Inf Syst Front
DOI 10.1007/s10796-012-9385-6

Harnessing global expertise: A comparative study of expertise profiling methods for online communities

Xiaomo Liu · G. Alan Wang · Aditya Johri · Mi Zhou · Weiguo Fan

Abstract Building expertise profiles in global online communities is a critical step in leveraging the range of expertise available in the global knowledge economy. In this paper we introduce a three-stage framework that automatically generates expertise profiles of online community members. In the first two stages, document-topic relevance and user-document association are estimated for calculating users' expertise levels on individual topics. We empirically compare two state-of-the-art information retrieval techniques, the vector space model and the language model, with a Latent Dirichlet Allocation (LDA) based model for computing document-topic relevance as well as the direct and indirect association models for computing user-document association. In the third stage we test whether a filtering strategy can improve the performance of expert profiling. Our experimental results using two real datasets provide useful insights on how to select the best models for profiling users' expertise in online communities that can work across a range of global communities.

Keywords Expert finding · Online communities · Information retrieval · Global expertise

1 Introduction

One of the foremost advantages of globally distributed work is the opportunity to harness expertise across the world (Hinds and Kiesler 2002; Papazafeiropoulou 2004). Firms often cite improved access to expertise, a necessity in the knowledge economy, as the primary reason for internationalizing as going global allows them to hire the best resources wherever they are available. Furthermore, they are able to leverage the expertise across the globe by combining their human resources through the use of information technology. Scholars have documented the benefits of global expertise sharing since it brings diversity and therefore more and better ideas (Cummings 2004). But scholars have also found that utilizing global expertise is a complex and complicated task due to diversity, time zone differences, social and cultural differences, and linguistic differences (Hinds & Kiesler, 2002; Remus and Wiener, 2009). These barriers increase interpersonal and organizational breakdowns among teams (Majchrzak et al. 2000) and adversely affect knowledge sharing. A potential solution recommended by Kotlarsky and Oshri (2005) is to develop collective knowledge and transactive memory (i.e., a set of core knowledge possessed by group members coupled with an awareness of who knows what) that can lead to novel and productive work practices (Johri 2011). In this paper we test an instantiation of this suggestion by assessing the technical

X. Liu
Department of Computer Science, Virginia Tech,
Blacksburg, VA 24061, USA

X. Liu
e-mail: xiaomliu@vt.edu

G. A. Wang (✉)
Department of Business Information Technology, Virginia Tech,
Blacksburg, VA 24061, USA
e-mail: alanwang@vt.edu

A. Johri
Department of Engineering Education, Virginia Tech,
Blacksburg, VA 24061, USA

A. Johri
e-mail: ajohri@vt.edu

M. Zhou
School of Management, Xi'an Jiaotong University,
Xi'an, Shanxi 710049, People's Republic of China

M. Zhou
e-mail: zhoumiii@vt.edu

W. Fan
Department of Accounting and Information Systems,
Virginia Tech,
Blacksburg, VA 24061, USA

W. Fan
e-mail: wfan@vt.edu

Published online: 21 September 2012

feasibility of a range of approaches that can build automatic profiles of experts based on their contributions to online communities thereby assisting global expertise sharing.

With advances in information technology and its application across a range of organizations (Ackerman et al. 2003), online communities have emerged as an important mechanism for seeking professional expertise and collaboratively finding solutions to problems. Online communities have become a primary forum for exchange of experience and knowledge at a global scale as they extend across traditional national borders (Palvia 1998). One distinctive characteristic of online communities that makes collaboration on them different from traditional organizations is the easy accessibility of both social and technical cues present as a result of online interactions by users (Johri 2006). It is still an open question as how to effectively leverage the knowledge base and rich social capital in online communities for expertise profiling, particularly in a global context.

The rest of the paper is organized as follows. We first review knowledge management and social network literature related to knowledge sharing and expertise identification in online communities. Following an introduction to existing expert profiling methods, we propose an expert finding framework for online communities and identify alternative techniques for building expertise profiles within our framework. Empirical evaluations are used to compare the performance of alternative expertise profiling methods and make recommendations on the optimal approach. Finally, we discuss our findings, limitations, and future directions.

2 Related work

2.1 Knowledge management and social network theories related to online communities

Online communities are an instantiation of an electronic network of practice that consists of a large, loosely knit, collection of geographically distributed individuals engaged in a shared practice through computer-mediated communication (Wasko and Faraj 2005). Participants in online communities might not be personally acquainted, but they are still capable of sharing a great deal of knowledge (Brown and Duguid 2000, 2001). Members of these communities comprise of individuals who share common interests and voluntarily work together to expand their understanding of a knowledge domain through learning and sharing (Lin et al. 2009). Online communities rely on cooperating members as primary resources, who collaboratively share knowledge and help build a community knowledge repository. They provide a virtual media environment where individuals may seek and share knowledge across time and space. Similar to other asynchronous computer-medicated communication systems, online communities are rich in social media because social interactions are not limited by physical copresence and user interaction is continuously captured digitally.

Two widely adopted social network theories can be used to explain the motivation behind the voluntary knowledge sharing activities in online communities. The Social Capital Theory (Nahapiet and Ghoshal 1998), building on the work of Bourdieu (1990), defines social capital as the interpersonal relationships developed over time that facilitate coordination and collaboration for mutual benefit. It comprises of three dimensions including a structural dimension (overall pattern of connections between people), a relational dimension (assets such as trust created through relationships), and a cognitive dimension (shared language, culture, and norms). The theory suggests that social capital strongly influences the extent to which knowledge sharing occurs. The Social Cognitive Theory (Bandura 1986), on the other hand, considers human behavior as a triadic, dynamic, and reciprocal interaction of personal factors and the social network. Chiu, Hsu, and Wang (2006) provide empirical evidence in support of the two social network theories that both outcome expectations and social capital are helpful in explaining the motivation of knowledge sharing in virtual communities. A similar distinction has also been forwarded by Ren et al. (2007) who argue that users develop common identity and social bonds and these mechanisms serve as community building blocks. Therefore, the area of content and the availability of likeminded others and experts is critical for online communities. These two aspects, therefore, can also be leveraged to grow and sustain community participation.

2.2 Finding expertise in online communities

Expertise finding is an important knowledge management task in corporate, governmental and virtual organizations (Ardichvili et al. 2003; Thomas et al. 2001). Expert finding mainly has two aspects including expertise identification ("Who are the experts on Topic X?") and expertise selection ("What does Expert Y know?") (McDonald and Ackerman 2000). However, most of existing systems focus only on one of the two aspects although scholars recommend that both aspects needs to be taken into account (Pipek et al. 2012). Existing expert finding methods can be categorized into two categories: link-based and content-based methods. The link-based methods construct affiliation networks (Newman et al. 2002) based on user-to-user or user-to-document associations. Link-based algorithms such as PageRank, HITS and social network centrality measures can be used to calculate the social importance of a user based on his or her structural properties in the affiliation network (Campbell et al. 2003;

Zhang et al. 2007). However, the construction of an affiliation network is usually inefficient when the network is huge, which is always the case in online communities. Content-based expert finding methods identify the expertise of a candidate using the documents authored by the candidate. They consider the relevance between the authored documents and a topic as an expertise indicator. Traditional information retrieval methods, such as the vector space model (Demartini et al. 2009) and language model (Balog et al. 2009), often use terms to represent the contents of documents and to compute the relevance between a document and a search query. However, these terms sometimes cannot accurately capture the semantics of documents. Documents with similar semantic meanings but different term vocabulary will not be correctly associated using those methods. Recently, Latent Dirichlet Allocation (LDA) based expert finding techniques have been introduced to capture the semantic topics embedded in a collection of documents, also known as a corpus. They estimate the topic distributions of authors given their authored documents. These probabilities indicate the expertness of each candidate on each topic (Mimno and McCallum 2007; Steyvers et al. 2004). These expert finding techniques have also been shown effective in global online communities where languages other than English are used (e.g., Yang et al. 2008; Herzig and Taneva 2010).

3 Research goal

Automated expertise profiling techniques often employ a combination of information retrieval (IR), text mining, and data mining techniques to synthesize and build expert profiles using indicators such as authored documents and social interaction (Reichling et al. 2009). Although some techniques have been applied to online question-answering communities, their performance is unknown in the context of broader global online communities that attract participation from members across the globe. This study aims to propose an expertise-profiling framework for such communities. We identify alternative techniques and compare their performance using real online community datasets. Our framework focuses on both expertise identification and selection and will extract knowledge domains based on the semantic connections among the words in a text corpus. The framework consists of three stages including a document-topic relevance model, a user-document association model, and a filtering strategy. The findings of our study will provide practical guidance on how to identify expertise profiles in online communities. The study will also make a broad impact when being applied to expert finding in globally distributed collaborations.

4 An expert profiling framework for online communities

We first define the terms that will be used in our expert finding framework for online communities.

Definition 1. A *social corpus* in an online community is a document collection $D=\{d_1,\ldots, d_N\}$ that associates with a set of community members *$U=\{u_1,\ldots, u_M\}$. This relation is denoted by a matrix of user-document associations $A \in \mathbb{R}^{N\times M}$ where $A(d, u)$ quantifies the social tie between a member u and a document d.*

Definition 2. An *expertise profile* of a member u is defined as a vector of his/her expertness levels on all knowledge domains (i.e., topics) $Z=\{z_1,\ldots, z_T\}$ contained in an online community as profile(u)⇒<$K(z_1, u),\ldots, K(z_T, u)$>, where $K(z, u)$ represents the estimated expertise level of member u on domain z. All members' profiles are aggregated as a matrix of expertise scores $K \in \mathbb{R}^{T\times M}$.

Given above definitions, we propose an expert profiling framework for online communities. Figure 1 illustrates the major components of the proposed framework. The inputs to the framework include online posting documents (d) with authors (u) identified. Given the document corpus D, we can use the LDA model to identify a set of latent knowledge topic domains Z from the corpus. The rest of the framework computes an expertise level K for each member u on each topic domain z. The expertise level K is a mapping function $K=f(R, A)$, where $R(z, d)$ indicates the relevance of document d to topic domain z (determined by Stage 1) and $A(d, u)$ represents the association between a document d and a member u (determined by Stage 2). The filtering strategy is an attempt to remove irrelevant documents and improve the quality of document-topic relevance determination. In the rest of this section we introduce the LDA model for knowledge domain identification as well as the three major stages for computing the expertise level K for each member u on each topic domain z.

4.1 Identifying knowledge domains

The LDA model is a probability based generative model that considers each document as a mixture of a small number of topics (Blei et al. 2003; Griffiths and Steyvers 2004). In general, the LDA model measures the probability of generating a document d with S_d words $\{w_1,\ldots,w_{S_d}\}$ by a mixture of topics $\{z_1,\ldots,z_T\}$

$$p(w_1,\ldots,w_{S_d}) = \prod_{i=1}^{S_d} \sum_{j=1}^{T} p(w_i|z_j)p(z_j|d),$$

where $p(w_i \mid z_j)=\phi^{(j)}$ refers to a multinomial distribution of word w_i over a topic z_j with a Dirichlet prior α, and $p(z_j \mid$

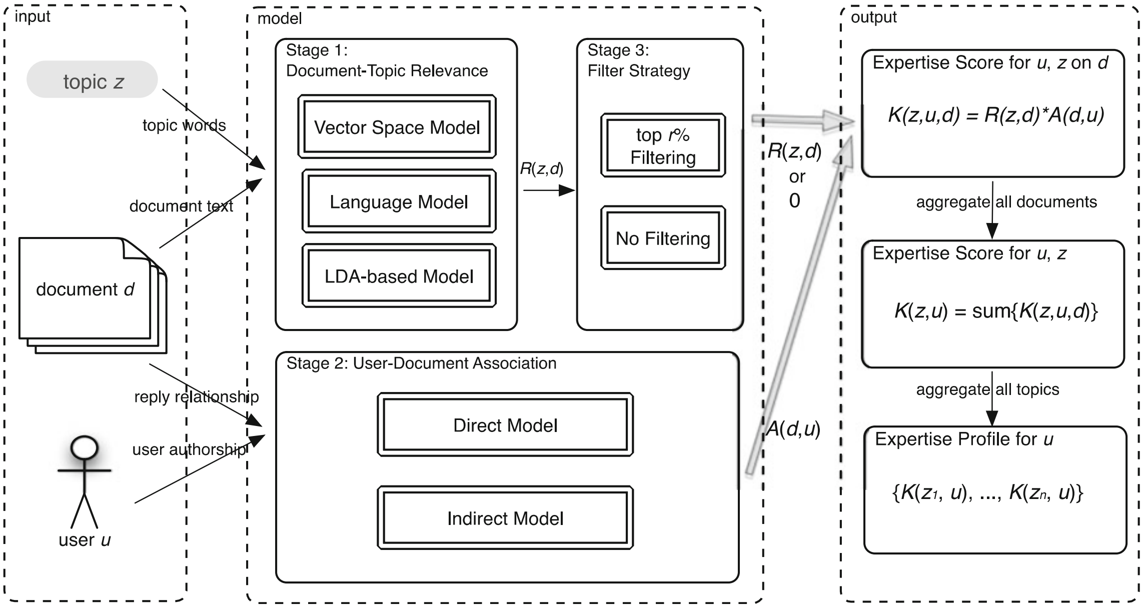


Fig. 1 An expertise profiling framework for online communities

d)=$\theta^{(d)}$ is another multinomial distribution of a topic z_j over a document d with a Dirichlet prior β. The statistical model is conditioned on three parameters including the Dirichlet hyperparameters α and β and the number of topics T. Griffiths and Steyvers (2004) recommend that β is set to 0.1 and α=50/T, resulting a fine-grained decomposition of the corpus into topics. Due to the space limitation, please refer to their work for the discussion of the two parameters. We now still need to determine the number of knowledge domains. Following the procedure described in (Blei et al. 2003; Heinrich 2005), we train our LDA model with different settings of the number of topics using training data D_{train}. The trained model that fits held-out test data D_{test} the best indicates the optimal setting for the topic number. Perplexity is a commonly used measure for measuring the likelihood of the held out test data given an LDA model. It is defined as

$$\text{perplexity}(D_{test}|T) = \exp\left(-\frac{\sum_{d=1}^{|D_{test}|}\log p(w_1,\ldots,w_{S_d})}{\sum_{d=1}^{|D_{test}|}S_d}\right).$$

The measure decreases monotonically in the likelihood of the test data. Thus, the optimal number of topics T in a corpus is determined by:

$$T = \arg\min_{T}\{\text{perplexity}(D_{test}|T)\}.$$

4.2 Document-topic relevance

There are different techniques for calculating the relevance score between a topic and a document. If we view each topic z as a query q_z of n words, the relevance score between a topic and a document indicates the likelihood of a document containing a certain topic. The expertise level of a candidate with regard to a topic is the aggregation of all the relevance scores of the candidate's associated documents. Two IR models, namely the Vector Space Model and language model, and a modified LDA–based model can be used to estimate the document-topic relevance.

The Vector Space Model (VSM) is widely used for measuring the relevance between a document and a query. In this model each document d is represented as a weight vector $\vec{d}$, in which each term weight $\omega_{i,d}$ is computed based on the TF-IDF scheme, i.e., $\omega_{i,d}$=TF$_{i,d}$×IDF_i (Lee and Seamons 1997). TF is term frequency calculated as the normalized occurrence count $c_{i,d}$ of word w_i in document d. IDF is inverse document frequency, a measure of the general importance of a term. They are defined as

$$TF_{i,d} = \frac{c_{i,d,}}{\sum_k c_{k,d}} \quad \text{and } IDF_i = \log\left(\frac{|D)}{1+\left|\{j{:}w_i{\in}d_j\}\right|}\right).$$

Similarly, each topic z can be represented as a weight vector $\vec{q_z}$, in which each element is the probability of a word belonging to that topic, $\omega_{i,z} = p(w_i|z)$. Thus, the relevance between d and z can be calculated as the cosine similarity of the two vectors such as

$$R_{vector}(z,d) \Leftrightarrow sim(z,d) = \frac{\vec{q_z}\cdot\vec{d}}{\|\vec{q_z}\|\cdot\|\vec{d}\|} = \frac{\sum_{i=1}^{|q_z|}\omega_{i,z}\cdot_{i,d}}{\sqrt{\sum_{i=1}^{|q_z|}\omega_{i,z}^z}\cdot\sqrt{\sum_{i=1}^{|d|}\omega_{i,d}^z}}.$$

Using VSM the expertise level of a member u on a topic z is the accumulative topic relevance of all the documents associated with u,

$$K_{vector}(z,u) = \sum_d R_{vector}(z,d)A(d,u).$$

The drawback of this model is that it does not take into account the semantic relationships between words. Thus, documents with similar meanings but different vocabularies may not result in a good match.

In the expert finding technique proposed by Balog et al. (2009), a statistical language model is used to estimate the probability $p(u|q_z)$ of u being an expert given a query q_z. Using the Bayes' theorem, this probability is indirectly calculated from $p(q_z|u)$:

$$p(u|q_z) = \frac{p(q_z|u)p(u)}{p(q_z)} \mu \; p(q_z|u).$$

Assuming that both $p(u)$ and $p(q_z)$ are constant probabilities, they can be ignored here. $p(q_z|u)$ is then obtained by simply taking the integral over all associated documents,

$$p(q_z|u) = \sum_d p(q_z|d)p(d|u).$$

If we consider R_{lang} $(z,d) \Leftrightarrow p(q_z|d)$ and $A(d,u) \Leftrightarrow p(d|u)$, the expertise score under this model can be calculated in the same form as the VSM,

$$K_{lang}(z,u) = \sum_d R_{lang}(z,d)A(d,u).$$

In the language model, the probability $p(q_z|d)$ of document d generating query q_z is computed using each word w_i of q_z individually Assuming that words are independent to each other, the probability can be computed as

$$p(q_z|d) = \prod_{i=1}^{|q_z|} p(w_i|d).$$

Thus, the computation is reduced to calculating the probability of document d generating word w_i. It is usually approximated using the maximum likelihood estimation (MLE),

$$p(w_i|d) = (1-\lambda)\widehat{p}_{mle}(w_i|d) + \lambda\widehat{p}_{mle}(w_i|D),$$

where $\widehat{p}_{mle}(w_i|d)$ is the probability of word w_i generated by document d and $\widehat{p}_{mle}(w_i|D)$ is the probability of word w_i generated by the entire corpus D . They are smoothed using the Jelink-Mercer method with a coefficient λ (Zhai and Lafferty 2004). Although this model has been verified to be effective on TREC enterprise corpora, the performance of its application to online communities remains to be tested.

The LDA model that we introduced earlier has a built-in mechanism to capture document-topic relevance. It learns the distributions $p_{\text{lda}}(z|d)$ of documents over topics. The expertise level of a candidate person u on a topic z is calculated indirectly through all documents associated with the candidate,

$$\begin{aligned} K_{lda}(z,u) &= \sum_d p_{lda}(z|d)p(d|u) \\ &= \sum_d p_{lda}(z,d)A(d,u) \end{aligned}.$$

Although directly solving $p_{\text{lda}}(z|d)$ is intractable, it can be approximated effectively using numerical algorithms such as Gibbs Sampling (GS)

$$R_{lda}(z,d) \Leftrightarrow p\left(z|d,\widehat{\phi}\right) = \frac{n_{d,z}+\alpha}{\sum_z \left(n_{d,z}+\alpha\right)},$$

where $n_{d,z}$ is the number of times a word in document d assigned to topic z during the sampling process (Griffiths and Steyvers 2004). The LDA based method has the ability to match documents with similar semantic topics. Hence, we expect this model to perform better than VSM and the language model.

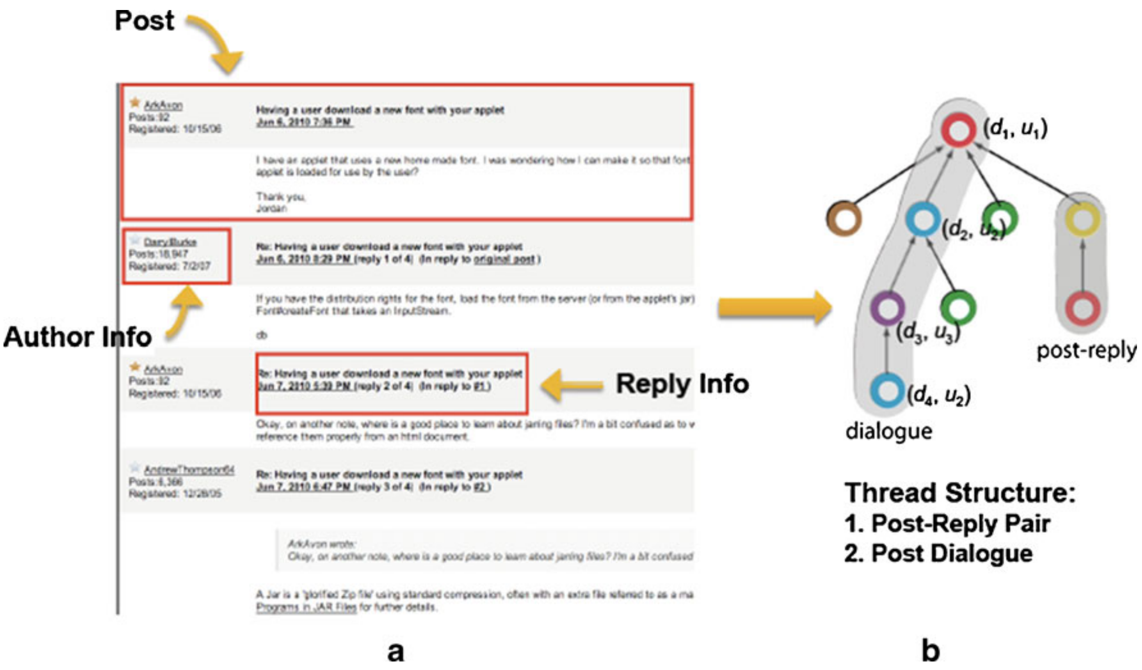
4.3 User-document association

With the relevance between each document and each topic domain known, we still need to know the user-document association before we can associate each user to each topic domain through his or her associated documents. Two user-document association models that can be used to determine $A(d, u)$, the strength of a social tie between document d and person u. Such a tie can be defined as a probability $p(d|u)$ of document d being associated with person u.

A direct association model calculates user-document association as a probability $p(d|u)$ using authorship. Unlike research papers or news articles that may have multiple authors, an online posting normally has a single author. We can build a binary association between online users and their authored documents as

$$A_{\text{direct}}(d,u) \Leftrightarrow p(d|u) = \begin{cases} 1 & \text{if } u \text{ wrote } d \\ 0 & \text{otherwise} \end{cases}.$$

An indirect association model finds the documents that are not authored by an expert candidate but are related. A typical discussion thread in an online community consists of an original question posting and possibly several replies. Postings in the discussion are relevant not only to their authors, but also other participants in the same discussion. A recent study showed that documents indirectly connecting to a person can propagate relevance and be used to help assess one's expertise (Griffiths and Steyvers 2004). Figure 2 shows an online discussion thread example involving four postings (i.e., documents) and three users $\{(d_4, u_2)\rightarrow(d_3, u_3)\rightarrow(d_2, u_2)\rightarrow(d_1, u_1)\}$. The closest distance l between u_2



ArkAxon
Posts:92
Registered: 10/15/06

Having a user download a new font with your applet
Jun 6, 2010 7:36 PM

I have an applet that uses a new home made font. I was wondering how I can make it so that font
applet is loaded for use by the user?

Thank you,
Jordan

DarryBurke
Posts:18,947
Registered: 7/2/07

Re: Having a user download a new font with your applet
Jun 6, 2010 8:29 PM (reply 1 of 4) (In reply to original post)

If you have the distribution rights for the font, load the font from the server (or from the applet's jar)
Font#createFont that takes an InputStream.

db

ArkAxon
Posts:92
Registered: 10/15/06

Re: Having a user download a new font with your applet
Jun 7, 2010 5:39 PM (reply 2 of 4) (In reply to #1)

Okay, on another note, where is a good place to learn about jaring files? I'm a bit confused as to v
reference them properly from an html document.

AndrewThompson64
Posts:6,366
Registered: 12/28/05

Re: Having a user download a new font with your applet
Jun 7, 2010 6:47 PM (reply 3 of 4) (In reply to #2)

> *ArkAvon wrote:*
> *Okay, on another note, where is a good place to learn about jaring files? I'm a bit confused*

A Jar is a 'glorified Zip file' using standard compression, often with an extra file refered to as a ma
Programs in JAR Files for further details.

Fig. 2 A thread example in an online community **a** A typical discussion thread with multiple posts; **b** The thread structure derived from the post-reply relationships (*circles* represent posts *d* and *colors* represent authors *u*)

and d_1 is one step. We revise the direct user-document association model in order to consider indirect relevance propagation,

$$A_{indirect}(d,u) \Leftrightarrow p(d|u)$$
$$= \begin{cases} 1 & \textit{if } u \text{ wrote } d \\ (1-\sigma)^l & \text{if } u \text{ is } l \text{ steps from } d\,. \\ 0 & \text{otherwise} \end{cases}$$

where σ is a decay factor representing the strength of the propagated association. We arbitrarily set σ=0.5 as a compromise between strongly and weakly indirect association.

4.4 Filtering strategy

The expertise score$K(z,u) = \sum_d R(z,d)A(d,u)$ is the sum of the product of document-topic relevance and user-document association over all the authored documents of a candidate. When determining the expertise level of a candidate on a topic, the document-topic relevance is very important because only those documents highly relevant to the topic are indicative to the author's expertise on this topic. Therefore, we propose to keep the top *r*% most relevant documents when computing the document-topic relevance. Assuming that relevant documents are equally distributed over *T* topics, we only keep the top 1/*T* documents relevant to each topic.

5 Empirical evaluations

Our expert finding framework consists of three input stages: document-topic relevance, user-document association, and a filtering strategy. Using real online community datasets, we conducted empirical evaluations to study the effect of alternative techniques with regard to the performance of expert profiling in online communities. Table 1 summarizes the inputs and outputs of all candidate models used in our comparative study.

5.1 Data collection and knowledge domain identification

We collected data from two popular online communities: *Sun forums* (now *Oracle forum*s) and *Apple Discussions*. We selected the largest sub-forum from each community, i.e., the *Java Programming* sub-forum from *Sun forums* and the *Phone and Messaging (iphone)* sub-forum from *Apple discussions*. We crawled all discussion threads in the two sub-forums published as of Jun 6, 2009 for *Java* and April 1, 2010 for *iPhone*. The characteristics of these two datasets are reported in Table 2. Our data pre-processing steps include term indexing and stop-word removal.

We first identified knowledge domains that existed in the two datasets separately. To learn the best LDA model, we used a 10-fold cross-validation approach (McLachlan et al. 2004). The discussion threads in each dataset were randomly partitioned into 10 folds. Of the 10 folds, a single fold was retained for testing while the other 9 folds were used as

Table 1 Candidate models for expertise finding in online communities

Input & Output	Model	Algorithm
Input stage 1:	Vector Space Model (R_{vector})	$R_{vector}(z,d) \Leftrightarrow sim(z,d) = \frac{\overrightarrow{q}_z \cdot \overrightarrow{d}}{\left\Vert \overrightarrow{q}_z \right\Vert \cdot \left\Vert \overrightarrow{d} \right\Vert} = \frac{\sum_{i=1}^{\lvert qz \rvert} \omega_{i,z} \cdot \omega_{i,d}}{\sqrt{\sum_{i=1}^{\lvert qz \rvert} \omega_{i,z}^2 \cdot \sqrt{\sum_{i=1}^{\lvert d \rvert} \omega_{i,d}^2}}}$
Document-Topic Relevance (R)	Language Model(R_{lang})	$R_{lang}(z,d) \Leftrightarrow p(q_z \vert d) = \prod_{i=1}^{\lvert q_z \rvert} p(w_i \vert d)$
	LDA-based Model (R_{lda})	$R_{lda}(z,d) \Leftrightarrow p\left(z \vert d, \widehat{\phi}\right) = \frac{n_{d,z}+\alpha}{\sum_z \left(n_{d,z}+\alpha\right)}$
Input stage 2:	Direct Association Model (A_{direct})	$A_{\text{direct}}(d,u) \Leftrightarrow p(d \vert u) = \begin{cases} 1 & \text{if } u \text{ wrote } d \\ 0 & \text{otherwise} \end{cases}$
User-Document Association (A)	Indirect Association Model (A_{indirect})	$A_{indirect}(d,u) \Leftrightarrow p(d \vert u) = \begin{cases} 1 & \text{if } u \text{ wrote } d \\ (1-\sigma)^l & \text{if } u \text{ is } l \text{ steps from } d \\ 0 & \text{otherwise} \end{cases}$
Input stage 3:	Without Filtering	Keep all documents for calculating K
Filter Strategies	With Filtering	Keep top r% documents according to each R model
Output: Domain Expertise (K)	Mapping Function K=f(R, A)	$K(z,u) = \sum_d R(z,d)A(d,u)$

training data. Using the commonly used parameters α=50/T and β=0.1 recommended in previous studies (Griffiths and Steyvers 2004; Wei and Croft 2006), we discovered 270 topics in the *Java* dataset and 200 topics in the *iPhone* dataset. We illustrated 5 randomly selected topics with their manually labeled domains as well as the top 10 keywords in Table 3.

5.2 Experimental design

To get a gold standard for our datasets in order to evaluate the performance of our expert finding framework, we recruited four computer science graduate students to manually annotate the domain expertise of expert candidates using a 5-rating scale (Zhang et al. 2007). In addition, we added one additional rating level 0 due to the fact that there may not be adequate information for evaluation. Table 4 summarizes the rating scale.

To reduce the burden on annotators in rating expertise level, we used a sampling strategy to randomly select a small number of members, topics, and threads so that an efficient evaluation could be done without losing generalizability. For each corpus we randomly selected 5 topics. We then used a pooling method (Griffiths and Steyvers 2004) to find the top 200 most relevant documents given each selected topic. The pooled documents were merged to a sample corpus. We randomly selected 30 members among those who appeared in the sample corpus and produced at least 10 posts. The minimum activity requirement provided annotators adequate information to evaluate a member. Our final sample contained 150 member-topic records. We divided the four graduate students into two-person groups. Within each group the two students rated each sampled member individually after reading his/her participating threads. If there was discrepancy in the ratings, we asked them to reconcile between them first and then took the average if reconciliation failed. Finally, we calculated Cohen's Kappa Coefficient (k) and Pearson's correlation coefficient (ρ) to assess the inter-coder agreement. As Table 5 shows, the average k and ρ were above 0.75 and 0.9, respectively, and had satisfactory agreement (Fleiss 1981).

5.3 Performance metrics

We considered three use cases of expertise profiling in online communities: (1) topic-centric profiling seeks a

Table 2 Data characteristics

Dataset	# of threads	# of posts	# of members	Avg. replies per thread	Avg. # of repliers per thread	# of words in dataset	Size of vocabulary
Java	70,488	440,708	36,687	4.1723	2.5137	7,374,557	125,114
iPhone	49,343	271,823	55,108	4.1306	3.1543	4,962,015	56,157

Table 3 Sample topics

Java					***iPhone***				
Topic 15 GUI with swing	Topic 37 Thread synchronization	Topic 83 Parsing xml files	Topic 106 Multimedia: video & audio	Topic 109 Querying database	Topic 28 Using Google map	Topic 32 Screen protection	Topic 66 Keyboard & word type	Topic 165 Sound issue	Topic 168 Phone plan
swing	thread	xml	sound	database	Google	screen	type	volume	Plan
component	wait	parse	play	connect	map	scratch	word	speaker	data
frame	synchronize	document	video	sql	direction	glass	keyboard	Hear	unlimited
jframe	lock	dom	media	jdbc	search	clean	correct	sound	monthly
panel	timer	element	capture	driver	view	water	language	low	family
layout	run	transform	audio	oracle	location	protector	suggest	lound	pay
label	execute	sax	jmf	mysql	street	plastic	English	ear	at&t
container	sleep	tag	demo	db	traffic	film	learn	lounder	cost
pane	concurrent	xerc	music	query	route	cloth	letter	issue	include
manager	notify	schema	song	pool	pin	cover	dictionary	echo	line

ranked list of experts given a knowledge domain. This scenario is useful for those online community participants who cannot find answers to their questions and would consult an expert in the related areas; (2) profile-centric profiling retrieves a ranked list of knowledge domains given a member. This scenario provides a complete expertise profile for a given member. It will benefit participants who would know more about a member before making a contact; (3) community-centric profiling produces a ranked list of topic-member combinations based on the expertise level. This scenario is useful for the administrators of online communities who get to know the top knowledge domains and associated expert members within the online community. We used the following metrics to evaluate the effectiveness of our expert finding framework in online communities.

Rank correlation. We converted the human annotators' expertise ratings and automatically calculated expertise scores into two separate ranked lists. Rank correlation uses statistical methods to measure the degree to which two ranked lists are correlated. The correlation signifies the tendency of the values of one ranking to be in the same order of the values of the other ranking (Melucci 2007). In this study we chose two commonly used statistics, Spearman's ρ and Kendall's τ (Zar 2009), to measure the correlation between the ranking provided by the human annotators and that calculated by our expert ranking algorithm.

Table 4 Domain expertise rating scales

Rate	Category	Description
5	Expert	Knows core knowledge on a domain and always provides insightful answers
4	Professional	Can answer most questions and knows some sub-topics of a domain very well
3	User	Has some advanced knowledge and can give relatively good answers
2	Learner	Knows basic concepts, but is not good at advanced knowledge in a domain
1	Newbie	Starts to learn the knowledge of a domain, asks questions on basic concepts
0	N/A	There is not enough information (i.e. no post) on a knowledge domain

Graded relevance. Rank correlation compares the ranking orders of the entire lists. However, it does not answer the question if real domain experts are actually ranked higher on the lists. Normalized discounted cumulative gain (nDCG) is commonly used in information retrieval to measure the gain of a ranked list based on the position of the ranked elements (Cormack et al. 1998). The nDCG value at a particular rank position p is calculated as:

$$nDCG_p = \frac{DCG_p}{IDCG_p} = \frac{rate'_1 + \sum_{i=2}^{p} rate'_i/\log_2 i}{rate_i + \sum_{i=2}^{p} rate_i/\log_2 i}$$

Where rate$'_i$ is the calculated rating at the i^{th} position ranked by a candidate model and rate$_i$ is the rating ranked by the gold standard.

5.4 Experimental results

We first studied the performances of three document-topic relevance models, namely the VSM (R_{vector}), language model (R_{lang}) and LDA-based model (R_{lda}) along with the direct user-document association model. Filtering strategy was not applied at this moment. The top 10 keywords in each topic were used as a query for the VSM and language model.[1] We observed that the LDA-based model outperformed both the

[1] We experiment various values for n words. Our results show that top-10 is a near optimal choice.

Table 5 Inter-coder agreement

Java			iPhone		
	k	ρ		*k*	ρ
Topic 15	0.79	0.94	Topic 28	0.8	0.96
Topic 37	0.76	0.92	Topic 32	0.76	0.91
Topic 83	0.73	0.94	Topic 66	0.79	0.94
Topic 106	0.86	0.94	Topic 165	0.74	0.94
Topic 109	0.74	0.95	Topic 168	0.78	0.95
All Topics	0.77	0.95	All Topics	0.78	0.94

VSM and language model on nearly all metrics (Table 6), the only exception being the topic-centric profiling task in the *Java* corpus. The rank correlations of the LDA-based model indicate good ranking performance in the *iPhone* corpus (when ρ≈0.8 and τ≈0.7) along with a less satisfactory ranking performance in the *Java* corpus (when ρ≈0.7 and τ≈0.6). The nDCG values of the LDA-based model were near perfect for both corpora. The result shows that the LDA-based model performed the best comparing to the VSM and language model. After examining what caused ranking errors in the VSM and language models, we found that these two models tended to favor short documents over long ones when determining document-topic relevance. Both models calculate document relevance using term frequencies normalized by the document length. Two documents with very similar content can have a significant difference in relevance calculation simply because one is much longer than the other. Those users who frequently made very short messages would receive inflated expertise scores using the VSM and language models. The LDA-based model does not consider document length. Thus, the document length would not have an effect in its relevance computation. We also noticed that the LDA-based model and VSM had comparable performance for the topic-centric task on the Java data set while the LDA-based model outperformed VSM for the same task on the IPhone data set. The performance difference across the two data sets was caused by the difference in document length in the two data sets. We observed that the average number of words in a discussion thread in the Java data set is notably less than that in the IPhone data set. Therefore, the VSM is less likely to produce biased relevance judgments for long documents in the Java data set.

To achieve statistical significance in performance comparison we applied Fisher's *z*-transformation to compare rank correlations (Zar 2009) and conducted paired *t*-tests for nDCG. The *p*-values in Table 7 illustrate that the LDA-based model significantly outperformed the language model in all three use-cases while significantly outperformed the

Table 6 The performance comparison of the three *R* models (best results in bold)

	Java					*iPhone*				
	Topic-centric									
	ρ	τ	nDCG$_5$	nDCG$_{10}$	nDCG$_{15}$	ρ	τ	nDCG$_5$	nDCG$_{10}$	nDCG$_{15}$
***R*$_{\text{vector}}$**	0.6971	**0.5741**	**0.930**	**0.9430**	0.9376	0.7700	0.6391	0.9481	0.9438	0.9530
***R*$_{\text{lang}}$**	0.5506	0.4232	0.8430	0.8774	0.8898	0.5294	0.4741	0.7454	0.8299	0.8560
***R*$_{\text{lda}}$**	**0.7077**	0.5628	0.9282	0.9285	**0.9430**	**0.8440**	**0.7325**	**0.9842**	**0.9733**	**0.9804**
	Profile-centric									
	nDCG$_1^u$	*nDCG*$_2^u$	*nDCG*$_3^u$	*nDCG*$_4^u$	*nDCG*$_5^u$	*nDCG*$_1^u$	*nDCG*$_2^u$	*nDCG*$_3^u$	*nDCG*$_4^u$	*nDCG*$_5^u$
***R*$_{\text{vector}}$**	0.8477	0.9197	0.9346	0.9541	0.9666	0.9658	0.9407	0.9552	0.9624	0.9750
***R*$_{\text{lang}}$**	0.8488	0.8965	0.9010	0.8949	0.9471	0.8706	0.8949	0.9165	0.9408	0.9553
***R*$_{\text{lda}}$**	**0.9681**	**0.9610**	**0.9667**	**0.9751**	**0.9830**	**0.9833**	**0.9415**	**0.9600**	**0.9724**	**0.9773**
	Community-centric									
	ρ	τ	nDCG$_{10}$	nDCG$_{30}$	nDCG$_{50}$	ρ	τ	nDCG$_{10}$	nDCG$_{30}$	nDCG$_{50}$
***R*$_{\text{vector}}$**	0.7020	0.5335	0.9592	0.9298	0.9302	0.7963	0.6383	0.8985	0.9143	0.9236
***R*$_{\text{lang}}$**	0.5156	0.3797	0.8620	0.8864	0.8572	0.5803	0.4422	0.7850	0.7857	0.8248
***R*$_{\text{lda}}$**	**0.7406**	**0.5715**	**0.9702**	**0.9572**	**0.9381**	**0.8436**	**0.7015**	**0.9543**	**0.9633**	**0.9617**

ρ and τ do not apply to the profile-centric task because the rank lists are too short to achieve statistical significance

Table 7 Significance testing for R model comparisons (p-values)

	Java Corpus					*iPhone* Corpus				
	$nDCG^z$	$nDCG^u$	ρ	τ	nDCG	$nDCG^z$	$nDCG^u$	ρ	τ	nDCG
R_{vector} *vs* R_{lda}	0.6998	**0.0095**	0.2045	0.3315	**0.0019**	**<0.0001**	0.3661	0.1240	0.1801	**<0.0001**
R_{lang} *vs* R_{lda}	**<0.0001**	**0.0061**	**0.0065**	0.0741	**<0.0001**	**<0.0001**	**0.0070**	**<0.0001**	**0.0016**	**<0.0001**

Bold p-values indicate statistically significant difference in performance comparison

VSM in community-centric profiling. Therefore, we chose the LDA-based model as the optimal document-topic relevance model in our framework.

While using the LDA-based document-topic relevance model, we compared the effect of the two user-document association models on the performance of expert profiling. Table 8 shows the effect of each document-topic relevance model on the performance of expert profiling in online communities. The direct association model outperformed the indirect model in most of the performance measures. In the cases where it did not, the differences were not statistically significant. The direct association model performed consistently better across the three tasks than the indirect association model on the Java data set. It had better performance in terms of nDCG across the three tasks on the IPhone data set. However, it only outperformed the indirect association model in the profile-centric task in terms of ρ and τ. The performance differences on the IPhone data set were not statistically significant. It is partly because the Java forum has a much more diverse range of topics than the IPhone forum. On the Java data set, the indirect association model may incorrectly calculate a member's expertise score on one topic using the documents that belong to very different topics. The performance differences between the two models were less obvious on the IPhone data set because the topics in the IPhone data set were more focused and related to each other. It is interesting to learn that indirectly associated documents do not help in LDA-based expert finding, although they were reported to be beneficial for the language model (Serdyukov et al. 2008).

We also tested whether or not our filtering strategy improved the performance of expertise profiling in online communities when the LDA-based document-topic relevance and direct user-document association were applied. Table 9 summarizes the performance differences with the number of retaining relevant document varying between 5 % and 50 %. Compared with the baseline results where no filtering was applied, expert profiling with filtering had a significant performance increase when the top 15–20 % most relevant documents were used. The percentage is consistent with our assumption that relevant documents are evenly distributed across all topics (r=1/5=20 %). When the filtering was more or less restrictive than that with the optimal document-topic relevance threshold, most of the performance measures got worse. When the filtering was less restrictive, less relevant documents may have introduced noise in expertise assessment. When it was more restrictive, some relevant documents helpful in expertise assessment may have been removed.

Table 8 The performance comparison of the two user-document association models

	Java Corpus					*iPhone* Coprus				
	ρ	τ	$nDCG_5$	$nDCG_{10}$	$nDCG_{15}$	ρ	τ	$nDCG_5$	$nDCG_{10}$	$nDCG_{15}$
	Topic-centric									
A_{direct}	**0.7077**	**0.5628***	**0.9282**	**0.9285**	**0.9430**	0.8440	0.7325	**0.9842**	**0.9438**	**0.9530**
$A_{indirect}$	0.6734	0.5321	0.8988	0.9127	0.9296	**0.8559**	**0.7391**	0.9807	0.9418	0.9512
	Profile-centric									
A_{direct}	**0.9681**	**0.9610**	**0.9667**	**0.9751**	**0.9830**	**0.9833**	**0.9415**	**0.9600**	**0.9724**	**0.9773**
$A_{indirect}$	0.9583	0.9602	0.9649	0.9725	0.9817	0.9821	0.9403	0.9500	0.9696	0.9757
	Community-centric									
A_{direct}	**0.7406**	**0.5715***	**0.9702***	**0.9572***	**0.9381**	0.8436	0.7012	**0.9543**	**0.9633**	**0.9617**
$A_{indirect}$	0.7094	0.5444	0.9227	0.9199	0.9232	**0.8507**	**0.7045**	0.9540	0.9609	0.9558

Bold numbers indicate the best performance between the two user-document association models

* p-value <0.05

Table 9 Testing of the filtering strategy

	Java			*iPhone*		
	Topic-centric nDCG$_5$/ρ/τ	Profile-centric nDCG$_5$	Community-centric nDCG$_{10}$/ρ/τ	Topic-centric nDCG$_5$/ρ/τ	Profile-centric nDCG$_5$	Community-centric nDCG$_{10}$/ρ/τ
No filtering	0.928/0.708/0.563	0.983	0.970/0.741/0.572	0.974/0.844/0.733	0.977	0.954/0.844/0.702
r=top 50 %	0.963/0.750/0.607	0.983	0.971/0.782/0.615	0.975/0.862/0.748	0.977	0.968/0.857/0.719
r=top 25 %	0.959/0.782/0.643	0.988	0.974/0.809/0.645	0.981/0.866/0.754	0.977	0.968/0.860/0.722
r=top 20 %	0.962/0.792/**0.657***	0.988	0.974/**0.825***/**0.667**	0.985/**0.872**/**0.759**	0.977	0.968/**0.863**/**0.727***
r=top 15 %	**0.964****/**0.796***/0.654	**0.990**	0.974/0.824/**0.667****	**0.992***/0.866/0.750	**0.978**	0.968/0.861/0.722
r=top 10 %	0.963/0.787/0.653	0.979	**0.981***/0.824/0.662	0.989/0.831/0.718	0.974	0.968/0.822/0.686
r=top 5 %	0.945/0.737/0.621	0.956	0.974/0.756/0.620	0.984/0.731/0.634	0.922	**0.970***/0.698/0.577

Bold numbers indicate the best performance for each measure across different filtering strategies

* *p*-value <0.05, ** *p*-value <0.01

6 Conclusions and future directions

Global online communities are increasingly become a central part of the global expertise ecosystem. In this paper we presented results from a comparative study of approaches that can be used to generate expertise profiles from data available on online communities and aid in expertise sharing. Both the communities whose data we examined attract visitors from across the globe. We first introduced a framework that builds profiles based on each person's domains (i.e., topics) expertise and then evaluated the expertise using a three-stage procedure: (1) estimating document topic relevance (*R*); (2) building user-document association (*A*); and (3) applying a filtering strategy to remove irrelevant documents. Based on this framework, we provided alternative models that can be used at each stage.

In our experiments we evaluated three *R* models, two *A* models and a filtering strategy using two real online community datasets. To the best of our knowledge, this is the first comparative study on expertise profiling methods in online communities and can be applied across a range of global communities. Our experimental results assess the appropriateness of a model at each stage. The LDA model outperformed the VSM and language model for evaluating document topic relevance. It is preferable to choose direct model for building the user-document association since the indirect model impaired the LDA model to achieve the best performances. Finally, the filtering strategy improved the performance of the LDA model.

One limitation of this study is its implicit assumption that highly relevant documents truly reflect the expertise level of the contributor. However, it is likely that two contributors having the same expertise level may receive different ratings under our expertise-profiling framework due to the difference in their responsiveness. Although responsiveness has been found directly related to the experts' accessibility (McDonald and Ackerman 1998; 2000), it should be considered under a new dimension separated from the expertise level in future studies.

We believe that our results are promising in identifying the optimal expert profiling method in global online communities that have extensive variations in their participation. Although our empirical evaluations were conducted using English-based online communities developed in the United States, we believe our findings are still applicable to global online communities developed in other countries. With the development of machine translation services such as the Google Translate, documents written in different languages can be easily translated into one language such as English before our expert finding approach can be applied. However, our expert finding approach is built on the assumption that one's expertise can be revealed through the contents of the authored documents and social interactions with other members of the community. The cultural factors such as degree of collectivism and competitiveness among members may influence the knowledge sharing strategies of participants in different countries (Ardichvili et al. 2006). The practitioners of global online communities should be aware of the cultural barriers to knowledge sharing that may exist in a cultural group. They could build preventive measures in the design of their online community systems in order to motivate their community numbers. We hope that future studies can follow the framework and continue this line of research on building user expertise profiles in other virtual collaboration environment.

Our work may also be extended beyond online communities to other scenarios where the relation of document authors and topic domains embedded in the documents is important. One such example is electronic discovery. The litigation process of a lawsuit often needs to process a large amount of electronic documents such as email, instant messages, electronic documents, and databases. Electronic discovery refers to that process that reviews and finds evidence relevant to the lawsuit from electronically stored information (Holley et al. 2010). Our proposed framework can be used to find evidence domains embedded in the electronic documents with authors (e.g., email, instant messages) and authors who are highly associated with each evidence domain. Another domain in which this method might be applicable is "knowledge discovery" where researchers and policy makers need to build actionable knowledge from a large data corpus. It is critical in such endeavors to be able to associate ideas with experts and to understand how different experts contribute to a core set of ideas. We believe that our work has great potential in e-discovery as well as other applicable domains.

Acknowledgment This work was partially funded by the U.S. National Science Foundation (Grant #0935124, #1122609, and #0840719) and the National Science Foundation of China (Grant #70872089 and #71072129). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the funding agencies.

References

Ackerman, M., Pipek, V., & Wulf, V. (2003). *Sharing Expertise: Beyond Knowledge Management*. Cambridge: MIT Press.

Ardichvili, A., Page, V., & Wentling, T. (2003). Motivation and barriers to participation in virtual knowledge-sharing communities of practice. *Journal of Knowledge Management, 7*(1), 64–77.

Ardichvili, A., Maurer, M., Li, W., Wentling, T., & Stuedemann, R. (2006). Cultural influences on knowledge sharing through online communities of practice. *Journal of Knowledge Management, 10* (1), 94–107.

Balog, K., Azzopardi, L., & Rijke, M. D. (2009). A language modeling framework for expert finding. *Information Processing and Management, 45*(1), 1–19.

Bandura, A. (1986). *Social Foundations of Thought and Action: A Social Cognitive Theory*. Englewood-Cliffs: Prentice Hall.

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

Bourdieu, P. (1990). *The Logic of Practice*. Stanford: Stanford University Press.

Brown, J. S., & Duguid, P. (2000). *The Social Life of Information*. Boston: Harvard Business School Press.

Brown, J. S., & Duguid, P. (2001). Knowledge and organization: a social-practice perspective. *Organization Science, 12*(2), 198–213.

Campbell, C. S., Maglio, P. P., Cozzi, A., & Dom, B. (2003). Expertise Identification Using Email Communications. In *Proceedings of the 12th International Conference on Information and Knowledge Management* (pp. 528–531). New Orleans.

Chiu, C., Hsu, M., & Wang, E. (2006). Understanding knowledge sharing in virtual communities: an Integration Of Social Capital and Social Cognitive Theories. *Decision Support Systems, 42*, 1872–1888.

Cormack, G. V., Palmer, C. R., & Clarke, C. L. A. (1998). Efficient Construction of Large Test Collections. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 282–289). Melbourne.

Cummings, J. N. (2004). Work groups, structural diversity, and knowledge sharing in a global organization. *Management Science, 50* (3), 352.

Demartini, G., Gaugaz, J., & Nejdl, W. (2009). A vector space model for ranking entities and its application to expert search. *Advances in Information Retrieval, 5478*, 189–201.

Fleiss, J. L. (1981). *Statistical Methods for Rates and Proportions*. New York: John Wiley.

Griffiths, T., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Science, 101*(Suppl 1), 5228–5235.

Heinrich, G. (2005). *Parameter Estimation for Text Analysis*, Technical Report, Web: http://www.arbylon.net/publications/text-est.pdf.2005.

Herzig, D., and Taneva, H. (2010). Multilingual Expert Search using Linked Open Data as Interlingual Representation. In *Notebook Papers of the CLEF 2010 Labs and Workshops*, Padua, Italy, Web: http://www.clef2010.org/resources/proceedings/clef2010labs_submission_59.pdf

Hinds, P. J., & Kiesler, S. (2002). Preface. In P. J. Hinds & S. Kiesler (Eds.), *Distributed Work* (pp. xi–xviii). Cambridge: MIT Press.

Holley, J. O., Luehr, P. H., Smith, J. R., & Schwerha, J. J., IV. (2010). Electronic discovery. In E. Casey (Ed.), *Handbook of Digital Forensics and Investigation* (pp. 135–209). Burlington: Elsevier Academic Press.

Johri, A. (2011). Sociomaterial bricolage: the creation of location-spanning work practices by global software developers. *Information and Software Technology, 53*(9), 955–968.

Johri, A. (2006). Interpersonal assessment: Assessing peer knowledge and behavior in online learning environments. In T. S. Roberts (Ed.), *Self, Peer, And Group Assessment in E-Learning* (pp. 283–312). Hershey: Idea Group Publishing.

Kotlarsky, J., & Oshri, I. (2005). Social ties, knowledge sharing and successful collaboration in globally distributed system development projects. *European Journal of Information Systems, 14*(1), 37–48.

Lee, D. L., & Seamons, K. (1997). Document ranking and the vector space model. *IEEE Software, 14*(2), 67–75.

Lin, H., Fan, W., & Zhang, Z. (2009). A qualitative study of Web-based knowledge communities: examining success factors. *International Journal of E-Collaboration, 39*(3).

Majchrzak, A., Rice, R., Malhotra, A., King, N., & Ba, S. (2000). Technology adaptation: the case of a computer-supported inter-organizational virtual team. *MIS Quarterly, 24*(4), 569–600.

McDonald, D. W., & Ackerman, M. S. (1998). Just Talk to Me: A Field Study of Expertise Location. In *Proceedings of the 1998 ACM Conference on Computer Supported Cooperative Work* (pp. 315–324). Seattle.

McDonald, D. W., & Ackerman, M. S. (2000). Expertise Recommender: A Flexible Recommendation System and Architecture. In *Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work* (pp. 231–240). Philadelphia.

McLachlan, G. J., Do, K.-A., & Ambroise, C. (2004). *Analyzing Microarray Gene Expression Data*. Hoboken: John Wiley & Sons, Inc.

Melucci, M. (2007). On rank correlation in information retrieval evaluation. *ACM SIGIR Forum, 41*(1), 18–33.

Mimno, D., & McCallum, A. (2007). Expertise Modeling for Matching Papers with Reviewers. In *Proceeding of the 13th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 500–509). San Jose.

Nahapiet, J., & Ghoshal, S. (1998). Social capital, intellectual capital, and the organizational advantage. *Academy of Management Review, 23*(2), 242.

Newman, M. E. J., Watts, D. J., & Strogatz, S. H. (2002). Random graph models of social networks. *Proceedings of the National Academy of Sciences, 99*(suppl. 1), 2566–2572.

Palvia, P. (1998). Global information technology research: past, present, and future. *Journal of Global Information Technology Management, 1*(2), 6–29.

Papazafeiropoulou, A. (2004). Inter-country analysis of electronic commerce adoption in south eastern Europe: policy recommendations for the region. *Journal of Global Information Technology Management, 7*(2), 54–69.

Pipek, V., Wulf, V., & Johri, A. (2012). Bridging artifacts and actors: expertise sharing in organizational ecosystems. *Journal of Computer-Supported Cooperative Work, 21*(2–3), 261–282.

Remus, U., & Wiener, M. (2009). Critical success factors for managing offshore software development projects. *Journal of Global Information Technology Management, 12*(1), 6–29.

Ren, Y., Kraut, R. E., & Kiesler, S. (2007). Applying common identity and bond theory to the design of online communities. *Organizational Studies, 28*(3), 379–410.

Reichling, T., Veith, M., & Wulf, V. (2009). Expert recommender: Designing for a network organization. In J. M. Carroll (Ed.), *Learning in Communities: Interdisciplinary Perspectives on Human Centered Information Technology* (pp. 139–171). London: Springer-Verlag.

Serdyukov, P., Rode, H., & Hiemstra, D. (2008). Modeling Multi-step Relevance Propagation for Expert Finding. In *Proceeding of the 17th ACM Conference on Information and Knowledge Management* (pp. 1133–1142). Napa Valley.

Steyvers, M., Smyth, P., & Rosen-Zvi, M. (2004). Probabilistic Author-topic Models for Information Discovery. In *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 306–315). Seattle.

Thomas, J. C., Kellogg, W. A., & Erickson, T. (2001). The knowledge management puzzle: human and social factors in knowledge management. *IBM Systems Journal, 40*(4), 863–884.

Wasko, M. M., & Faraj, S. (2005). Why should I share? Examining social capital and knowledge contribution in electronic networks of practice. *MIS Quarterly, 29*(1), 35–57.

Wei, X., & Croft, W.B. (2006). LDA-Based Document Models for Ad-Hoc Retrieval. In *Proceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 178–185). Seattle.

Yang, J., Adamic, L. A., & Ackerman, M. S. (2008). Competing to Share Expertise: the Taskcn Knowledge Sharing Community. In *Proceeding of the International Conference on Weblogs and Social Media* (pp. 161–170). Seattle.

Zar, J. H. (2009). *Biostatistical Analysis*. Englewood Cliffs: Prentice-Hall.

Zhai, C., & Lafferty, J. (2004). A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems, 22*(2), 179–214.

Zhang, J., Ackerman, M. S., & Adamic, L. (2007). Expertise Networks in Online Communities: Structure and Algorithms. In *Proceedings of the 13th International World Wide Web Conference* (pp. 221–230). Banff.

Xiaomo Liu is a Ph.D. candidate in Computer Science at Virginia Tech and as Software Engineering at Microstrategy Inc. He holds an M.S. in Computer Science from Virginia Tech and an M.S. in Complex Adaptive Systems from Chalmers University of Technology, Sweden. His research interests include data and text mining, topic modeling, social media, and knowledge dissemination.

Dr. G. Alan Wang is an Assistant Professor in Business Information Technology at Virginia Tech. He received a Ph.D. in Management Information Systems from the University of Arizona, an M.S. in Industrial Engineering from Louisiana State University, and a B.E. in Industrial Management & Engineering from Tianjin University. His research interests include heterogeneous data management, data cleansing, data mining and knowledge discovery, and decision support systems. He has published in Communications of the ACM, IEEE Transactions of Systems, Man and Cybernetics (Part A), IEEE Computer, Group Decision and Negotiation, Journal of the American Society for Information Science and Technology, and Journal of Intelligence Community Research and Development.

Dr. Aditya Johri is an Assistant Professor in the Department of Engineering Education at Virginia Tech. He holds a Ph.D. in Education from Stanford University. His current research examines formal and informal learning in information technology intensive environments. His research has been recognized by a U.S. National Science Foundation Career Award and a Dean's Award for Outstanding New Assistant Professor at Virginia Tech, among others. More information about him is available at: http://filebox.vt.edu/users/ajohri.

Dr. Mi Zhou is an Assistant Professor of School of Management at Xi'an Jiaotong University, P.R. China. She received her Ph.D. in Management Science and Engineering from the School of Management at Xi'an Jiaotong University, P.R. China, in July 2007, a M.Sce. in Management Science and Engineering at Xi'an Jiaotong University, P.R. China, in 1998, and a B.E. in Mechanical Engineering from Nanjing Polytechnic University, P.R. China, in 1992. Her research interests focus on the information management, social relationship and knowledge management (knowledge transfer, knowledge sharing, knowledge creation), models of online knowledge communities.

Dr. Weiguo Fan is a Full Professor of Accounting and Information Systems and Full Professor of Computer Science (courtesy) at Virginia Tech) He received his Ph.D. in Business Administration from the Ross School of Business, University of Michigan, Ann Arbor, in 2002, a M. Sce in Computer Science from the National University of Singapore in 1997, and a B. E. in Information and Control Engineering from the Xi'an Jiaotong University, P.R. China, in 1995. His research interests focus on the design and development of novel information technologies — information retrieval, data mining, text/web mining, business intelligence techniques — to support better business information management and decision making. He has published more than 100 refereed journal and conference papers. His research has appeared in journals such as Information Systems Research, Journal of Management Information Systems, IEEE Transactions on Knowledge and Data Engineering, Information Systems, Communications of the ACM, Journal of the American Society on Information Science and Technology, Information Processing and Management, Decision Support Systems, ACM Transactions on Internet Technology, Pattern Recognition, IEEE Intelligent Systems, Pattern Recognition Letters, International Journal of e-Collaboration, and International Journal of Electronic Business.