Mach Learn (2011) 82: 211–237
DOI 10.1007/s10994-010-5212-9

Topic level expertise search over heterogeneous networks

Jie Tang · Jing Zhang · Ruoming Jin · Zi Yang · Keke Cai · Li Zhang · Zhong Su

Received: 1 June 2009 / Accepted: 1 May 2010 / Published online: 17 September 2010

Abstract In this paper, we present a topic level expertise search framework for heterogeneous networks. Different from the traditional Web search engines that perform retrieval and ranking at document level (or at object level), we investigate the problem of expertise search at topic level over heterogeneous networks. In particular, we study this problem in an academic search and mining system, which extracts and integrates the academic data from the distributed Web. We present a unified topic model to simultaneously model topical aspects of different objects in the academic network. Based on the learned topic models, we investigate the expertise search problem from three dimensions: ranking, citation tracing analysis, and topical graph search. Specifically, we propose a topic level random walk method for ranking the different objects. In citation tracing analysis, we aim to uncover how a piece of

Editors: S.V.N. Vishwanathan, Samuel Kaski, Jennifer Neville, and Stefan Wrobel.

J. Tang (✉) · J. Zhang · Z. Yang
Department of Computer Science and Technology, Tsinghua University, Beijing, China
e-mail: jietang@tsinghua.edu.cn

J. Zhang
e-mail: zhangjing@keg.cs.tsinghua.edu.cn

Z. Yang
e-mail: yz@keg.cs.tsinghua.edu.cn

R. Jin
Department of Computer Science, Kent State University, Kent, OH 44241, USA
e-mail: jin@cs.kent.edu

K. Cai · L. Zhang · Z. Su
IBM, China Research Lab, Beijing, China

K. Cai
e-mail: caikeke@cn.ibm.com

L. Zhang
e-mail: lizhang@cn.ibm.com

Z. Su
e-mail: suzhong@cn.ibm.com

work influences its follow-up work. Finally, we have developed a topical graph search function, based on the topic modeling and citation tracing analysis. Experimental results show that various expertise search and mining tasks can indeed benefit from the proposed topic level analysis approach.

Keywords Social network · Information extraction · Name disambiguation · Topic modeling · Expertise search · Association search

1 Introduction

Heterogeneous networks are becoming prevalent in many real-world applications. For example in a heterogeneous academic network, there are different objects such as authors, conferences, and papers; in a product review system, there are objects like products, users, and reviews. The emerging complex networking data poses many fundamental challenges for search and mining of them.

Traditional keyword-based search essentially matches the queried keywords with documents, and object-oriented search extracts attributes for specific objects (e.g., product) and performs search at the object level. In contrast with the traditional search, our topic level search tries to extract the "semantics" of documents/queries and match them at the topic level. The fundamental issue in the topic level search is how to model the object/document with semantic topics. The problem becomes more challenging with the prevalence of heterogeneous networks, which usually consist of different types of objects. In addition, we need also consider how to employ the learned latent topical information to help the search and mining tasks.

Academic network is a typical heterogeneous network. Recently, the fast growing repository of scientific literature has posed many challenging issues in literature management. Preliminary statistics show that there are more than 1 million researchers, 3 million publications, and 32 million citation relationships in Computer Science (Tang et al. 2010). Previously, several systems have been developed for academic search, for instance DBLP[1], CiteSeer[2], Google Scholar[3], Libra[4], and Dblife[5]. However, most of these systems are simply modeling documents based on unigram language model (Zhai and Lafferty 2001). Libra (Nie et al. 2005) considers papers, authors, and conferences as different objects and utilizes a PopRank (by extending PageRank, Page et al. 1999) to rank the different objects. However, its search model is still based on keyword matching and does not consider topical aspects of the academic data. As a result, an object that is highly popular for one topic (even "spam" topic) may dominate the results of another topic in which it is less authoritative. We argue that existing modeling methods are insufficient for an in-depth analysis of such a large-scale heterogeneous network.

We begin with several motivating examples drawn from the academic daily life. First, when starting a work on a new research topic, a researcher, especially a beginning researcher, often wants to have a quick overview of the research topic: who are the experts on this topic?

[1] http://www.informatik.uni-trier.de/~ley/db/.

[2] http://citeseerx.ist.psu.edu/.

[3] http://scholar.google.com.

[4] http://libra.msra.cn/.

[5] http://dblife.cs.wisc.edu/.

what are the best papers? what are the authoritative publication venues in this research area? and what are the most well-known research labs? Next, to further the research, the researcher wants to have an in-depth understanding of the research field: what are the hot sub-topics discussed in this field? how this field has been evolving in the past years? what are the relationships between different research works? and how a piece of research work influences its follow-up work? Moreover, when got a new idea, the researcher is usually eager to know which papers he should refer to. Additionally, researchers often want to be informed with the research trend of the academic community, e.g., what are the hot research fields? which fields grew up quickly in the past years and which ones taper off? who are the most active researchers in a specific field?

To summarize, the fundamental challenges include: How to capture the topical "semantics" for each object in the academic network? and How to improve the quality of search and mining over the networking data with the learned topic models?

Recently, statistical topic models have attracted much attention from the research community. Topic models such as probabilistic Latent Semantic Indexing (pLSI) (Hofmann 1999) and Latent Dirichlet Allocation (LDA) (Blei et al. 2003) have been proposed to model latent topics among documents and have been successfully applied to multiple text mining tasks (McCallum et al. 2007; Mimno and McCallum 2007; Steyvers et al. 2004; Zhai et al. 2004). Inspired by the recent success of topic models for text mining, in this work, we are investigating a paradigm shift to enable search/analysis at topic level for the academic network with the following contributions:

- We present a probabilistic topic model to simultaneously model the different objects in the heterogeneous academic network by incorporating the link information (citation relationship).
- Based on the topic modeling results, we investigate how to improve the performance of expertise search, and then study two interesting issues: citation tracing analysis and topical graph search.
- We evaluate the effectiveness of the approaches and experimental results show that the topic level analysis can effectively improve the performance of expertise search (+15.4% in terms of MAP) and citation relationship categorization (+16.3%).

We have developed Arnetminer.org[6], a topic level academic search engine, in which we aim to provide comprehensive services for research communities. Currently, the system collects information of over 1 million researchers, 3 million publication papers (including 1 million full papers), and 8 thousand conferences. Services such as expertise search, people association search, topical graph search, and topic browser are provided in Arnetminer. The system is in operation on the internet since 2006 and receives a large amount of accesses from 190 countries.

The paper is organized as follows. In Sect. 2, we introduce the extracted heterogeneous academic network. In Sect. 3, we explain our approach for topical analysis over the heterogeneous network. In Sect. 4, we describe how to take advantage of the discovered topical aspects for expertise search. In Sect. 5, we introduce how we employ parallelization techniques to improve the topical analysis efficiency. Section 6 gives the experimental results. Section 7 reviews the related work. We conclude the paper in Sect. 8.

[6]http://arnetminer.org.

2 Data preparation

The academic data is distributed on the Web. For searching and mining the academic network, we need to first extract the networking data from the Web. We define the data model of the academic network (as shown in Fig. 1).

Some of the academic data can be extracted from structured data sources such as the publication information from DBLP; while other data needs to be extracted from unstructured Web pages such as researchers' homepages. We propose a unified approach to extract researcher profiles from the researchers' homepages. We integrate the publication data from online databases. We extract the organization information from Wikipedia using regular expressions.

Our technique contribution includes the unified approach for researcher profiling (Tang et al. 2007) and the approach for dealing with the name disambiguation problem in the integration (Zhang et al. 2007a). The unified approach for research profiling explored in this paper is based on a Tree-structured Conditional Random Fields (TCRFs) (Tang et al. 2006).

Researcher profiling Specifically, the researcher profiling approach consists of three steps: relevant page identification, preprocessing, and tagging. In relevant page identification, given a researcher name, we first get a list of web pages by a search engine (we use Google API) and then identify the homepage/introducing page using a classifier. The performance of the classifier is 92.39% in terms of F1-measure. In preprocessing, we separate the text into tokens and assign possible tags to each token. The tokens form the basic units in the following tagging step and the pages form the sequences of units. In tagging, given a sequence of units, we determine the most likely corresponding sequence of tags by using a trained tagging model. The type of the tags corresponds to the profile property (as shown in Fig. 1).

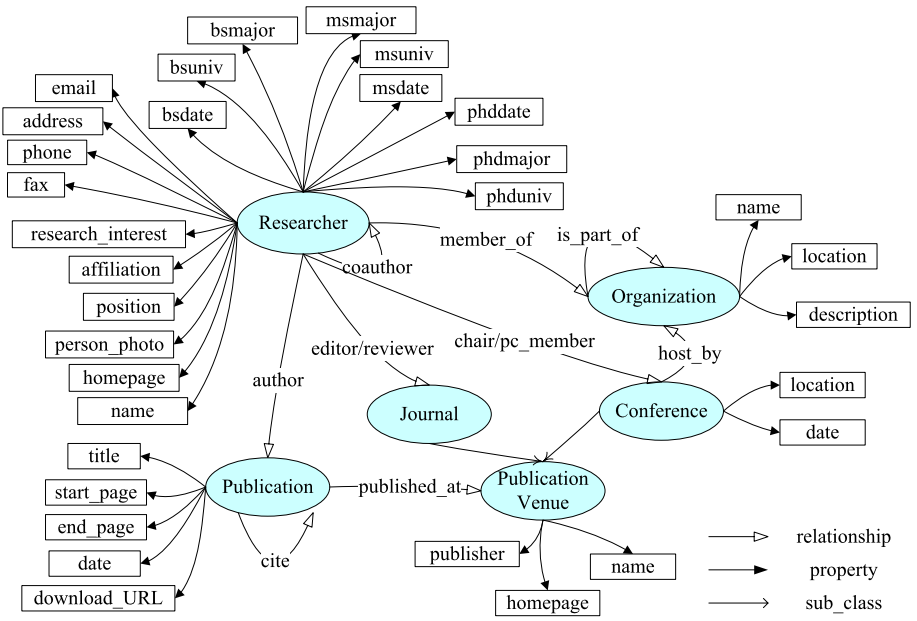

Fig. 1 The schema of academic network

As the tagging model, we use Tree-structured Conditional Random Fields (TCRFs) (Tang et al. 2006). TCRFs can model dependencies across hierarchically laid-out information. In researcher profile extraction, an identified homepage can be represented as an Document Object Model (DOM) tree. The root node corresponds to the Web page, a leaf node denotes a word token, and an inner node denotes a coarse information block (e.g., a block containing contact information). For parameter estimation, as the graphical structure in TCRFs can be a tree with cycles, exact inference will be expensive. We propose using the Tree-based Reparameterization (TRP) algorithm (Wainwright et al. 2001) to compute the approximate inference.

We evaluate the performance of the proposed approach on 2,000 randomly chosen researchers' homepages. Our approach can reach 86.70% (in terms of F1-measure) on average. We compare our method with several state-of-the-art methods, i.e., rule learning based method (Amilcare), classification based method (SVM-based method), and linear-chain CRFs. Our approach significantly outperforms (+3.4%–33.2%) the baseline methods for profile extraction.

Integration We collect the publication data from online databases including DBLP, ACM Digital library, Citeseer, and others. For integrating researcher profiles and the publication data, the author name is used as the identifier. Thus it is necessary to deal with the name ambiguity problem. The task of name disambiguation is defined as follows: Given a person name a, we denote all papers having the author name a as $P = \{p_1, p_2, \ldots, p_n\}$. Suppose there existing k actual researchers $\{y_1, y_2, \ldots, y_k\}$ having the name a, our task is to assign each of these n papers to its real researcher y_i. We propose a probabilistic framework for name disambiguation based on Hidden Markov Random Fields (HMRF) (Zhang et al. 2007a). The method effectively improves (+8%) the performance of name disambiguation, by comparing with the baseline methods on two real-world data sets.

Heterogeneous academic network The extracted/integrated data is stored into an academic network base. With the profiling and integration methods, we have already collected 1,048,504 researcher profiles, 3,258,504 publications, 8,042 conferences, and 34,215,473 paper-paper citation relationships, and 57,443,857 coauthor relationships. A detailed introduction about how the academic network has been constructed can be referred to (Tang et al. 2010). Based on the academic network, we have developed an academic search system, called Arnetminer. Services such as expertise search, citation tracing analysis, topical graph search, and topic browser are provided.

3 Topical analysis

The objective of topical analysis is to discover latent topics ("semantic" aspects) associated with each object in the academic network. Traditional keyword-based analysis tends to be overly specific in terms of matching words. Topic analysis can alleviate the problem by representing each object at the topic level. We present two topic models: Author-Conference-Topic (ACT) model and Citation-Tracing-Topic (CTT) model. In the next section, we will discuss how to take advantage of the discovered topics in academic search, citation tracing analysis, and topical graph search.

Before introducing the two topic models, we first give the definition of topic. Each topic z is defined as $\{(w_1, P(w_1|z)), \ldots, (w_{N1}, P(w_{N1}|z))\}$. The definition means that a topic is represented by a mixture of words and their probabilities belonging to the topic. The topic

Table 1 Notations

Symbol	Description
M	Number of papers
P	Number of citation contexts
V	Number of unique words in papers
T	Number of topics
K	Number of categories
N_d	Number of words in paper d
N_p	Number of words in citation context p
d	A paper
p	A citation context
w_{di}	The i-th word in paper d
w_{pi}	The i-th word in citation context p
z_{di}	The topic assigned to the i-th word in paper d
c_{pi}	The category assigned to the i-th word in citation context p
z^s_{pi}	The topic chosen from the citing paper for the i-th word in citation context p
z^t_{pi}	The topic chosen from the cited paper for the i-th word in citation context p
θ_d	Multinomial distribution over topics specific to paper d
ϕ_z	Multinomial distribution over words specific to z
$\psi_{z^s z^t}$	Multinomial distribution over relationship categories specific to the topic pair (z^s, z^t)
λ_c	Multinomial distribution over words specific to the category c

definition can be also extended to other information sources. For example, we can extend the topic definition by publication venues, i.e., $\{(c_1, P(c_1|z)), \ldots, (c_{N1}, P(c_{N1}|z))\}$. Further, each object can be associated with a set of topic distribution, e.g., a researcher a is associated with $\{P(z|a)\}_z$. Table 1 lists the major notations used throughout this paper.

3.1 Author-conference-topic model

Basically, the academic network is composed of three types of objects: authors, papers, and conferences. Our goal is to discover the latent topic distribution associated with each object. Modeling the different sources can be done in many different ways, for example, using the state-of-the-art language model (LM) (Baeza-Yates and Ribeiro-Neto 1999) or using a separated pLSI (Hofmann 1999) or LDA (Blei et al. 2003) for each type of object. However, the different types of objects in the academic network are often intertwined. Separately learning the model for each type of object cannot take advantage of the correlation between them, thus may result in unsatisfactory performance. Some preliminary experimental results (Tang et al. 2008a) confirm this assumption. Our main idea in this work is to use a unified probabilistic model to model papers, authors, and publication venues simultaneously.

The proposed topic model is called Author-Conference-Topic (ACT) model. For simplicity, we use conference to denote conference, journal, and book hereafter. The model simulates the process of writing a scientific paper using a series of probabilistic steps. In essence, the topic model uses a latent topic layer as the bridge to connect the different types of objects. More accurately, for each object it estimates a mixture of topic distribution which represents the probability of the object being associated with every topic. For example, for each paper, we have a set of probabilities $\{P(z_i|d)\}$, respectively denoting how likely paper d discusses a topic z_i.

We use Gibbs sampling for parameter estimation. During parameter estimation, the algorithm keeps track of a $A \times T$ (author by topic) count matrix, a $T \times V$ (topic by word) count matrix, and a $T \times C$ (topic by conference) count matrix. Given these three count matrices, we can estimate the probability of a topic given an author $P(z|a)$ or θ_{az}, the probability of a word given a topic $P(w|z)$ or ϕ_{zw}, and the probability of a conference given a topic $P(c|z)$ or ψ_{zc}. Interested reader can refer to (Tang et al. 2008a) for more details.

3.2 Citation-tracing-topic model

The ACT model can model the topical aspects associated with objects, but ignores the link information. We further investigate how to model the citation relationship in the topic model. Generally, we hope that the topic model can capture the characteristics of the citation relationship and its connected papers, for example, the correlation between the relationship category (e.g. "Basic theory" and "Comparable work") and the topic distribution of the papers. We propose a novel topic model, called Citation-Tracing-Topic (CTT) model, to extract topics in the paper and to categorize the citation relationship in a unified way. The basic idea is to use two correlated generative processes to fulfill the two subtasks simultaneously. Figure 2 shows the graphical representation of the CTT model. The first process (the left/right column of Fig. 2) is to model the topic distribution within each paper. The second process (the middle column of Fig. 2) is to model the citation relationship between the source paper and the target paper. Specifically, each citation relationship corresponds to a *Citation Context*, which is defined by the context words surrounding the citation position, e.g., the words "...We use Cosine computation (Andrieu et al. 2003) to evaluate the similarity ..." would be the citation context between the source paper and the target paper "(Andrieu et al. 2003)". Each citation context is denoted as p. We use the correlation between the two processes to

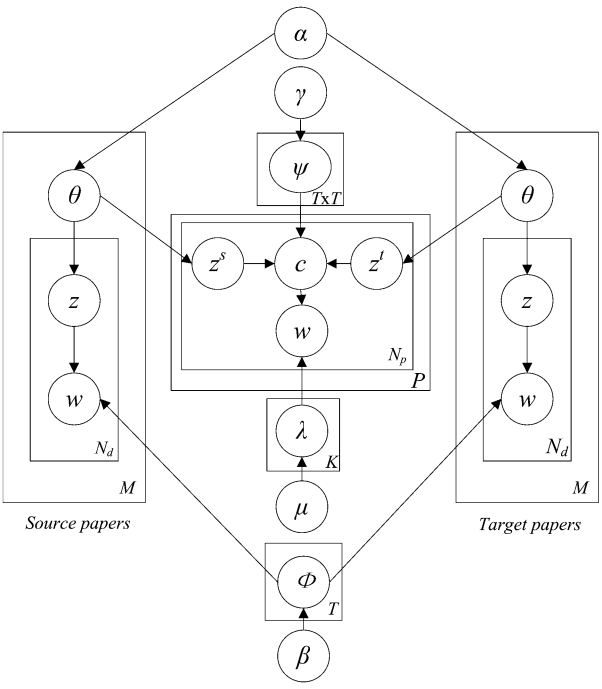

Fig. 2 Graphical representation of the CTT model

model dependencies between the topic distribution and the relationship category. Formally, the generative process of the CTT model is described as:

- For each word w_{di} in paper d:
  - draw a topic z_{di} from a multinomial $Mult(.|\theta_d)$;
  - draw the word w_{di} from a multinomial $Mult(.|\phi_z)$;
- For each word w_{pj} in citation context p:
  - draw a topic z^s from a multinomial $Mult(.|\theta_{p(s)})$ specific to the source paper $p(s)$ of the citation context p;
  - draw a topic z^t from a multinomial $Mult(.|\theta_{p(t)})$ specific to the target paper $p(t)$ of the citation context p;
  - draw a category c from a multinomial $Mult(.|\psi_{z^s z^t})$
  - draw the word w_{pj} from multinomial $Mult(.|\lambda_c)$;

where θ, ϕ, ψ, and λ are multinomial distributions respectively specific to paper d, topic z, topic-pair (z^s, z^t), and relationship category c. We assume that the four multinomials are sampled from Dirichlet distributions with priors α, β, γ, and μ.

Parameter estimation There are four sets of unknown parameters in the CTT model: (1) the distribution Θ of M paper-topics and the distribution Φ of T topic-words; (2) the distribution Ψ of $T \times T$ topic-pair-categories and the distribution Λ of K category-words; (3) the corresponding topic z_{di} for each word w_{di} in the paper d; and (4) the chosen topic pair (z^s, z^t) and the category c_{pj} for each word w_{pj} in the citation context p. It is usually intractable to exactly estimate the parameters in such a probabilistic model. A variety of algorithms have been proposed to conduct approximate estimation, for example variational EM methods (Blei et al. 2003), Gibbs sampling (Griffiths and Steyvers 2004; Steyvers et al. 2004), and expectation propagation (Griffiths and Steyvers 2004; Minka 2003). We chose Gibbs sampling for its ease of implementation. Instead of estimating the model parameters directly, we evaluate (a) the posterior distribution on just z and then use the results to infer Θ and Φ for the first generative process; (b) the posterior distribution on topic pair (z^s, z^t), and category c, and then use the sampling results to infer Ψ and Λ for the second generative process. More specifically, we begin with the joint distribution of variables $\mathbf{w}$, $\mathbf{w}_{pair}$, $\mathbf{z}$, $\mathbf{z}^s$, $\mathbf{z}^t$, $\mathbf{c}$ given α, β, γ, and μ as:

$$
\begin{aligned}
&P(\mathbf{w}, \mathbf{w}_{pair}, \mathbf{z}, \mathbf{z}^s, \mathbf{z}^t, \mathbf{c}|\alpha, \beta, \gamma, \mu) \\
&= \int_\theta \int_\phi \int_\lambda \int_\psi P(\mathbf{w}, \mathbf{z}|\theta, \phi) P(\mathbf{w}_{pair}, \mathbf{z}^s, \mathbf{z}^t, \mathbf{c}|\theta, \lambda, \psi) \\
&\quad \times P(\theta|\alpha) P(\phi|\beta) P(\lambda|\mu) P(\psi|\gamma) d\theta d\phi d\lambda d\psi \\
&= \prod_{z=1}^{T} \frac{\Delta(\mathbf{n}_z + \beta)}{\Delta(\beta)} \prod_{d=1}^{M} \frac{\Delta(\mathbf{n}_d + \alpha)}{\Delta(\alpha)} \prod_{z^s=1}^{T} \prod_{z^t=1}^{T} \frac{\Delta(\mathbf{n}_{z^s z^t} + \gamma)}{\Delta(\gamma)} \prod_{c=1}^{C} \frac{\Delta(\mathbf{n}_c + \mu)}{\Delta(\mu)}
\end{aligned} \tag{1}
$$

where $\mathbf{n}_z$ denotes a set of numbers with each n_{zv} corresponding to the number of times that word w_v has been generated by topic z; $\mathbf{n}_d$ denotes a set of numbers with each n_{dz} corresponding to the number of times that topic z has been generated from document d; $\mathbf{n}_{z^s z^t}$ denotes a set of numbers with each $n_{z^s z^t c}$ corresponding to the number of times that category c has been generated by topic pair (z^s, z^t); $\mathbf{n}_c$ denotes a set of numbers with each n_{cv} corresponding to the number of times that word w_v has been generated by category c;

function $\Delta(\mathbf{n}_d + \alpha)$, for example, is defined as:

$$\Delta(\mathbf{n}_d + \alpha) = \frac{\prod_{z=1}^{T} \Gamma(n_{dz} + \alpha)}{\Gamma(\sum_{z=1}^{T}(n_{dz} + \alpha))} \tag{2}$$

where $\Gamma(.)$ is a gamma function.

And then using the chain rule, we can obtain the conditional probability $P(z_{di}|\mathbf{z}_{-di}, \mathbf{w}, \alpha)$ for sampling the topic for each word in the first generative process:

$$P(z_{di}|\mathbf{z}_{-di}, \mathbf{w}, \alpha) = \frac{n_{dz_{di}}^{-di} + \alpha}{\sum_z (n_{dz}^{-di} + \alpha)} \frac{n_{z_{di}w_{di}}^{-di} + \beta}{\sum_v (n_{z_{di}v}^{-di} + \beta)} \tag{3}$$

where n_{dz} is the number of times that topic z has been sampled from paper d; n_{zw} is the number of times that word w has been generated by topic z; the number n^{-di} with the superscript $-di$ denotes a quantity, excluding the current instance.

For the second generative process, with an analogous process, we can first sample a topic-pair (z^s, z^t) from the source and the target papers for each word in the citation context p by:

$$P(z_{pj}^s, z_{pj}^t|\mathbf{z}_{-pj}^s, \mathbf{z}_{-pj}^t, \mathbf{c}) = \frac{n_{p(s)z_{pj}^s}^{-pj} + \alpha}{\sum_z (n_{p(s)z}^{-pj} + \alpha)} \frac{n_{p(t)z_{pj}^t}^{-pj} + \alpha}{\sum_z (n_{p(t)z}^{-pj} + \alpha)} \frac{n_{z_{pj}^s z_{pj}^t c_{pj}}^{-pj} + \gamma}{\sum_c (n_{z_{pj}^s z_{pj}^t c}^{-pj} + \gamma)}. \tag{4}$$

We then sample the category c_{pj} for each word w_{pj} by:

$$P(c_{pj}|\mathbf{c}_{-pj}, \mathbf{w}_p, \mathbf{z}^s, \mathbf{z}^t, \gamma) = \frac{n_{z_{pj}^s z_{pj}^t c_{pj}}^{-pj} + \gamma}{\sum_c (n_{z_{pj}^s z_{pj}^t c}^{-pj} + \gamma)} \frac{n_{c_{pj}w_{pj}}^{-pj} + \mu}{\sum_v (n_{c_{pj}w_v}^{-pj} + \mu)}.$$

After training the CTT model, we can obtain the probability $P(z|d)$ of a topic z given paper d, the probability $P(w|z)$ of a word w given topic z, and the probability $P(c|z^s, z^t)$ of a category c given a topic-pair (z^s, z^t).

As for the hyperparameters α, β, γ, and μ, one could estimate the optimal values using a Gibbs EM algorithm (Andrieu et al. 2003; Minka 2003) or a variational EM method (Blei et al. 2003). For some applications, topic models are sensitive to the hyperparameters (Asuncion et al. 2009) and it is necessary to get the right values for the hyperparameters. As for citation tracing analysis, we found that the estimated topic models are not very sensitive to the hyperparameters. Thus, for simplicity, we took a fixed value.

4 Topic level expertise search

Based on the topic analysis results, the system provides various search/analysis services, e.g., research profile search, expertise search, citation tracing analysis, academic suggestion, and topic browser. We explained some of them in our prior work (Tang et al. 2008a). In this paper, we will focus on those not covered in the prior work, i.e. ranking at the topic level, citation tracing analysis, and topical graph search.

4.1 Ranking at the topic level

One of the most important module in a search engine is to estimate the relative importance of each object, i.e. to rank the objects. PageRank (Page et al. 1999) is one of the state-of-the-art algorithm for this purpose by analyzing link structures. However, the conventional PageRank does not consider the topic information. To break this limitation, we propose a *topical random walk* algorithm. The basic idea is to integrate the topic modeling results into a random walk framework for improving the object ranking quality (Tang et al. 2008b). We consider two methods to integrate the discovered topics into the random walk framework for ranking objects in the academic network.

Random walk (RW) We first briefly introduce the process of random walk over the academic network. The academic network can be considered to be composed of three composite networks. At the center is a directed graph of paper citations $G_d = (V_d, E_{dd})$, where V_d includes all papers, and the directed edge $(d_1, d_2) \in E_{dd}$ suggests the paper d_1 cites the paper d_2. The relationships between authors and papers are modeled by a bipartite graph $G_{ad} = (V_a \cup V_d, E_{ad})$ where V_a is the set of authors, and author-paper relationships are recorded in the edge-set E_{ad}. Similarly, we can define another bipartite graph $G_{cd} = (V_c \cup V_d, E_{cd})$ between publication venues and papers.

We can define a random walk over the academic network. The transition probability between nodes can be defined in different ways, e.g. by an average scoring scheme or based on the topic modeling results. Given this, similar to PageRank, the random walk ranking score for each object x can be defined as:

$$r[x] = \frac{\xi}{|V|} + (1 - \xi) \times \sum_{(x,y)\in E} \lambda_{yx} r[y] P(x|y) \tag{5}$$

where $|V|$ is the number of nodes in the network; ξ is a random jump parameter; λ_{yx} is the transition probability between the type of node y and the type of node x; $P(x|y)$ is the probability between two specific nodes y and x. A similar definition has been previously used for ranking objects in heterogeneous networks (Nie et al. 2005).

Proposed 1: Random walk with topic nodes (RWTN) The first proposed method for combining the topic model with random walk is to integrate the discovered topics into the random walk. It augments the academic network with topic nodes V_z. Let $G_{td} = (V_z \cup V_d, E_{zd})$ be a bipartite graph between papers and topics, where V_z is the set of topic nodes estimated by the topic model, and if paper d can be generated from topic z with a probability $P(d|z) > \epsilon$ (where ϵ is a parameter to control the density of the constructed network), then we have an edge $(z, d) \in E_{zd}$. Similarly, we can define edges E_{cz} between conferences and topics and edges E_{az} between authors and topics. We conduct random walk on the new network. The random walk can make use of the topic distribution associated with each object.

In this method, we consider that after the random surfer walks to a topic node from some other node, he/she will has different transition probabilities to walk back to different types of nodes. The transition probabilities are calculated using the ACT topic model (e.g., $P_{ACT}(a|z)$), more specifically,

$$P(z_i|a_j) = \theta_{a_j z_i}, \tag{6}$$

$$P(a_j|z_i) = \frac{P(z_i|a_j)P(a_j)}{P(z_i)}, \tag{7}$$

$$P(z_i|d_j) = \frac{1}{A_d} \sum_{x \in \mathbf{a}_d} \theta_{xz_i}, \tag{8}$$

$$P(d_j|z_i) = P(\mathbf{w}_d|z_i) = \prod_{i=1}^{N_d} P(w_{di}|z_i), \tag{9}$$

$$P(c_j|z_i) = \psi_{z_i c_j}, \tag{10}$$

$$P(z_i|c_j) = \frac{P(c_j|z_i)P(z_i)}{P(c_j)} \tag{11}$$

where θ and ψ are obtained by parameter estimation for the ACT model; $P(z_i)$, $P(a_j)$, and $P(c_j)$ can be obtained by counting the number after Gibbs sampling.

Note that we can adjust the different parameters λ to weight how the random walk and the topic model will affect the final rank. If λ from the other nodes to the topic nodes are very small, then the dominant portion of the ranking will be determined by the random walk. If λ from the other nodes to the topic nodes are very large, then the dominant portion of the ranking will be determined by the topic model.

Proposed 2: Random walk at topic level (RWTL) Another proposed method is to run the random walk at the topic level. In RWTL, for each object x in the heterogeneous academic network, we introduce a vector of ranking scores $\{r[x, z]_{z=1}^{T}\}$. The random walk is still performed at the original academic network G, but at the topic level. That is, for any two linked nodes, we consider random walk between the two nodes within the same topic and across different topics. Formally, the topic level ranking score of paper d is defined as:

$$\begin{aligned} r[d, z_k] = {} & \xi \frac{1}{|V|} P(z_k|d) \\ & + (1-\xi) \sum_{d':d' \to d} \left[\epsilon P(d|d', z_k) + (1-\epsilon) \frac{1}{T} \sum_{j \neq k} p(d, z_k|d', z_j) \right] \end{aligned} \tag{12}$$

where $p(z|d)$ is the probability of topic z generated by paper d and it is calculated by a similar equation to (8); ξ is the random jump factor; parameter ϵ represents the preference to the topic-intra transition or the topic-inter transition; the probability $P(d|d', z_k)$ is defined as $\frac{1}{E_{d'*}}$; here $E_{d'*}$ denotes the set of edges pointing from paper d' to the other nodes; the probability $P(d, z_k|d', z_j)$ is defined as $P(z_k|d)P(z_j|d')$.

Similarly, we can define the ranking scores for authors and conferences. Further we combine the ranking scores with the relevance score the topic model by multiplication.

Academic search In the academic network, expertise search can be decomposed into four sub-tasks: person search, expert search, publication search, and conference search. The user can input any types of queries (e.g., a person name such as "Jie Tang", a general query such as "data mining", or a composite query such as "Jie Tang KDD mining"). The system first parses the query and identifies "semantics" of the query (e.g., "author:Jie Tang; conf:KDD; keyword:mining"), and then retrieves objects by combining the topic model and the word-based language model, e.g. (formulas for other objects can be defined similarly.)

$$P(q|d) = P_{LM}(q|d) \times P_{ACT}(q|d) \tag{13}$$

where $P_{LM}(q|d)$ is the generating probability of query q from paper d by the language model and $P_{ACT}(q|d)$ is the generating probability by the ACT model. The two probabilities are respectively calculated by:

$$P_{LM}(q|d) = \prod_{w \in q} \left(\frac{N_d}{N_d + \eta} \cdot \frac{tf(w,d)}{N_d} + \left(1 - \frac{N_d}{N_d + \eta}\right) \cdot \frac{tf(w,\mathbf{D})}{N_{\mathbf{D}}} \right), \quad (14)$$

$$P_{ACT}(q|d,\theta,\phi) = \prod_{w \in q} \left(\frac{1}{A_d} \sum_{x \in \mathbf{a}_d} \sum_{z=1}^{T} P(w|z,\phi_z) P(z|x,\theta_x) \right) \quad (15)$$

where N_d is the number of word tokens in document d, $tf(w,d)$ is the frequency (i.e., occurring times) of word w in d, $N_{\mathbf{D}}$ is the number of word tokens in the entire collection, and $tf(w,\mathbf{D})$ is the frequency of word w in the collection $\mathbf{D}$; η is the Dirichlet smoothing factor and is commonly set according to the average document length in the collection (Zhai and Lafferty 2001); $\mathbf{a}_d$ denotes all authors of paper d; A_d is the number of authors, i.e., $A_d = |\mathbf{a}_d|$.

The topic model captures the *general* meaning of the query through its associated topics and the language model describes the *specific* meaning in terms of matching words. Thus the combination of the two models results in a balance between generality and specificity. Further, we combine the random walk ranking score with the relevance score. The simplest method is to multiply the two scores, e.g.:

$$R[x] = r[x] \times P(q|x) \quad (16)$$

where $r[x]$ is the random walk ranking score and $P(q|x)$ is the relevance score of the query q to the object x.

For combining the ranking score of the first proposed topical random walk method with the relevance score, we can use the same equation as (16). For combining the ranking score of our second topical random walk method with the relevance score, we can sum up either all topic level ranking scores or only the top-K ranking scores, and then multiply the obtained score with the relevance score by a similar equation as (16).

4.2 Citation tracing analysis

The goal of citation tracing analysis is to construct a citation tracing graph based on paper contents and their citation relationships. An ideal citation graph would provide the following information to the user: (1) topical aspects discussed in each paper; (2) semantics of the citation relationship between two papers; and (3) the strength of each citation relationship. With such a graph, a researcher could easily trace the origins of an idea/technique, analyze the evolution and impact of a research topic, filter the citations by certain categories of citation relationships, as well as zoom in and zoom out the citation tracing graph with the degree of influence.

Specifically, we define three categories of citation relationships: "Basic theory" (the source paper is based on the theory introduced in the target paper), "Comparable work" (the source paper is a comparable work of the target paper, e.g., an enhancement of an existing algorithm), and "Other" (neither "Basic theory" nor "Comparable work"). We use the CTT model to simultaneously model the topic distribution of each paper and the mixture of latent topics and relationship categories. Next, we propose a method to categorize the citation relationship based on the estimated topic models. Then, we present a method to calculate the influential strength of a citation relationship by considering the topic distribution of the source and the target papers as well as the topic-category mixture.

Citation relationship categorization With the topical analysis by the CTT model, we can obtain the probability of a topic given a paper θ_{dz}, the probability of a word given a topic ϕ_{zv}, the probability of a category given a topic-pair $\psi_{z^s z^t c}$, and the probability of a word in the citation contexts given a category λ_{cv}. For determining the category of a citation relationship, we can make use of the topic modeling results to calculate the posterior probability of the category of the citation relationship given a citation context by:

$$P(c|p) = \sum_{z^s=1}^{T} \sum_{z_t=1}^{T} P(c|z^s, z^t) P(z^s|p(s)) P(z^t|p(t)) \tag{17}$$

where $p(s)$ and $p(t)$ denotes the source paper and the target paper of citation context p respectively. The intuition is that the relationship category is determined by topics presented in the citing and the cited papers. The last two probabilities on the right part of (17) represent the topic distribution in the citing and the cited papers, respectively; while the first probability $P(c|z^s, z^t)$ captures the correlation between the relationship category and the distribution of topic-pair.

Citation influence estimation The other task in the citation tracing analysis is to estimate the influence of each citation relationship (citation context). We use two different methods to calculate the influential strength: one is category independent and the other is category dependent.

In the category independent method, we use KL-divergence, a standard measure of the difference between two probability distributions, to estimate the strength of the citation relationship. The intuition is: if two papers describe a similar content (a small divergence between their topic distributions), then the cited paper may have strong influence on the citing paper. Thus the influential strength (*IS*) of each citation context p based on the KL-divergence measure is defined as:

$$IS(p) = D_{KL}(\theta_{p(s)} \| \theta_{p(t)}) = \sum_{z=1}^{T} \theta_{p(s)z} \log \frac{\theta_{p(s)z}}{\theta_{p(t)z}}. \tag{18}$$

The other way to measure the influential strength is based on the learned models. Basically, we use the mixture of the topic distribution and the category distribution in the citation context to calculate the influential strength. This measure is derived from the observation: different categories of citation relationships have different influential strengths. We therefore define the influential strength (*IS*) based on the sampling results of the citation context as:

$$IS(p) = \frac{1}{N_p} \sum_{j=1}^{N_p} P(c_{pj}|z^s_{pj}, z^t_{pj}) P(z^s_{pj}|p(s)) P(z^t_{pj}|p(t)) \tag{19}$$

where z^s_{pj}, z^t_{pj}, c_{pj} is the sampled topic pair and the sampled category for the word w_{pj} in the citation context p.

Generating citation tracing graph Figure 3 shows a snippet of the constructed citation tracing graph. In Fig. 3, the model identifies the topic distribution of each paper, e.g. the paper "Self-Indexing Inverted Files for Fast Text Retrieval" (Moffat and Zobel 1996) has a high topical distribution on "Ranking and Inverted Index" (Topic 31). The model further

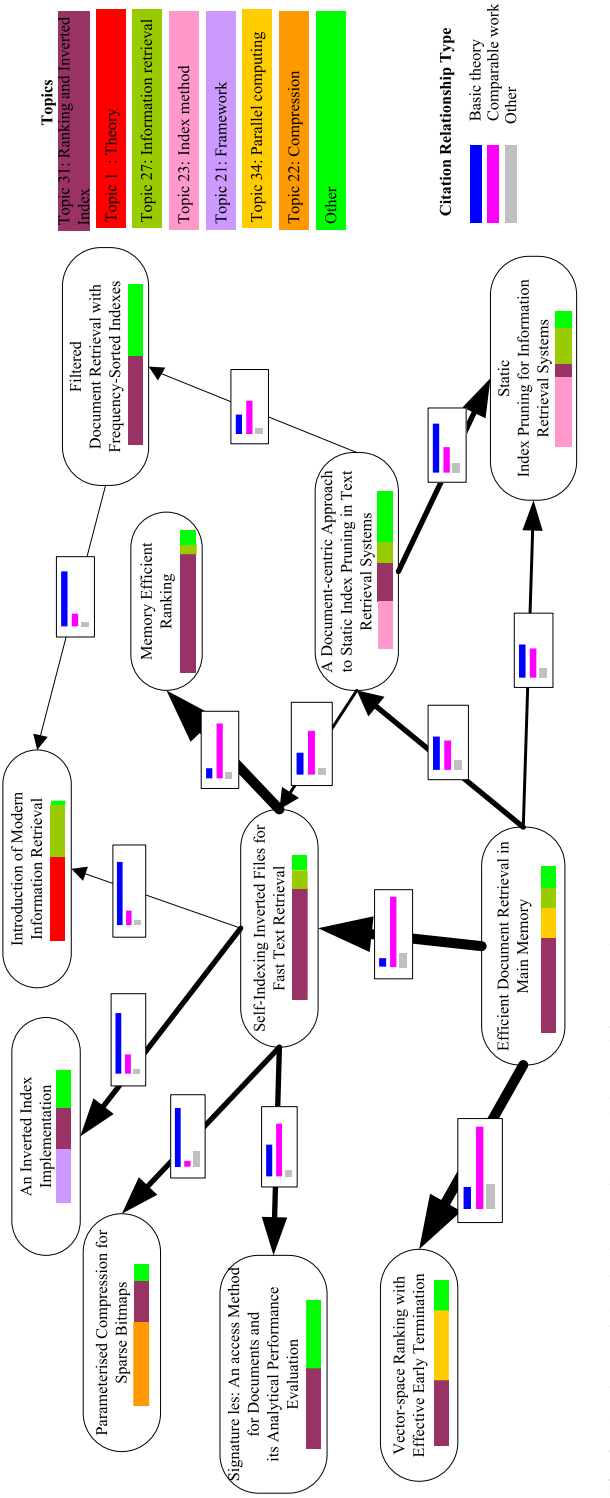

Fig. 3 A snippet of the citation tracing graph learned from our dataset

identifies that it has a strong "Comparable work" relationship with the paper "Memory Efficient Ranking" (Moffat et al. 1994), a strong "Basic theory" relationship with the paper "An Inverted Index Implementation" (McDonell 1977), and a weak "Basic theory" relationship with the paper "Introduction of Modern Information Retrieval" (Salton and McGill 1986). We see that with such a graph, an in-depth understanding of a research area can be easily grasped at the first glance.

Based on the citation tracing graph, one can provide powerful citation tracing analysis. As for the example in Fig. 3, we can filter the other categories of citation relationships and focus on the "Comparable work" based citation network. We can also use one paper as the center and obtain a bird's-eye graph by filtering the not-so-influential citation relationships.

4.3 Topical graph search

Based on the topic modeling and citation tracing analysis, we investigate a novel problem of topical graph search. Given a query, different from the conventional search engine that returns a list of documents, the topical graph search returns a series of topic-based graphs. These topic-based graphs can help users to quickly identify whether the returned information by a search engine is what they need and what kind of information (topics) contained in the returned documents, before taking a close look at the documents.

Processing flow With the learned models by CTT, we design and implement the topical graph search. Figure 4 shows the processing flow of the topical graph search.

To retrieve citation tracing graphs, we employ a three-step method. In the first step, we retrieve relevant papers to a given query. This can be done using any retrieval method. In this paper, we use (16). Specifically, we calculate the relevance of each query word to a paper using the hidden topics. Then we multiply the relevance of all query words and obtain the relevance of the query to each paper. Papers with a higher relevance score (larger than a threshold) are chosen as candidates. In the second step, we identify major topics from these candidate papers. The topic distribution $\{P(z|d)\}_z$ is employed to calculate the probability of a topic given all candidates, by accumulating the distributions for different topics. Topics with the highest probability are selected as major topics. Finally, we generate a citation tracing graph for each topic, which consists of related papers and citation relationships between these papers. Each citation relationship is associated with an influence score and a category.

We generate a semi-structured summary for each citation tracing graph. The summary mainly includes the most impact papers, highly used keywords/keyphrases, active authors, and hot conferences.

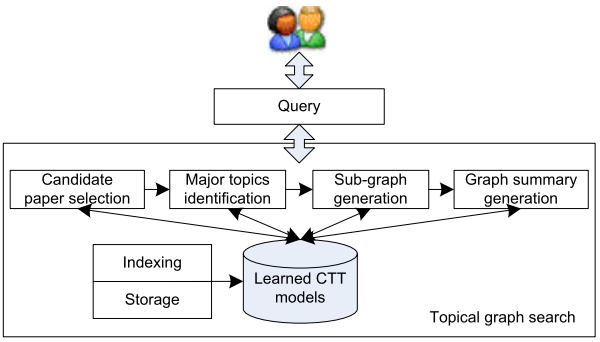

Fig. 4 Processing flow in the topical graph search

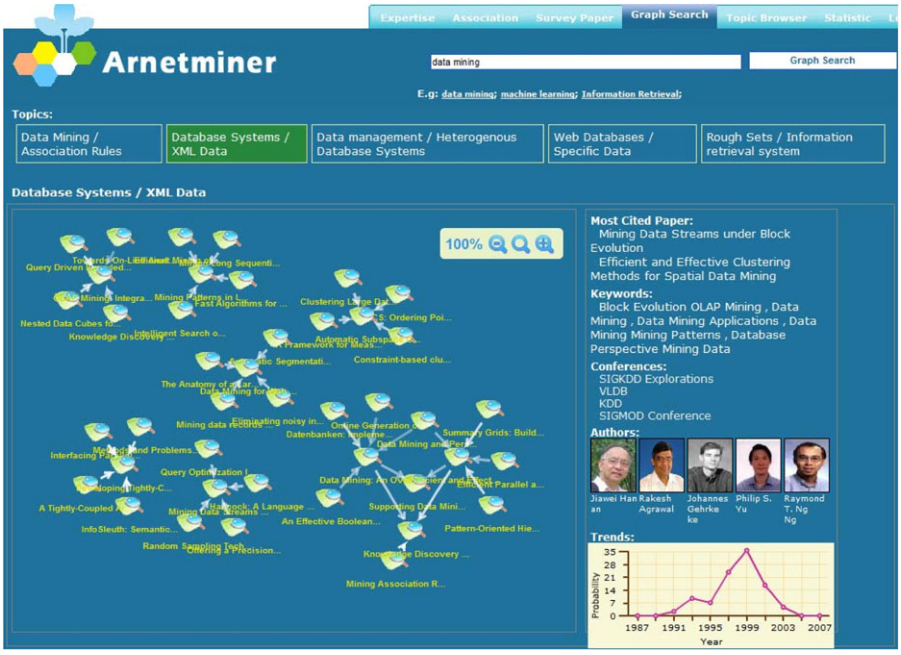

Fig. 5 Demonstration of topical graph search

Demo Figure 5 shows a screenshot of the topical graph search. The query is "data mining". Five relevant topics (e.g., "data mining", "database", and "data management") are identified, and for each topic, a citation tracing graph is constructed (as shown in Fig. 5). For each graph, a semi-structured summary is generated and displayed on the right of the graph. The demo is publicly available at http://arnetminer.org/.

5 Parallelization

We employ parallelization techniques to improve the runtime performance of the system. Generally speaking, parallelization can be utilized in every module in our system. We employed a paralleled program to crawl and extract the academic data from the Web. We can also use parallelization techniques for storage and access of the academic networking data. But so far, we have found that the bottleneck of the system performance is the topical analysis. For example, it needs more than one week to train the ACT model. In this section, we use training of the ACT model as an example to introduce how we employ parallelization techniques to improve the efficiency of topical analysis.

Inspired by the distributed inference for LDA (Newman et al. 2007), we implement a distributed inference algorithm over multiple processors for the ACT model. The basic idea is to conduct the inference in a "distribute-and-merge" way. In distribution, given P processors, we partition the author-specific (author by topic) $A \times T$ count matrix to the P processors, with $A_p = A/P$ rows on each processor and duplicate the other (topic by word, topic by conference) matrices to each processor. We conduct Gibbs sampling on each processor for a number of internal iterations independently. The duplicated matrices will be updated independently. In merge, we combine the count matrices to guarantee the consistence between

them. More accurately, we respectively update each element of two duplicated (topic by word, topic by conference) matrices by:

$$n_{zv}^{(new)} = n_{zv}^{(old)} + \sum_{p=1}^{P} (n_{zv}^{(p)} - n_{zv}^{(old)}), \tag{20}$$

$$n_{zc}^{(new)} = n_{zc}^{(old)} + \sum_{p=1}^{P} (n_{zc}^{(p)} - n_{zc}^{(old)}) \tag{21}$$

where the number $n^{(old)}$ with the superscript (*old*) denotes the count before distribution and the number $n^{(new)}$ with the superscript (*new*) denotes the count after merging. The number $n^{(p)}$ denotes the count obtained after the independent sampling on each processor.

6 Experimental results

We evaluate the proposed methods in the context of Arnetminer system (http://arnetminer.org). We perform three types of experiments for topical analysis, academic search, and citation relationship categorization.

6.1 Topical analysis

We perform topic model estimation on the entire Arnetminer data (1,048,365 researcher and 3,225,343 papers). We preprocessed each paper by (a) removing stopwords and numbers; (b) removing words that appear less than three times in the corpus; and (c) downcasing the obtained words. Finally, we obtained $V = 870{,}729$ unique words and a total of 107,532,489 words in the paper data set and a total of 50,584,237 words in the citation contexts.

In our experiments of topic model, the number of topics was fixed at $T = 200$ and the hyperparameters α, β, γ, and μ were set with $\alpha = 0.01$, $\beta = 0.01$, $\gamma = 0.01$, and $\mu = 0.1$ respectively. Each parameter was tuned by minimizing the perplexity (Blei et al. 2003), a standard measure for estimating the performance of a probabilistic model (the lower the better), with the other parameters fixed. We ran 5 independent Gibbs sampling chains for 2,000 iterations each after a burn-in process of 200 iterations. Each chain is randomly initialized. When training the topic model on a Server with one Intel Xeon processor (3.0 GHz) and 4 GB memory, the run time per chain was about one week. When distributed training on 8 processors, the run time per chain was about 54 hours. A complete topic modeling result can be found at http://arnetminer.org/topicBrowser.do. In Table 2, we illustrate five topics discovered by the ACT model on the Arnetminer data set.

6.2 Academic search

We conduct experiments to evaluate whether the topic level analysis can help academic search.

Data sets As there is no a standard data set with ground truth and also it is difficult to create such a data set of ground truth, for the evaluation purpose, we first select 60 most frequent queries from the query log of the Arnetminer system; then we remove the overly specific or lengthy queries (e.g., 'A Convergent Solution to Subspace Learning') and normalize similar

Table 2 Five topics discovered by the ACT mode on the Arnetminer publication data set. Each topic is shown with the top 10 words and their corresponding conditional probabilities. Below are top 7 authors and top 7 conferences associated with each topic. The titles are our interpretation of the topics. (CL—Computational Linguistics, JMLR—Journal of Machine Learning Research, MLSS—Machine Learning Summer School, JAIR—J. Artif. Intell. Res., and LNLP—Learning for Natural Language Processing.) The table is derived from (Tang et al. 2008a)

Topic #5 "Natural language processing"		Topic #10 "Semantic web"		Topic #16 "Machine learning"		Topic #19 "Support vector machines"		Topic #24 "Information extraction"	
language	0.034820	semantic	0.068226	learning	0.058056	support	0.082669	learning	0.065259
parsing	0.023766	web	0.048847	classification	0.018517	vector	0.071373	information	0.043527
natural	0.019029	ontology	0.043160	boosting	0.015881	machine	0.064076	extraction	0.033592
learning	0.015871	knowledge	0.041497	machine	0.017797	kernel	0.026897	web	0.019311
approach	0.012712	learning	0.013431	feature	0.013904	regression	0.020544	semantic	0.011860
grammars	0.012712	framework	0.012095	classifiers	0.013904	neural	0.016308	text	0.010618
processing	0.011923	approach	0.011427	margin	0.013245	classification	0.012072	rules	0.010618
text	0.011923	based	0.010758	selection	0.012586	networks	0.011366	relational	0.009376
machine	0.011133	management	0.010090	algorithm	0.025086	model	0.010660	logic	0.009376
probabilistic	0.010343	reasoning	0.009502	kernels	0.011269	algorithm	0.009269	programming	0.008755
Yuji Matsumoto	0.001389	Steffen Staab	0.005863	Robert E. Schapire	0.004033	Bernhard Scholkopf	0.003929	Raymond J. Mooney	0.010346
Eugene Charniak	0.001323	Enrico Motta	0.004365	Yoram Singer	0.003318	Johan A.K. Suykens	0.003536	Andrew McCallum	0.004074
Rens Bod	0.001323	York Sure	0.003713	Thomas G. Dietterich	0.002472	Vladimir Vapnik	0.002947	Craig A. Knoblock	0.003492
Brian Roark	0.001190	Nenad Stojanovic	0.001824	Bernhard Scholkopf	0.001496	Olvi L. Mangasarian	0.002947	Nicholas Kushmerick	0.002457
Suzanne Stevenson	0.001124	Alexander Maedche	0.001824	Alexander J. Smola	0.001301	Joos Vandewalle	0.002030	Ellen Riloff	0.002199
Anoop Sarkar	0.001058	Asuncion Gomez-Perez	0.001694	Ralf Schoknecht	0.001236	Nicola L.C. Talbot	0.001768	William W. Cohen	0.002134
Claire Cardie	0.000992	Frank van Harmelen	0.001563	Michael I. Jordan	0.001106	Bart De Moor	0.001703	Dan Roth	0.001487

Table 2 (*Continued*)

Topic #5 "Natural language processing"		Topic #10 "Semantic web"		Topic #16 "Machine learning"		Topic #19 "Support vector machines"		Topic #24 "Information extraction"	
ACL	0.253487	ISWC	0.125291	NIPS	0.289761	Neural Computation	0.096707	AAAI	0.295846
COLING	0.234435	EKAW	0.122379	JMLR	0.206583	NIPS	0.094388	IJCAI	0.192995
CL	0.118136	IEEE Intelligent Systems	0.071418	ICML	0.156389	ICANN	0.084338	ICML	0.060567
ANLP	0.060423	CoopIS/DOA/ ODBASE	0.065594	COLT	0.096157	JMLR	0.083565	KDD	0.058551
CoRR	0.058674	K-CAP	0.054674	Neural Computation	0.023017	Neurocomputing	0.071197	JAIR	0.046451
COLING-ACL	0.036814	ESWS	0.023369	MLSS	0.011545	Machine Learning	0.067331	ECML	0.033006
NAACL	0.035065	WWW	0.016817	Machine Learning	0.010827	ESANN	0.045686	IIWeb	0.026956

queries (e.g., 'Web Service' and 'Web Services' to 'Web Service'); finally we obtain 43 queries. We use 7 queries and conduct evaluation on a subset of the data (including 14,134 authors, 10,716 papers, and 1,434 conference) from the Arnetminer system. For evaluation, we use the method of pooled relevance judgments (Buckley and Voorhees 2004) together with human judgments. Specifically, for each query, we first pool the top 30 results from three similar (academic search) systems (Libra, Rexa, and Arnetminer) into a single list. Then, two faculties and five graduate students from CS provided human judgments. Four grade scores (3, 2, 1, and 0) are assigned respectively representing best relevance, relevance, marginal relevance, and not relevance. For example, for annotating persons, assessments were carried out mainly in terms of how many top conference/journal papers he or she has published, how many times his/her papers have been cited, and what distinguished awards he or she has been awarded. Finally, the judgment scores were averaged to construct the ground truth.

Experimental setting We conduct evaluation in terms of P@5 (Precision for the top five returned results), P@10, P@20, R-pre, and mean average precision (MAP) (Buckley and Voorhees 2004; Craswell et al. 2005).

We use BM25 (Robertson et al. 1996), language model (LM) (Baeza-Yates and Ribeiro-Neto 1999), pLSI (Hofmann 1999), LDA (Blei et al. 2003), and the Author-Topic (AT) model (Rosen-Zvi et al. 2004; Steyvers et al. 2004) as baseline methods. BM25 is a state-of-the-art method for information retrieval. In BM25, we use the method in (Robertson et al. 1996) to calculate the relevance of a query and a paper. For language model, we use (14) to calculate the relevance between a query term and a paper and for pLSI, LDA, AT, we use a similar equation to (15) to calculate the relevance of a term and a paper. We also compare with the results obtained by combining BM25 or LM with random walk using the multiplication combination.

To learn the topic model, for pLSI we estimated the topic distribution using the EM algorithm (Hofmann 1999). For LDA and AT, we performed model estimation with the same setting as that for the ACT models. We empirically set the number of topics as $T = 80$ for all topic models.

Results Table 3 shows the experimental results of retrieving papers, conferences, and authors using our proposed methods and the baseline methods on the collected evaluation queries. +RW denotes integration of a method into the random walk. +RWTN denotes to combine the proposed random walk with topic nodes method, and +RWTL denotes to combine the proposed random walk at topic level. We see that our proposed topic models outperform the baseline methods (BM25, LM, pLSI, LDA, and AT). Without random walk, the improvements of our proposed topic models over the baseline methods range from 7.4% to 14.3% in terms of MAP. Based on all other evaluation measures, our methods consistently perform better than the baseline methods. We can also see that ACT+RWTL (random walk at topic level) achieves the best performance in terms of most evaluation measures. ACT+RW is better than ACT+RWLT in terms of P@10 and P@20. It is difficult to tell which method (ACT+RWLT and ACT+RW) is better. As both perform the ranking at the topic level, we can conclude that topic level search can achieve a better performance. We conducted sign tests on the results, which indicates that improvements are statistically significant ($p \ll 0.01$). We also tried different topic numbers (80, 120, 140) and found that the topic level expertise search methods are consistently superior to the baseline methods.

For comparison purpose, we evaluated the results returned by Libra (libra.msra.cn) and Rexa (rexa.info), two academic search engines. The average MAP obtained by Libra and

Table 3 Performance of academic ranking approaches (%). LM—language model; pLSI—probabilistic Latent Semantic Indexing; LDA—Latent Dirichlet Allocation; AT—Author Topic Model; ACT—Author-Conference-Topic model; +RW—combining random walk

Method	Object	P@5	P@10	P@20	R-pre	MAP
BM25	Paper	42.9	45.7	41.4	12.0	47.2
	Author	77.1	47.1	26.4	67.5	85.5
	Conference	51.4	38.6	22.9	48.8	66.0
	Average	57.1	43.8	30.2	42.8	66.2
BM25+RW	Paper	71.4	55.7	46.4	15.7	67.2
	Author	62.9	47.1	26.4	64.6	71.9
	Conference	51.4	34.3	22.1	48.8	58.1
	Average	61.9	45.7	31.7	43.1	65.7
LM	Paper	40.0	38.6	37.1	10.0	46.4
	Author	65.7	44.3	25.0	58.8	73.4
	Conference	51.4	32.9	21.4	47.6	63.1
	Average	52.4	38.6	27.9	38.8	61.0
LM+RW	Paper	62.9	55.7	44.3	12.9	65.3
	Author	71.4	48.6	25.7	64.6	83.8
	Conference	60.0	35.7	22.1	53.6	64.6
	Average	64.8	46.7	30.7	43.7	71.2
pLSI	Paper	32.5	33.8	30	9.7	40.4
	Author	65.0	40.0	22.5	60.4	75.5
	Conference	47.5	36.3	21.3	45.1	54.1
	Average	48.3	36.7	24.6	38.4	56.7
LDA	Paper	31.4	48.6	42.9	13.5	45.8
AT	Paper	42.9	48.6	42.9	13.1	49.3
	Author	82.9	45.7	25.7	73.5	78.1
	Average	62.9	47.1	34.3	43.3	63.7
ACT	Paper	42.9	45.7	43.6	16.6	51.0
	Author	91.4	50.0	26.4	80.0	89.6
	Conference	62.9	41.4	23.6	60.7	72.3
	Average	65.7	45.7	31.2	52.4	71.0
ACT+RW	Paper	68.6	61.4	50.7	17.1	66.6
	Author	80.0	51.4	27.1	77.6	87.4
	Conference	62.9	42.9	23.6	59.5	72.0
	Average	70.5	51.9	33.8	51.4	75.4
ACT+RWTN	Paper	45.7	40.0	38.6	13.4	52.2
	Author	71.4	44.3	24.3	65.4	71.5
	Conference	51.4	32.9	20.0	53.6	60.7
	Average	56.2	39.1	27.6	44.1	61.4
ACT+RWTL	Paper	71.4	48.6	37.1	16.0	70.3
	Author	82.3	50.0	25.7	79.4	89.1
	Conference	64.3	45.4	24.1	64.2	73.9
	Average	72.7	48.0	29.0	53.2	77.8

Rexa on our collected queries are respectively 48.3% and 44.9%. We see that our methods clearly outperform the two systems.

6.3 Citation relationship category

We evaluate the performance of citation relationship categorization. This provides us with another opportunity to quantitatively compare the proposed topical analysis approach with the existing work. Specifically, we apply the proposed approach and the multi-class SVM based baseline method to predict the category of each citation relationship.

Evaluation measures and baseline methods We conduct two experiments. The first experiment was to train the topic model and to categorize the citation relationship without using the prior information and the second experiment was with the prior information. In both experiments, we evaluate the performance of relationship categorization by Precision, Recall, and F1-measure, Accuracy, and Area Under the ROC Curve (AUC). The AUC score represents the area under ROC curve, which details the rate of true positives against false positives over the range of possible thresholds (Brefeld and Scheffer 2005). The area of that curve is the probability that a randomly drawn positive example has a higher decision function value than a random negative example.

We define a baseline method based on the method proposed by Nanba and Okumura (Nanba and Okumura 1999). Specifically, we used the words appearing in the citation context as the features and defined the feature values by the number of occurrences. Then a classification model was learned for the three categories of citation relationship using the multi-class SVM (Tsochantaridis et al. 2004). We use SVM-light, which is available at http://svmlight.joachims.org/. We choose polynomial kernel, because our preliminary experimental results show that it works best for our current task. We use the default values for the parameters in SVM-light. For categorizing the citation relationship, we apply the learned classification model to each citation relationship. We also compare with another method by combining the multi-class SVM with extracted latent topics of each paper. Specifically, we use LDA, a state-of-the-art topic model (Blei et al. 2003), to extract topics from papers. We then use topics extracted from the citing paper and cited paper as features of a citation context to learn the classification model. We call this method multi-class SVM+LDA.

Results Table 4 shows the result of the relationship categorization. For multi-class SVM, multi-class SVM+LDA, and Supervised CTT, we conduct five-fold cross-validation experiments as follows. We split the human labeled data into five average sets and used four of them for training and the remaining one for test. We evaluate the average performance of these methods on the data set. For unsupervised CTT, we remove the human annotated labels, and applied (17) to derive the relationship category.

We see from Table 4 that the extracted latent topics can be useful for identifying the relationship category. For example, by combining topics extracted by LDA into multi-class SVM, we obtain an improvement of +3.68% in terms of F1-score. By further integrating the mixture of category-topic distribution discovered by our proposed model, we can obtain significant improvements (+16.33% than multi-class SVM and +12.65% than multi-class SVM+LDA in terms of F1-measure, respectively). This indicates that the proposed approach is effective for identifying the category of the citation relationship. We have also found that the performance is not satisfactory without using any prior information: the performance of Unsupervised CTT is less than half of Supervised CTT by F1-measure). This result is consistent with that of (Mei et al. 2007). It means that incorporation of the prior information into our approach is necessary.

Table 4 The performance of categorizing citation relationship by Multi-class SVM and our approach (%)

Approach	Category	Precision	Recall	F1-measure	Accuracy	AUC
Multi-class SVM	Basic theory	64.09	53.60	58.24	88.50	84.33
	Comparable work	77.66	83.34	80.33	76.59	83.39
	Other	64.63	59.58	61.77	79.71	80.50
	Avg.	68.79	65.51	66.78	81.60	82.74
Multi-class SVM+LDA	Basic theory	66.31	58.56	61.71	89.24	86.52
	Comparable work	80.69	85.47	82.93	79.71	85.94
	Other	69.58	64.36	66.75	82.42	84.76
	Avg.	72.19	69.46	70.46	83.79	85.74
Unsupervised CTT	Basic theory	16.81	43.09	24.19	59.82	52.72
	Comparable work	60.88	31.48	41.50	48.81	54.01
	Other	29.74	34.73	32.04	59.57	54.95
	Avg.	35.81	36.44	32.58	56.07	53.89
Supervised CTT	Basic theory	57.82	93.92	71.58	88.91	90.11
	Comparable work	96.23	80.06	87.40	86.69	86.56
	Other	89.68	91.02	90.34	94.66	74.74
	Avg.	81.24	88.33	83.11	90.08	83.81

We conduct sign tests for each subtask on the results, which indicates that all the improvements of Supervised CTT over Multi-class SVM and Multi-class SVM+LDA are statistically significant ($p \ll 0.01$).

7 Related work

7.1 Random walk

Random walk theory gained popularity in 1973, originally used for examining stock prices and gained popularity in computer science due to the large number of Web-based networks becoming available. Considerable researches have been conducted on analyzing link structures to better understand the Web-based networks. For example, the page rank algorithm is a state-of-the-art algorithm proposed by Brin and Page to estimate the importance of a Web page (Page et al. 1999). The basic idea in PageRank is to calculate the importance of each Web page based on the scores of the pages pointing to the page and thus Web pages pointed by many high quality pages become more important. HITS is another classic algorithm for ranking Web pages based on link analysis (Kleinberg 1999). HITS divides the notion of importance of Web pages into two related attributes: hub and authority. HITS calculates two scores respectively for hub and authority by reinforcing them via the linkage between pages. The basic idea is that a good authority page would be pointed by many high hub-scored pages and a good hub page should also point to high authority-scored pages.

Many research efforts have been made to extend the algorithms. For instance, Xi et al. (2004) proposed a unified link analysis framework called link fusion to consider both the inter- and intra-type link structure among multi-type inter-related data objects. Nie et al. (2005) propose an object-level link analysis model, called PopRank, to rank the objects within a specific domain. Liu et al. (2005) build a weighted, directed co-authorship network using the co-authorship relationships in digital libraries, and propose

an AuthorRank algorithm to rank authors. See also (Zhang et al. 2007b; Garfield 1972; Dom et al. 2003). However, most existing methods only perform ranking by using the link information between web pages, but do not consider the topic information.

Our work is very different from the existing research since we address heterogeneous networks whereas most of the previous work focuses on homogeneous networks (that is, the type of objects in the network is unique, e.g., only web pages). Moreover, we develop different methods to conduct random walk at the topic level. We note that some efforts (Xi et al. 2004, 2005; Nie et al. 2005) have also been placed for addressing the heterogeneous networks. However, none of them consider the topic information in the random walk.

7.2 Topic model

Considerable work has been conducted for learning topics from text. For example, Hofmann (1999) proposes the probabilistic latent semantic indexing (pLSI) and applies it to information retrieval (IR).

Blei et al. (2003) introduce a three-level model, called Latent Dirichlet Allocation (LDA). The generative process of LDA closely resembles pLSI except that in pLSI, the topic mixture is conditioned on each document while in LDA, the topic mixture is drawn from a conjugate Dirichlet prior that remains the same for all documents. However, the two models does not consider the link information, thus they cannot model the citation relationships in our problem.s

Several extensions of the topic model have been proposed, for example, the Author model (McCallum 1999), and the Author-Topic model (Rosen-Zvi et al. 2004; Steyvers et al. 2004). The major difference of the proposed CTT model from the existing model is that we incorporate the citation relationship and the citation category information into the topic modeling process.

McCallum et al. have also studied several topic models in social network analysis (McCallum et al. 2007). They propose the Author-Recipient-Topic (ART) model, which learns topic distributions based on emails sent between people.

Compared with these prior work, in this paper, we present two topic models: ACT and CTT models. The former is to simultaneously model topics of different objects in the academic network and the latter is to model topic distribution of links.

7.3 Academic search

For academic search, several research issues have been intensively investigated, e.g. expert finding and paper suggestion.

Expert finding is one of the most important issues for mining social networks. For example, both Nie et al. (2007) and Balog et al. (2006) propose extended language models to address the expert finding problem. TREC also provides a platform for researchers to empirically evaluate their methods for expert finding (Craswell et al. 2005). McNee et al. (2002) employ collaborative filtering in citation network to recommend citations.

In addition, a few systems have been developed for academic search such as, scholar.google.com, libra.msra.cn, citeseer.ist.psu, and Rexa.info. Though much work has been performed, to the best of our knowledge, the problem of topic level expertise search over heterogeneous networks has not been sufficiently investigated. Our system addresses all these problems holistically. The proposed approach can achieve a better performance on expertise search than existing system by taking advantage of the topic information.

8 Conclusion

In this paper, we introduce the notion of topic level expertise search and propose methods to address this problem. Specifically, we present two topic models to model the heterogeneous academic network. The first ACT model can model the different objects simultaneously; while the second CTT model can model the link (citation relationship) into the topic model. Based on the topic modeling results, we study three important search/mining issues: expertise search, citation tracing analysis, and topical graph search. Experimental results on expertise search and citation relationship categorization show that the topic level analysis can effectively improve the performance of academic search (+15.4% in terms of MAP) and citation relationship categorization (+16.3%).

The general problem of topic level expertise search represents an new and interesting research direction in heterogeneous social networks. There are many potential future directions of this work. It would be interesting to further investigate novel topic models for in-depth analysis of the academic network. It would be also interesting to study the influence between different types of objects in the heterogeneous network. Another potential issue is to apply the proposed approaches to other domain such as company search, local search, and blog search.

Acknowledgements The work is supported by the Natural Science Foundation of China (No. 60703059, No. 60973102), Chinese National Key Foundation Research (No. 60933013), National High-tech R&D Program (No. 2009AA01Z138), and Chinese Young Faculty Research Fund (No. 20070003093). We also thank Prof. Philip Yu for his valuable suggestions.

References

Andrieu, C., de Freitas, N., Doucet, A., & Jordan, M. I. (2003). An introduction to mcmc for machine learning. *Machine Learning*, *50*, 5–43.

Asuncion, A., Welling, M., Smyth, P., & Teh, Y. W. (2009). On smoothing and inference for topic models. In *Proceedings of the twenty-fifth annual conference on uncertainty in artificial intelligence (UAI'09)* (pp. 27–34).

Baeza-Yates, R., & Ribeiro-Neto, B. (1999). *Modern information retrieval*. New York: ACM.

Balog, K., Azzopardi, L., & de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of the 29th ACM SIGIR international conference on information retrieval (SIGIR'2006)* (pp. 43–55).

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993–1022.

Brefeld, U., & Scheffer, T. (2005). Auc maximizing support vector learning. In *Proceedings of ICML'05 workshop on ROC analysis in machine learning*.

Buckley, C., & Voorhees, E. M. (2004). Retrieval evaluation with incomplete information. In *Proceedings of the 27th annual international ACM SIGIR conference on research and development in information retrieval (SIGIR'04)* (pp. 25–32).

Craswell, N., de Vries, A. P., & Soboroff, I. (2005). Overview of the trec-2005 enterprise track. In *TREC 2005 conference notebook* (pp. 199–205).

Dom, B., Eiron, I., Cozzi, A., & Zhang, Y. (2003). Graph-based ranking algorithms for e-mail expertise analysis. In *Data mining and knowledge discovery* (pp. 42–48).

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science*, *178*(4060), 471–479.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. In *Proceedings of the national academy of sciences (PNAS'04)* (pp. 5228–5235).

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22nd international conference on research and development in information retrieval (SIGIR'99)* (pp. 50–57).

Kleinberg, J. M. (1999). Authoritative sources in a hyperlinked environment. *Journal of the ACM*, *46*(5), 604–632.

Liu, X., Bollen, J., Nelson, M. L., & de Sompel, H. V. (2005). Co-authorship networks in the digital library research community. *Information Processing & Management*, *41*(6), 681–682.

McCallum, A. (1999). Multi-label text classification with a mixture model trained by em. In *Proceedings of AAAI'99 workshop on text learning*.

McCallum, A., Wang, X., & Corrada-Emmanuel, A. (2007). Topic and role discovery in social networks with experiments on enron and academic email. *Journal of Artificial Intelligence Research (JAIR)*, *30*, 249–272.

McDonell, K. J. (1977). An inverted index implementation. *The Computer Journal*, *20*(1), 116–123.

McNee, S. M., Albert, I., Cosley, D., Gopalkrishnan, P., Lam, S. K., Rashid, A. M., Konstan, J. A., & Riedl, J. (2002). On the recommending of citations for research papers. In *Proceedings of the 2002 ACM conference on computer supported cooperative work (CSCW'02)* (pp. 116–125).

Mei, Q., Ling, X., Wondra, M., Su, H., & Zhai, C. (2007). Topic sentiment mixture: modeling facets and opinions in weblogs. In *Proceedings of the 16th international conference on world wide web (WWW'07)* (pp. 171–180).

Mimno, D., & McCallum, A. (2007). Expertise modeling for matching papers with reviewers. In *Proceedings of the 13th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD'07)* (pp. 500–509).

Minka, T. (2003). Estimating a Dirichlet distribution. In *Technique report*. http://research.microsoft.com/minka/papers/dirichlet/.

Moffat, A., & Zobel, J. (1996). Self-indexing inverted files for fast text retrieval. *ACM Transactions on Information Systems*, *14*(4), 349–379.

Moffat, A., Zobel, J., & Sacks-Davis, R. (1994). Memory efficient ranking. *Information Processing and Management*, *30*(6), 733–744.

Nanba, H., & Okumura, M. (1999). Towards multi-paper summarization using reference information. In *Proceedings of the sixteenth international joint conference on artificial intelligence (IJCAI'99)* (pp. 926–931).

Newman, D., Asuncion, A., Smyth, P., & Welling, M. (2007). Distributed inference for latent Dirichlet allocation. In *Proceedings of the 19th neural information processing systems (NIPS'07)*.

Nie, Z., Zhang, Y., Wen, J.-R., & Ma, W.-Y. (2005). Object-level ranking: bringing order to web objects. In *Proceedings of the 14th international conference on world wide web (WWW'05)* (pp. 567–574).

Nie, Z., Ma, Y., Shi, S., Wen, J.-R., & Ma, W.-Y. (2007). Web object retrieval. In *Proceedings of the 16th international conference on world wide web (WWW'07)* (pp. 81–90).

Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). *The pagerank citation ranking: bringing order to the web* (Technical Report SIDL-WP-1999-0120). Stanford University.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gatford, M., & Payne, A. (1996). Okapi at trec-4. In *Text retrieval conference*.

Rosen-Zvi, M., Griffiths, T., Steyvers, M., & Smyth, P. (2004). The author-topic model for authors and documents. In *Proceedings of the 20th international conference on uncertainty in artificial intelligence (UAI'04)* (pp. 487–494).

Salton, G., & McGill, M. J. (1986). *Introduction to modern information retrieval*. McGraw-Hill, New York.

Steyvers, M., Smyth, P., & Griffiths, T. (2004). Probabilistic author-topic models for information discovery. In *Proceedings of the 10th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD'04)* (pp. 306–315).

Tang, J., Hong, M., Li, J., & Liang, B. (2006). Tree-structured conditional random fields for semantic annotation. In *Proceedings of the 5th international semantic web conference (ISWC'06)* (pp. 640–653).

Tang, J., Zhang, D., & Yao, L. (2007). Social network extraction of academic researchers. In *Proceedings of 2007 IEEE international conference on data mining (ICDM'07)* (pp. 292–301).

Tang, J., Zhang, J., Yao, L., Li, J., Zhang, L., & Su, Z. (2008a). Arnetminer: extraction and mining of academic social networks. In *Proceedings of the 14th ACM SIGKDD international conference on knowledge discovery and data mining (SIGKDD'08)* (pp. 990–998).

Tang, J., Jin, R., & Zhang, J. (2008b). A topic modeling approach and its integration into the random walk framework for academic search. In *Proceedings of 2008 IEEE international conference on data mining (ICDM'08)* (pp. 1055–1060).

Tang, J., Yao, L., Zhang, D., & Zhang, J. (2010, to appear). A combination approach to web user profiling. *ACM Transactions on Knowledge Discovery from Data*.

Tsochantaridis, I., Hofmann, T., Joachims, T., & Altun, Y. (2004). Support vector machine learning for interdependent and structured output spaces. In *Proceedings of the twenty-first international conference on machine learning (ICML'04)* (pp. 823–830).

Wainwright, M. J., Jaakkola, T., & Willsky, A. S. (2001). Tree-based reparameterization for approximate estimation on loopy graphs. In *Proceedings of the 13th neural information processing systems (NIPS'01)* (pp. 1001–1008).

Xi, W., Zhang, B., Chen, Z., Lu, Y., Yan, S., Ma, W.-Y., & Fox, E. A. (2004). Link fusion: a unified link analysis framework for multi-type interrelated data objects. In *Proceedings of the 13th international conference on world wide web (WWW'04)* (pp. 319–327).

Xi, W., Fox, E. A., Fan, W., Zhang, B., Chen, Z., Yan, J., & Zhuang, D. (2005). Simfusion: measuring similarity using unified relationship matrix. In *Proceedings of the 28th ACM SIGIR international conference on information retrieval (SIGIR'2005)* (pp. 130–137).

Zhai, C., & Lafferty, J. (2001). A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th ACM SIGIR international conference on information retrieval (SIGIR'01)* (pp. 334–342).

Zhai, C., Velivelli, A., & Yu, B. (2004). A cross-collection mixture model for comparative text mining. In *KDD'04* (pp. 743–748).

Zhang, D., Tang, J., & Li, J. (2007a). A constraint-based probabilistic framework for name disambiguation. In *Proceedings of the 16th conference on information and knowledge management (CIKM'07)* (pp. 1019–1022).

Zhang, J., Ackerman, M. S., & Adamic, L. (2007b). Expertise networks in online communities: structure and algorithms. In *Proceedings of the 16th international conference on world wide web (WWW'07)* (pp. 221–230).