the sequential framework from topic modeling to link importance is transformed to a mutual enhancement framework.

From the implementation view, we provide the matrix form of the parameter estimation equations. The parameters involved in the overall framework include topic-word distributions $\mathrm{B} = \{\beta\}_{K\times V}$, hidden variable $\Phi = \{\phi\}_{K\times N}$, topic prior distributions $\Pi = \{\pi\}_K$, and topical ranking $\Gamma = \{\gamma\}_{K\times N}$. Let $\mathrm{S} = \{s\}_{N\times V}$ denote the document-word matrix in which s_{iw} represents the time word w occurs in document i. Let $\mathrm{L} = \{l\}_{N\times N}$ denote the link structure among those documents in which $l_{ij} = 1$ represents that there is a link from document i to document j and $l_{ij} = 0$ represents there is not.

It can be proved that Equation 8, 9, 10 and 2 have the following four matrix forms respectively.

$$\mathrm{B} = \mathrm{B}. * \left(\mathrm{Y}(\mathrm{S}./(\mathrm{Y}^{\mathrm{T}}\mathrm{B}))\right) \tag{11}$$

where $\mathrm{Y} = (\xi\Gamma + (1-\xi)\Phi). * \left[\begin{array}{ccc}\Pi & \cdots & \Pi\end{array}\right]$, and $.*$ and $./$ represent element wise multiplication and division operation between two matrices respectively.

$$\Pi = \mathrm{diag}\{\mathrm{Y}(\mathrm{S}./(\mathrm{Y}^{\mathrm{T}}\mathrm{B}))\mathrm{B}^{\mathrm{T}}\} \tag{12}$$

where $\mathrm{diag}\{\cdot\}$ returns the main diagonal of a matrix.

$$\Phi = \mathrm{Y}. * \left(\mathrm{B}\left(\mathrm{S}./(\mathrm{Y}^{\mathrm{T}}\mathrm{B})\right)^{\mathrm{T}}\right) \tag{13}$$

$$\Gamma = \lambda\left(\alpha\Gamma + (1-\alpha)\mathrm{X}\right)\widehat{\mathrm{L}} + \frac{1-\lambda}{M}\Theta^{\mathrm{T}} \tag{14}$$

where $\mathrm{X} = \Theta^{\mathrm{T}}. * \left[\begin{array}{ccc}\mathrm{sum}(\Gamma) & \cdots & \mathrm{sum}(\Gamma)\end{array}\right]^{\mathrm{T}}$, $\mathrm{sum}(\cdot)$ returns sums along the columns of a matrix, and $\widehat{\mathrm{L}}$ is the row normalization matrix of link structure L.

According to the mutual enhancement framework and matrix forms of the updating rules presented above, the learning algorithm of LIMTopic is summarized in Algorithm 1. In the following, we present the three termination conditions in the algorithm.

Condition 1: This condition is to test whether the topical ranking Γ converges. We compute the differences between the topical ranking of the current iteration and the previous one, and sum these differences over all the cells. If the difference is lower than a predefined small value (1e-2 in our experiments), this condition is satisfied.

Condition 2: This condition is to test whether the ranking based topic modeling process converges. For each iteration, we compute the log-likelihood of the observed documents with respect to the current parameters B, Γ, Φ and Π via Equation 6, and then compute the relative change of the log-likelihood between two continuous iterations as the fraction of the difference between the two log-likelihoods to the average value of them. If the relative change is lower than a predefined small value (1e-4 in our experiments), this condition is satisfied.

Algorithm 1: The learning algorithm of LIMTopic

Input: A document network L with M documents including totally V unique words, and the expected number K of topics and parameter ξ.
Output: Topic-word distributions B, Document-topic distributions Θ.
initialization: Perform PLSA to obtain B and Θ;
repeat
  repeat
    $\Gamma = \lambda\left(\alpha\Gamma + (1-\alpha)\mathrm{X}\right)\widehat{\mathrm{L}} + \frac{1-\lambda}{M}\Theta^{\mathrm{T}}$;
    Normalize Γ such that $\forall z, i, \sum_{z=1}^{K}\sum_{i=1}^{N}\gamma_{zi} = 1$;
  until *Satisfying condition 1*;
  repeat
    $\mathrm{B} = \mathrm{B}. * \left(\mathrm{Y}(\mathrm{S}./(\mathrm{Y}^{\mathrm{T}}\mathrm{B}))\right)$;
    Normalize B such that $\forall z, \sum_{w=1}^{V}\beta_{zw} = 1$;
    $\Pi = \mathrm{diag}\{\mathrm{Y}(\mathrm{S}./(\mathrm{Y}^{\mathrm{T}}\mathrm{B}))\mathrm{B}^{\mathrm{T}}\}$;
    Normalize Π such that $\sum_{z=1}^{K}\pi_z = 1$;
    $\Phi = \mathrm{Y}. * \left(\mathrm{B}\left(\mathrm{S}./(\mathrm{Y}^{\mathrm{T}}\mathrm{B})\right)^{\mathrm{T}}\right)$;
    Normalize Φ such that $\forall z, \sum_{i=1}^{N}\phi_{zi} = 1$;
  until *Satisfying condition 2*;
  $\Theta = (\xi\Gamma + (1-\xi)\Phi)^{\mathrm{T}}. * \left[\begin{array}{ccc}\Pi & \cdots & \Pi\end{array}\right]^{\mathrm{T}}$;
  Normalize Θ such that $\forall i, \sum_{z=1}^{K}\theta_{iz} = 1$;
until *Satisfying condition 3*;
return B Θ;

Condition 3: This condition is to test whether the whole process reaches a (local) optimal solution. For each iteration, we propose to compute a novel measure called log-likelihood of the ranking-integrated document-word matrix with respect to the current parameters B and Θ.

The ranking-integrated document-word matrix is computed by using topical pagerank on the link structure and original document-word matrix. Specifically, the ranking-integrated document-word matrix R is computed by iteratively performing Equation 15.

$$R_{wi}^{(t)} = \lambda\sum_{j\in I_i}\frac{\alpha R_{wj}^{(t-1)} + (1-\alpha)\widehat{S}_{jw}R_{.j}^{(t-1)}}{|O_j|} + (1-\lambda)\frac{\widehat{S}_{iw}}{M} \tag{15}$$

where $\widehat{S}$ is the row normalized matrix of original document-word matrix S. Equation 15 is essentially the same as Equation 2 by replacing γ and θ with R and $\widehat{S}$ respectively. The ranking-integrated document-word matrix is actually an imaginary document-word matrix which encodes the observational information from both documents and links.

The log-likelihood L_{rank} of the ranking-integrated document-word matrix conditioned on the parameters B and Θ is computed as Equation 16.

$$L_{rank} = P(R|\mathrm{B},\Theta) = \mathrm{sum}(\mathrm{sum}(\ln(\mathrm{B}*\Theta). * \mathrm{R})) \tag{16}$$

where $\mathrm{sum}(\mathrm{sum}(\cdot))$ returns the sum of all the elements in a matrix. Higher value of L_{rank} indicates better fit to the ranking-integrated document-word matrix. Since LIMTopic takes the link importance into account, the log-likelihood of the ranking-integrated document-word matrix can be a better evaluation measure than that of the original document-word matrix. If the incremental quantity of the log-likelihood is lower than a predefined threshold (1e-3 in our experiments), condition 3 is satisfied.

The time complexity of Algorithm 1 is analyzed as follows. The complexity of computing matrix X is $O(K \times N)$, and that of computing $\alpha\Gamma+(1-\alpha)\mathrm{X}$ is also $O(K \times N)$. The most time consuming part of topical ranking (i.e. the first loop) is to compute the product between a $K \times N$ matrix (i.e. $\alpha\Gamma+(1-\alpha)\mathrm{X}$) and a $N \times N$ matrix $\widehat{L}$ with complexity $O(K \times N \times N)$, which can be reduced to $O(K \times E)$ in sparse networks where E is the number of links in the network. Let T_1 be the maximum number of iterations, then the complexity of topical ranking is $O(T_1 \times K \times E)$. For the second loop, the most time consuming part is to compute the product between a $N \times K$ matrix and $K \times V$ matrix, whose time complexity is $O(N \times K \times V)$. Assuming the second loop need T_2 iterations, then its complexity is $O(T_2 \times N \times K \times V)$. Finally, we assume the third loop need T_3 iterations, then the complexity of the overall process is $T_3 \times (T_1 \times K \times E + T_2 \times N \times K \times V)$. If we further assume that the iteration numbers T_1, T_2, T_3 and the number of topics K are all constants, then the time complexity of the overall process turns out to be $O(E + N \times V)$, which is linear in the total number of links and words in the observed document network.

5 Experiments

In this section, we conduct experimental studies of LIMTopic based models, RankTopic and HITSTopic, in various aspects, and compare it with some state-of-the-art topic models, namely PLSA, LDA, iTopic and RTM (Relational Topic Model) [18]. In the experiments, we use two genres of data sets, i.e. three public paper citation data sets and one twitter data set.

ArnetMiner: This is a subset of the Citation-network V1 (http://www.arnetminer.org/citation) released by ArnetMiner [4]. After some preprocessing, there are 6,562 papers and 8,331 citations left and 8,815 unique words in the vocabulary.

Citeseer: This data set consists of 3,312 scientific publications and 4,715 links. The dictionary consists of 3,703 unique words. These publications have been categorized into 6 classes according to their research directions in advance.

Cora: There are 2,708 papers, and 5,429 citations in this subset of publications. The dictionary consists of 1433 unique words. These publications have been labeled as one of 7 categories in advance.

Twitter: The twitter data we used is released by [35], which can be downloaded from http://arnetminer.org/heterinf. In this data set, users associated with their published tweets are regarded as documents and the '@' relationship among users as links. After some preprocessing like stop word removing, we obtain 814 users in total and 5,316 unique words in the vocabulary. There are 4,206 '@' relationships between those users.

Both Citeseer and Cora data sets used in our experiments is the same as that used in [34].

5.1 Convergence Discussion

Although both ranking and topic modeling converge, it is hard to prove whether LIMTopic with nonzero ξ converges or not in theory. The alternative iteration process is complex in general due to the orthogonal relation between ranking and topic modeling, and different ranking algorithms or different values of ξ may have varying convergence performance. The above characteristics lead to the difficulty of strict convergence analysis for the whole process, which is left as an open problem for future research.

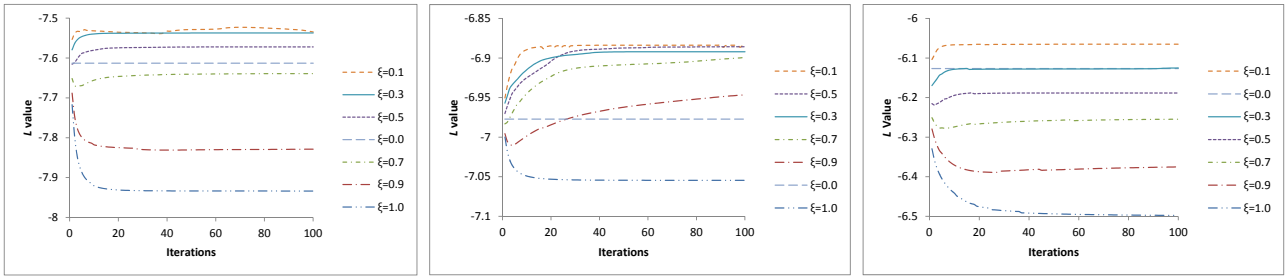
Instead, we explore the convergence of LIMTopic by plotting the L (i.e. Equation 6) curves with the iterations. Figure 5 shows some of the experimental results, and similar results are obtained from the left experiments. From the results, we can see that L curves do not always monotonically increase or decrease with the iterations. However, the L curves converge to fixed values under most ξ settings, indirectly reflecting that model parameters keep almost the same after enough iterations. From the results, we can also see that most L curves become flat after only a small number of iterations, i.e. less than 20 iterations.

5.2 Empirical Justification

In this subsection, we would like to empirically justify the necessity of integrating the link importance into topic modeling.

The difference between the results of link importance and traditional topic modeling is evaluated to justify the necessity of combining both of them. By normalizing the topical ranking vectors of each document to one, topical link importance can be converted to topic distributions of the documents. Formally, normalizing topical link importance Γ by column and transposing it, we get converted topic distributions $\widetilde{\Theta}$ of documents.

For each data set, we compute the KL-divergence between the topic distribution $\widetilde{\theta}_i$ converted from topical link importance and the original one θ_i for each document i. Some statistics of the results are shown in Table 1. From the statistics, we can see that there are significant differences between the converted and original topic distributions in all the four data sets. The divergences reach maximum value 13.33 and on average 1.02 in ArnetMiner data. The second largest average divergence occurs in Twitter data where the ratio of documents with non-zero divergence achieves 100%.



(a) RankTopic on ArnetMiner (b) RankTopic on Citeseer (c) HITSTopic on Cora

Fig. 5. L curves of LIMTopic with the iterations

TABLE 1
Statistics of the KL-Divergences between the converted topic distribution and the original topic distribution for all the documents.

KL-Divergence	SUM	AVG	MAX	NZRate[1]
ArnetMiner	6721.19	1.02	13.33	0.58
Citeseer	962.57	0.29	7.11	0.57
Cora	1666.80	0.62	12.95	0.58
Twitter	744.80	0.91	12.04	1.00

[1] NZRate represents the fraction of documents having non-zero KL-Divergence between their converted and original topic distributions.

The converted topic distribution is computed from the link structure while the original one is computed from the textual content of documents, i.e. they reflect different aspects of the document network. The significant inconsistency or divergence between them empirically indicates it is necessary to combine them together for better exploring the whole document network.

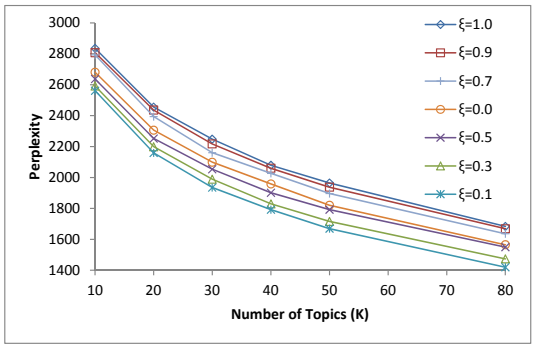
In the following subsections, we further investigate the practical performance of LIMTopic based models by using some well recognized measures and applications.

5.3 Generalization Performance

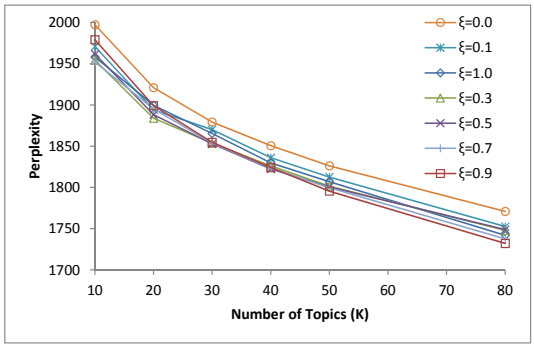
Perplexity [36] is a widely used measure for evaluating the generalization performance of a probabilistic model. Lower perplexity indicates better generalization performance.

In our experiments, we perform 10-fold cross validation. Before comparing RankTopic and HITSTopic with other topic models, we first study how the value of parameter ξ affects the generalization performance of RankTopic. Figure 6 shows parameter study results for some typical values of ξ on ArnetMiner and Twitter data. From the results, we observe the following phenomenons.

Both the results on ArnetMiner and Twitter data sets consistently show that RankTopic could obtain lower perplexity than the special case when ξ equals 0.0, which actually degenerates to PLSA but with additional termination condition for outside loop (see condition 3 in section 4.4). These results show that link based ranking can indeed be used to improve the generalization performance of basic topic models. However, we also observe



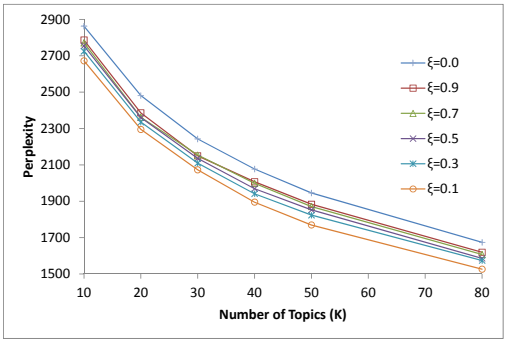
(a) Parameter Study on ArnetMiner for RankTopic



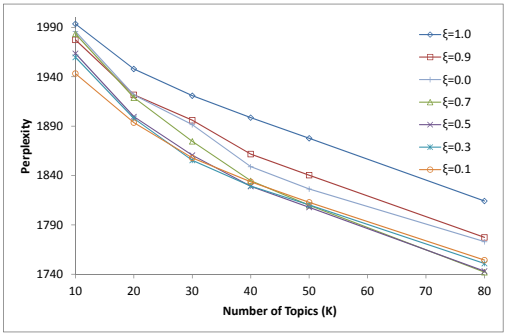
(b) Parameter Study on Twitter for RankTopic

Fig. 6. Perplexity results of RankTopic on ArnetMiner and Twitter data sets by setting some typical values of parameter ξ and K (number of topics) in RankTopic. All these results are average values computed under 10-fold cross validation.

different effects of ξ on RankTopic's generalization performance for different data sets. For ArnetMiner data, the lower the value of ξ, the better RankTopic's generalization performance except for $\xi = 0.0$. For Twitter data, the best generalization performance is obtained when $\xi = 0.9$ and perplexity is less sensitive to ξ except for the special case of $\xi = 0.0$. Whether RankTopic is sensitive to ξ may significantly depend on the consistency between links and texts and the noises in them. Nevertheless, we provide a tuning way for adapting RankTopic into practical senecios. For HITSTopic, we conduct the same experiments, and show the results in Figure 7. Due to the large difference between perplexity of $\xi = 1.0$ and those of other ξ values, the results of $\xi = 0.1$ is now shown

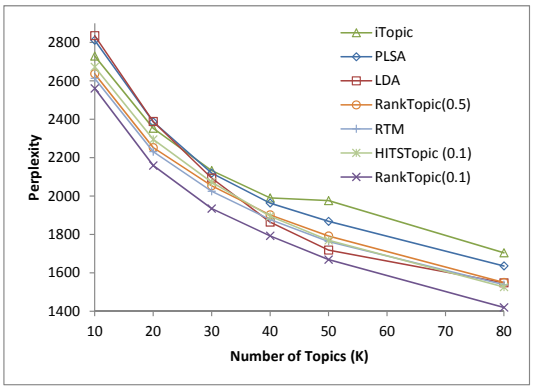


(a) Parameter Study on ArnetMiner for HITSTopic

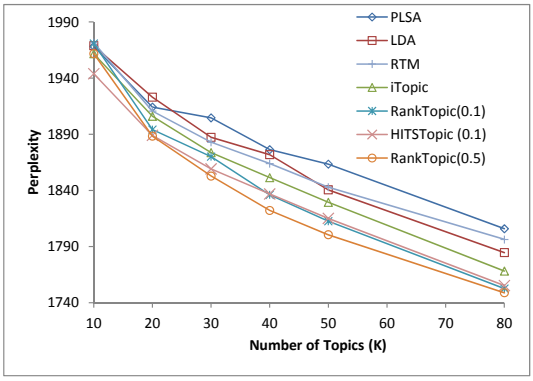


(b) Parameter Study on Twitter for HITSTopic

Fig. 7. Perplexity results of HITSTopic on ArnetMiner and Twitter data sets by setting some typical values of parameter ξ and K (number of topics) in HITSTopic. All these results are average values computed under 10-fold cross validation.



(a) Perplexity Comparison on ArnetMiner



(b) Perplexity Comparison on Twitter

Fig. 8. Perplexity results of RankTopic and some baseline topic models on ArnetMiner and Twitter data sets by setting various numbers of topics (K). All these results are averages computed under 10-fold cross validation. For RankTopic, the results of $\xi = 0.1$ and $\xi = 0.5$ are shown for comparison purpose. For HITSTopic, the result of $\xi = 0.1$ is shown.

for ArnetMiner data in this figure. From this figure, we can also see that HITSTopic has better generalization performance than the case when $\xi = 0.0$.

Figure 8 illustrates the perplexity results of the compared topic models. Results show that RankTopic with appropriately set ξ performs best among all the compared models, which indicates its superior generalization performance over the baseline topic models. The underlying reasons for the results are analyzed as follows. By introducing Dirichlet prior, LDA performs better than PLSA when K value increases. However, the prior adopted by LDA is non-informative. RankTopic can also be regarded as incorporating prior into PLSA, but topical ranking is more informative than Dirichlet prior. Both RTM and iTopic incorporate link structure into topic modeling. However, iTopic assumes that the neighbors of a node play equally important role in affecting the topics of that node, which is usually not the truth in practical document networks. The topics detected by RTM are governed by both link regression process and the document contents, but RTM does not model the weights of the two parts such that its generalization performance depends on the accuracy of links and contents. In contrast, RankTopic provides a turning weight of the incorporation of ranking such that it is more flexible than RTM. Notice that the generalization performance of HITSTopic is worse than that of RankTopic. The underline reason may be that topical pagerank may better serve as the conditional probability of a document on a topic than the sum of topical authority and topical hubness.

5.4 Document Clustering

Besides the generalization performance, topic models can also be evaluated by using their application performance. The most widely used applications of topic models include document clustering and classification. In this and subsequent subsection, we study the performance of LIMTopic based models on document clustering and classification respectively.

By using topic models, documents can be represented as topic proportion vectors, upon which document clustering can be performed. Specifically, we adopt k-means as the clustering algorithm. For a network, normalized cut (Ncut) [37], modularity (Modu) [38], are two well known measures for evaluating the clustering results. Lower normalized cut and higher modularity indicates better clustering result. When the background label information is known for documents, normalized mutual information (NMI) [39] can also be used to evaluate the

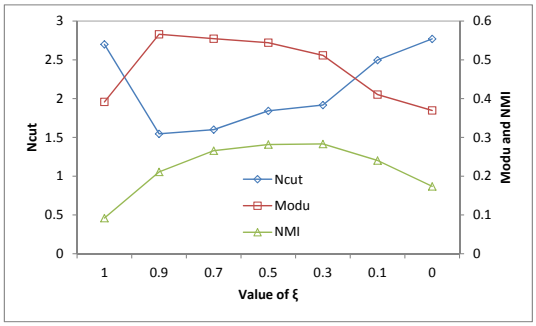


Fig. 9. Clustering performance of RankTopic with some typical values of ξ on Citeseer data. For Ncut, the lower the better. For both Modu and NMI, the higher the better. Notice that the left Y-axis is just for Ncut, while the right one is for both Modu and NMI.

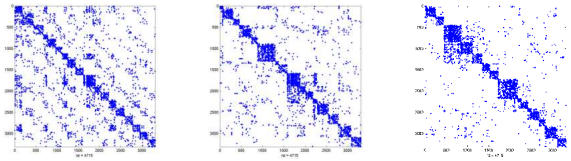
clustering result. The higher the NMI, the better the clustering quality. In these experiments, the number of clusters and topics are set to 6 for the Citeseer data, 7 for the Cora data, and 10 for twitter data. Both the numbers of clusters in Citeseer and Cora data are specified to be the known numbers of categories in them (see the data set part at the beginning of the experimental section), while that of the twitter data is set empirically. Although some parameter-free methods can be used to alleviate the cluster number setting issue, it is still a challenge to set an accurate number of clusters for an unexplored data set. In the experiments of Arnetminer data set, we find that both the normalized cut and modularity of all the compared models are not so significant, reflecting that Arnetminer data has no significant community structure in terms of citation network. The underlying reason is that we extract a subset of Arnetminer data in our experiments and this subset has no significantly dense parts in terms of link structure.

We first study the effect of parameter ξ. Figure 9 shows the results. For both Ncut and Modu, RankTopic with $\xi = 0.9$ performs best on Citeseer data. For NMI, RankTopic with $\xi = 0.3$ performs best on Citeseer data. Overall, RankTopic with $\xi = 0.5$ compromises among the three evaluation measures. We obtain similar results on Cora and Twitter data.

We then compare the clustering performance of RankTopic and HITSTopic with the baseline models. Table 2 reports our experimental results. For the purpose of comparison, results of RankTopic with $\xi = 0.5$ are selected to be shown. From the results, we can see that RankTopic performs better than PLSA, LDA, iTopic, topical ranking (TR) and is comparable with RTM. More importantly, RankTopic outperforms both of its ingredients, i.e. PLSA and topical ranking, which indicates that combining PLSA and ranking has much better clustering performance than each of them. Overall, the link combined topic models have better clustering performance than link ignored ones. NMI is not shown for Twitter since there is no background labels for users in that data.

TABLE 2
Clustering performance of different models on Citeseer and Cora data sets. For Ncut, the lower the better. For both Modu and NMI, the higher the better. For RankTopic, $\xi = 0.5$. For HITSTopic, $\xi = 0.3$. TR represents the topical ranking model.

Models	Citeseer			Cora			Twitter	
	Ncut	Modu	NMI	Ncut	Modu	NMI	Ncut	Modu
PLSA	2.92	0.35	0.14	4.85	0.16	0.11	5.88	0.35
LDA	2.68	0.38	0.21	4.30	0.24	0.19	4.76	0.37
iTopic	2.09	0.48	0.26	4.01	0.29	0.21	4.60	0.45
TR	1.99	0.50	0.17	4.74	0.18	0.14	5.37	0.38
RTM	1.63	0.54	0.31	2.98	0.47	0.32	4.24	0.47
RankTopic	1.60	0.55	0.28	3.01	0.47	0.30	2.77	0.53
HITSTopic	**0.52**	**0.68**	**0.32**	**1.66**	**0.61**	**0.48**	**2.72**	**0.60**



(a) iTopic (b) RankTopic (c) HITSTopic

Fig. 10. Clustering results of iTopic and RankTopic with $\xi = 0.5$ and HITSTopic with $\xi = 0.3$ on Citeseer data. The more a matrix looks like a block diagonal matrix, the better the clustering result summarizes the links.

From the results, we also see that HITSTopic performs significantly better than all the other models. The underlying reason is that the summation of topical authority and hubness of documents serves as more discriminated feature for clustering purpose than topical pagerank only.

We finally study the clustering results qualitatively in a visualized way. Since link structure can reflect the clusters of documents to some degree, the adjacency matrix of document network is taken for visualization. For example, clustering results of iTopic, RankTopic and HITSTopic on Citeseer data are illustrated in Figure 10. The clustering results for RankTopic with $\xi = 0.5$ and HITSTopic with $\xi = 0.3$ are shown for comparison. The documents clustered in the same class are arranged to be adjacent to each other in the visualized matrixes. The more a matrix looks like a block diagonal matrix, the better the clustering result summarizes the link structure. The results of PLSA and LDA look even worse than that for iTopic and that of RTM looks more or less the same as RankTopic. The visualization results are consistent with the quantitative results in Table 2.

However, there are large volume of community detection algorithms, such as spectral clustering [40] and PCL-DC [34], which aims at partitioning a network into clusters according to the links only. We do not compare RankTopic with them because the community detection algorithms directly perform clustering on links by optimizing measures like normalized cut and modularity. One drawback of those community detection algorithms is that they can only describe the community structure of the observational data but can not generalize the results to unseen data, which actually can be done by topic

modeling methods. In this sense, it is not fair to compare topic modeling methods with the community detection algorithms.

5.5 Document Classification

In this subsection, we study the performance of LIMTopic based models on document classification. We use an open source package, MATLABArsenal (http://finalfantasyxi.inf.cs.cmu.edu/), to conduct the following experiments. Due to the superior classification performance of SVM (Supporter Vector Machine), we select SVM_LIGHT with RBF kernel as the classification method, and set kernel parameter as 0.01 and cost factor as 3. However, classification method is not the main focus of this paper, and we just want to see how well different topic modeling results can serve as classification features. Recall that label information for publications in Citeseer and Cora data sets are known in advance, it is natural to choose the two data sets for classification purpose.

Similarly, we first study the effect of parameter ξ by empirically setting them to some typical values. Figure 11 shows the results on Citeseer Data. From the results, we see that RankTopic with $\xi = 0.3$ perform best in terms of classification accuracy. Overall, RankTopic performs well when ξ is at the middle of the range [0,1] and performs bad when ξ is close or equal to either 0 or 1. We obtain similar results on Cora data. Based on the results, we also compare RankTopic with the baseline models. Figure 12 shows the comparison results. It can be seen that the classification results built on topic features extracted by RankTopic are better than all the baseline topic models except RTM on Citeseer data set. Similar with the clustering results, the classification performance of RankTopic is comparable with RTM, which is one of competitive link combined topic models. From the classification results, we also see that HITSTopic are much better than RankTopic, from which we can further conclude that the summation of topical authority and hubness of documents are more discriminated than topical pagerank in machine learning tasks like classification and clustering.

Accuracy may not be a proper metric to evaluate a classifier when the class distributions are skewed over documents. In the following, we use some other well-known measures to compare diffident classifiers, namely Precision, Recall, F1 measure, Receiver Operation Curve (ROC) and Area Under the Curve (AUC). Since there are more than two classes in our problem, we compute the above measures except ROC for each class and average them as the overall performance of different classifiers. Table 3 shows the experimental results on Citeseer data. The ROC of different topic models for the 1st class of Citeseer data is shown in Figure 13, and we obtain similar results for other classes. From the experimental results, we can clear see that LIMTopic framework based models outperform the base line models in terms of

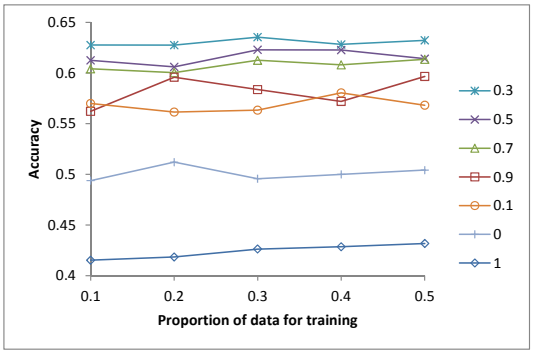
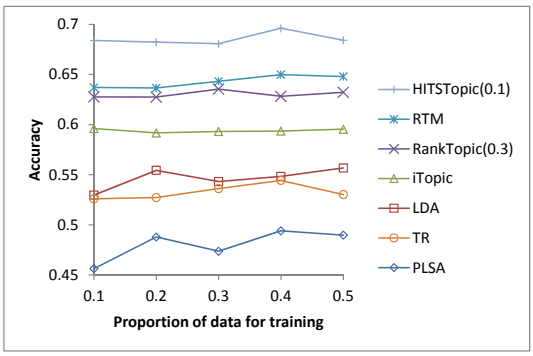
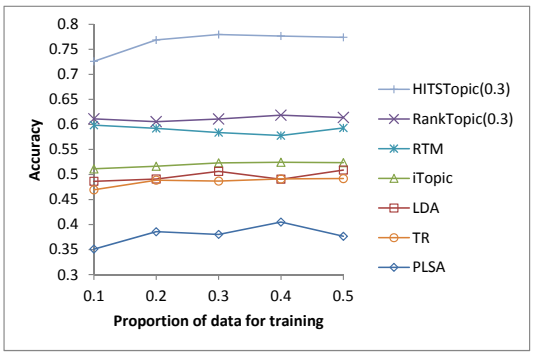


Fig. 11. Classification accuracy of RankTopic with some typical values of ξ on Citeseer data set for different proportions of training data. Accuracy is defined as the fraction of the correctly classified documents to the total number of documents. The higher the accuracy, the better the classification quality.



(a) Classification on Citeseer



(b) Classification on Cora

Fig. 12. Classification accuracy of different topic models on Citeseer and Cora data sets for different proportions of training data. The higher the accuracy, the better the model. TR represents topical ranking model.

all the above mentioned metrics. Similar results are obtained for Cora data.

From both the document classification and document clustering results, we conclude that topics detected by LIMTopic based models especially HITSTopic indeed serve as better features for documents than those detected by some baseline topic models, while RankTopic are comparable with one of the state-of-the-art link combined topic models RTM in both document clustering

TABLE 3
Average classification precision, recall, F1 and AUC of different topic models on Citeseer data

Models	Precision	Recall	F1	AUC
PLSA	0.3502	0.5133	0.4160	0.7366
LDA	0.4351	0.5806	0.4950	0.8213
iTopic	0.4555	0.6244	0.5250	0.8235
TR	0.3916	0.5713	0.4640	0.7842
RTM	0.4950	0.6516	0.5617	0.8401
RankTopic(0.3)	0.4884	0.6304	0.5478	0.8299
HITSTopic(0.1)	**0.5450**	**0.6783**	**0.6030**	**0.8789**

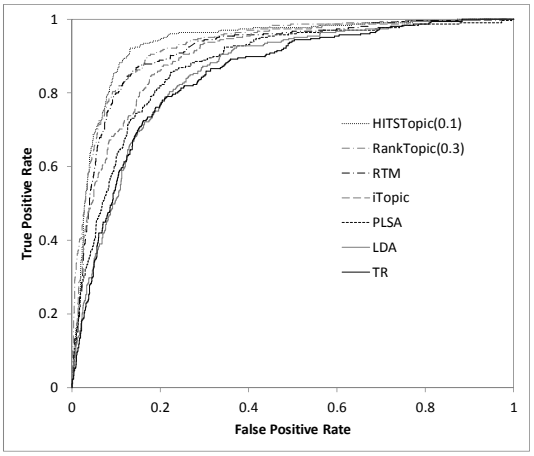


Fig. 13. ROC of different topic models for the 1st class of Citeseer data

and classification. Of course, to achieve the best performance, parameter ξ should be set properly, empirically ξ can be set to values close to 0.5.

5.6 Topic Interpretability

Topics detected by topic models are represented as a distribution over words in the vocabulary. The detected topics can be interpreted as high level concepts from their top probability words. The more easier the topics can be interpreted as meaningful concepts, the better the detected topics. We define the degree of how easy a topic can be interpreted as a semantically meaningful concepts as *topic interpretability*.

However, the interpreting process of a topic can be rather complicated, which depends on the domain knowledge and comprehensive ability of an interpreter. Nevertheless, there exist some methods that try to evaluate the topic interpretability in a quantitative way. One such method is to use point-wise mutual information (PMI) [41] between pairs of words to evaluate the topic coherence. Higher PMI reflects better topic interpretability. In our experiments, we represent each topic by using their top 10 words and compute PMI between those words. The PMI of a topic is computed as the average PMI of all pairs of top probability words of that topic.

From the parameter study, we find that when ξ is set to relatively low values, such as 0.1 and 0.3, the topic interpretability archives the best, while when ξ is set to

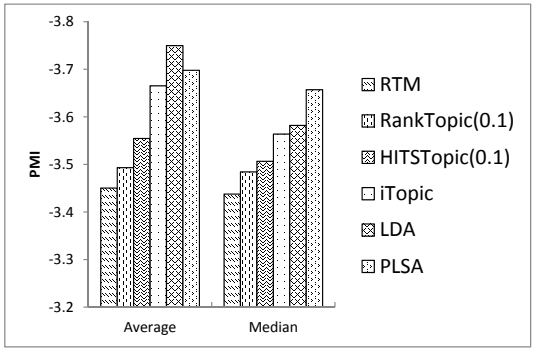


Fig. 14. The average and median PMI values of topics detected by different topic models. The lower the bar, the better topic interpretability. For RankTopic and HITSTopic, PMI for $\xi = 0.1$ is only shown.

0, the topic interpretability becomes worse. The results are consistent with those of generalization performance, which suggests that there are correlation between the generalization performance and topic interpretability of topic models. To compare the topic interpretability of different topic models, we compute the average and median of PMI values of topics detected by the baseline models. Figure 14 presents the comparison results, from which we can see that both RankTopic and HITSTopic performs better than some baseline topic models and are slightly worse than RTM in topic interpretability. Besides the quantitative evaluation of topic interpretability, we also compare the topics detected by RankTopic and one of the baseline models LDA in a manual way.

For example, Figure 15 shows one topic detected by LDA and two topics detected by RankTopic in Arnet-Miner data. The titles for the topics are manually given out according to the semantic of the top 10 words. Topic 4 detected by LDA is interpreted as Language by us. However, this topic is actually a mixture of two concepts. One is programming language, which is indicated by bold words. Another is natural language, which is indicated by underlined words. The two concepts are well distinguished by RankTopic as two topics, Topic 3 (Programming) and Topic 10 (Semantic). From the experiments, we also find out that RankTopic clearly discriminates topic Architecture detected by LDA as Computer Architecture and Service Oriented Architecture. Overall, all the 10 topics detected by RankTopic are easy to be interpreted to meaningfull research directions from the top probability words while some topics detected by LDA are difficult to be interpreted.

5.7 Document Network Summarization Performance

We would like to further empirically justify the prospect of integrating link based importance into topic modeling. Specifically, we see if LIMTopic can summarize the document network better than previous models.

We use the log-likelihood L_{rank} of ranking-integrated document-word matrix (see Equation **??**) as the fitness

Topic 4
Language
language
knowledge
programming
domain
development
oriented
context
semantics
semantic
languages

(a) LDA

Topic 3	Topic 10
Programming	Semantic
program	knowledge
code	semantic
programs	document
java	text
programming	documents
execution	retrieval
language	ontology
source	mining
type	content
flow	task

(b) RankTopic

Fig. 15. Example topics detected by LDA and RankTopic in ArnetMiner data set.

TABLE 4
Log-likelihoods of ranking-integrated document-word matrix with respect to different topic models.

log-likelihood	ArnetMiner	Citeseer	Cora	Twitter
PLSA	-88687	-31660	-11503	-45880
LDA	-85159	-31388	-10835	-45776
iTopic	-95815	-32860	-11758	-44999
RTM	-103837	-34268	-12960	-53072
HITSTopic	-85204	-30962	-10780	**-44184**
RankTopic	**-83566**	**-30746**	**-10693**	-44191

measure of a model for summarizing the whole document network. Notice that a document network include both the link structure and the document contents. L_{rank} is one of the unified ways to represent the whole document network. Therefore, it can be thought that higher L_{rank} value indicates that a model better fits to the document network.

Table 4 shows L_{rank} values of different topic models on our selected data sets. For RankTopic and HITSTopic, we choose the L_{rank} of $\xi = 0.7$ to illustrate, other ξ values except 0 produce more or less the same values. The comparison results show that LIMTopic based topic model outperforms almost all baseline models on all our selected data sets in L_{rank} value, reflecting that LIMTopic framework fits the ranking-integrated document-word matrix best among all the compared models. From the empirical results, LIMTopic reveals promising better performance for summarizing the whole document network than some state-of-the-art topic models in terms of L_{rank} measure.

6 Conclusion and Future Work

In this paper, we propose a framework LIMTopic to incorporate link based importance into topic modeling. As the instances of LIMTopic, RankTopic and HITSTopic is presented. To validate the effectiveness of LIMTopic based models, we have studied the performance of RankTopic and HITSTopic in various aspects, including generalization performance, document clustering and classification, topic interpretability, and document network summarization performance, and have compared LIMTopic with traditional topic models, PLSA and LDA, and link combined topic models, iTopic and RTM. Especially, we have investigated the model on a wide range of typical balancing parameter values and find out that LIMTopic is sensitive to that parameter and it is indeed necessary to introduce such parameter to combat link noises. Extensive experiments show that when properly setting balancing parameter ξ LIMTopic based model performs consistently better than all the baseline models in the above mentioned aspects on three public paper citation data sets and one twitter data set. Empirical results of KL-Divergences between topic distributions converted from topical link importance and those computed by basic topic model show that it is necessary to combine link based importance and topic modeling for better exploring document network, and we further show LIMTopic can better summarize the whole document network than other counterpart models in terms of a novel measure called L_{rank}. Moreover, we empirically show that LIMTopic's parameters tend to keep almost the same after enough iterations.

As future works, we will study how LIMTopic can benefit other applications, such as document retrieval and recommendation. Furthermore, we will implement LIMTopic framework in a distributed computing environment to make it scale up to large data sets. Moreover, we show the convergence of LIMTopic empirically, while leave the theoretical proof as an open problem for future research.

Acknowledgments

This work is supported by National Natural Science Foundation of China under Grants 70771043, 61173170 and 61300222, State Key Laboratory of Software Engineering under grants SKLSE2012-09-11, Innovation Fund of Huazhong University of Science and Technology under Grants 2012TS052, 2012TS053 and 2013QN120, Youth Foundation of National Computer network Emergency Response technical Team/Coordination Center of China under Grant 2013QN-19.

References

[1] F. Sebastiani, "Machine learning in automated text categorization," *ACM Computing Surveys*, vol. 34, no. 1, pp. 1–47, 2002.
[2] T. Hofmann, "Probabilistic latent semantic analysis," in *UAI*, 1999, pp. 289–296.
[3] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *Journal of Machine Learning Research*, vol. 3, no. Jan, pp. 993–1022, 2003.
[4] J. Tang, J. Zhang, L. Yao, J. Li, L. Zhang, and Z. Su, "Arnetminer: extraction and mining of academic social networks." in *KDD'08*, 2008, pp. 990–998.
[5] Y. Sun, J. Han, J. Gao, and Y. Yu, "iTopicmodel: Information network-integrated topic modeling," in *ICDM*, 2009, pp. 493–502.
[6] L. Nie, B. D. Davison, and X. Qi, "Topical link analysis for web search," in *SIGIR*, 2006, pp. 91–98.
[7] S. Brin and L. Page, "The anatomy of a large-scale hypertextual web search engine," *Computer Networks*, vol. 30, no. 1-7, pp. 107–117, 1998.
[8] J. M. Kleinberg, "Authoritative sources in a hyperlinked environment," *Journal of the ACM*, vol. 46, no. 5, pp. 604–632, 1999.
[9] N. Ma, J. Guan, and Y. Zhao, "Bringing pagerank to the citation analysis," *Information Processing and Management*, vol. 44, no. 2, pp. 800–810, 2008.
[10] Q. Mei, D. Cai, D. Zhang, and C. Zhai, "Topic modeling with network regularization," in *WWW*, 2008, pp. 101–110.

[11] D. M. Blei and J. D. Lafferty, "Dynamic topic models," in *ICML*, 2006, pp. 113–120.
[12] T. Iwata, T. Yamada, Y. Sakurai, and N. Ueda, "Online multiscale dynamic topic models," in *KDD*, 2010, pp. 663–672.
[13] X. Wang and A. McCallum, "Topics over time: a non-markov continuous-time model of topical trends," in *KDD*, 2006, pp. 424–433.
[14] Z. Yin, L. Cao, J. Han, C. Zhai, and T. S. Huang, "LPTA: A probabilistic model for latent periodic topic analysis," in *ICDM*, 2011, pp. 904–913.
[15] M. Rosen-Zvi, T. L. Griffiths, M. Steyvers, and P. Smyth, "The author-topic model for authors and documents," in *UAI*, 2004, pp. 487–494.
[16] R. Nallapati, A. Ahmed, E. P. Xing, and W. W. Cohen, "Joint latent topic models for text and citations," in *KDD*, 2008, pp. 542–550.
[17] C. X. Lin, Q. Mei, J. Han, Y. Jiang, and M. Danilevsky, "The joint inference of topic diffusion and evolution in social communities," in *ICDM*, 2011, pp. 378–387.
[18] J. Chang and D. M. Blei, "Hierarchical relational models for document networks," *The Annals of Applied Statistics*, vol. 4, no. 1, pp. 124–150, 2010.
[19] D. Duan, Y. Li, R. Li, R. Zhang, and A. Wen, "Ranktopic: Ranking based topic modeling," in *ICDM*, 2012, pp. 211–220.
[20] X. Wang, X. Jin, M.-E. Chen, K. Zhang, and D. Shen, "Topic mining over asynchronous text sequences," *Knowledge and Data Engineering, IEEE Transactions on*, vol. 24, no. 1, pp. 156–169, 2012.
[21] D. Duan, Y. Li, R. Li, Z. Lu, and A. Wen, "MEI: Mutual enhanced infinite community-topic model for analyzing text-augmented social networks," *The Computer Journal*, vol. 56, no. 3, pp. 336–354, 2013.
[22] C. Wang and D. M. Blei, "Collaborative topic modeling for recommending scientific articles," in *KDD*, 2011, pp. 448–456.
[23] C. Lin, Y. He, R. Everson, and S. Rüger, "Weakly supervised joint sentiment-topic detection from text," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 6, pp. 1134–1145, 2012.
[24] R. Nallapati, D. A. McFarland, and C. D. Manning, "Topicflow model: Unsupervised learning of topic-specific influences of hyperlinked documents," *Journal of Machine Learning Research - Proceedings Track*, vol. 15, pp. 543–551, 2011.
[25] T. H. Haveliwala, "Topic-sensitive pagerank: A context-sensitive ranking algorithm for web search," *IEEE Transactions on Knowledge and Data Engineering*, vol. 15, no. 4, pp. 784–796, 2003.
[26] J. Tang, J. Zhang, R. Jin, Z. Yang, K. Cai, L. Zhang, and Z. Su, "Topic level expertise search over heterogeneous networks," *Machine Learning*, vol. 82, no. 2, pp. 211–237, 2011.
[27] H. Sayyadi and L. Getoor, "Futurerank: Ranking scientific articles by predicting their future pagerank," in *SDM*, 2009, pp. 533–544.
[28] E. Yan, Y. Ding, and C. R. Sugimoto, "P-rank: An indicator measuring prestige in heterogeneous scholarly networks," *Journal of the American Society for Information Science and Technology*, vol. 62, no. 3, pp. 467–477, 2011.
[29] Y. Sun, Y. Yu, and J. Han, "Ranking-based clustering of heterogeneous information networks with star network schema," in *KDD*, 2009, pp. 797–806.
[30] Y. Sun, J. Han, P. Zhao, Z. Yin, H. Cheng, and T. Wu, "Rankclus: integrating clustering with ranking for heterogeneous information network analysis," in *EDBT*, 2009, pp. 565–576.
[31] S. Fortunato, "Community detection in graphs," *Physics Reports*, vol. 486, no. 3-5, pp. 75–174, 2010.
[32] U. von Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, no. 4, pp. 395–416, 2007.
[33] M. E. J. Newman, "Modularity and community structure in networks," *PNAS*, vol. 103, no. 23, pp. 8577–8582, 2006.
[34] T. Yang, R. Jin, Y. Chi, and S. Zhu, "Combining link and content for community detection: A discriminative approach," in *KDD*, 2009, pp. 927–935.
[35] L. Liu, J. Tang, J. Han, M. Jiang, and S. Yang, "Mining topic-level influence in heterogeneous networks," in *CIKM*, 2010, pp. 199–208.
[36] G. Heinrich, "Parameter estimation for text analysis," University of Leipzig, Tech. Rep., 2008.
[37] J. Shi and M. J., "Normalized cuts and image segmentation," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 8, pp. 888–905, 2000.
[38] M. E. J. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys Rev E Stat Nonlin Soft Matter Phys*, vol. 69, no. 2, 2004.
[39] R. Cilibrasi and P. M. B. Vitányi, "Clustering by compression," *IEEE Transactions on Information Theory*, vol. 51, no. 4, pp. 1523–1545, 2005.
[40] U. Luxburg, "A tutorial on spectral clustering," *Statistics and Computing*, vol. 17, pp. 395–416, 2007.
[41] D. Andrzejewski and D. Buttler, "Latent topic feedback for information retrieval," in *KDD*, 2011, pp. 600–608.

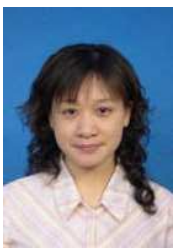
Dongsheng Duan is currently working in National Computer network Emergency Response technical Team/Coordination Center of China. He received his Ph.D. degree in Computer Science from Huazhong University of Science and Technology in 2013, and received his B.S. degree in Computer Science from Sichuan University, Chengdu, China, in 2008. His research interests are community detection, topic modeling and social network analysis.

Yuhua Li is currently an associate professor in the School of Computer Science and Technology at Huazhong University of Science and Technology, Wuhan, China. She received her Ph.D. degree in Computer Science from Huazhong University of Science and Technology in 2006. Her research interests include link mining, social network mining, graph mining, knowledge engineering, Semantic WEB and ontology. She is a senior member of China Computer Federation (CCF).

Ruixuan Li is currently a professor in School of Computer Science and Technology at Huazhong University of Science and Technology, Wuhan, China. He received his B.S., M.S. and Ph.D. in Computer Science from Huazhong University of Science and Technology in 1997, 2000 and 2004 respectively. His research interests include distributed data management, peer-to-peer computing, social network, and distributed system security. He is a member of IEEE and ACM.

Rui Zhang is currently an Associate Professor and Reader in the Department of Computing and Information Systems at The University of Melbourne, Australia. He received his B.S. from Tsinghua University in 2001 and his Ph.D. from National University of Singapore in 2006. His research interest is data and information management in general, particularly in areas of high-performance computing, spatial and temporal data analytics, moving object management, indexing techniques and data streams.

Xiwu Gu is currently a Lecturer in School of Computer Science and Technology at Huazhong University of Science and Technology, Wuhan, China. He received his B.S., M.S. and Ph.D. in Computer Science from Huazhong University of Science and Technology in 1989, 1998 and 2007 respectively. His research interests include distributed computing, data mining, social computing. He is a member of China Computer Federation (CCF).



Kunmei Wen is currently an associate professor in School of Computer Science and Technology at Huazhong University of Science and Technology, Wuhan, China. She received her B.S., M.S. and Ph.D. in Computer Science from Huazhong University of Science and Technology in 2000, 2003 and 2007 respectively. Her research interests include data management, social network, and information retrieval.