Knowl Inf Syst (2010) 23:29–54
DOI 10.1007/s10115-009-0202-6

REGULAR PAPER

Integrating multiple document features in language models for expert finding

Jianhan Zhu · Xiangji Huang · Dawei Song · Stefan Rüger

Received: 29 August 2008 / Revised: 30 December 2008 / Accepted: 14 February 2009 /
Published online: 26 March 2009

Abstract We argue that expert finding is sensitive to multiple document features in an organizational intranet. These document features include multiple levels of associations between experts and a query topic from sentence, paragraph, up to document levels, document authority information such as the PageRank, indegree, and URL length of documents, and internal document structures that indicate the experts' relationship with the content of documents. Our assumption is that expert finding can largely benefit from the incorporation of these document features. However, existing language modeling approaches for expert finding have not sufficiently taken into account these document features. We propose a novel language modeling approach, which integrates multiple document features, for expert finding. Our experiments on two large scale TREC Enterprise Track datasets, i.e., the W3C and CSIRO datasets, demonstrate that the natures of the two organizational intranets and two types of expert finding tasks, i.e., key contact finding for CSIRO and knowledgeable person finding for W3C, influence the effectiveness of different document features. Our work provides insights into which document features work for certain types of expert finding tasks, and helps design expert finding strategies that are effective for different scenarios. Our main contribution is to develop an effective formal method for modeling multiple document features in expert

J. Zhu (✉)
Department of Computer Science, University College London,
Gower Street, London WC1E 6BT, UK
e-mail: jianhan.zhu@ucl.ac.uk; j.zhu@adastral.ucl.ac.uk

X. Huang
School of Information Technology, York University, Toronto M3J 1P3, Canada
e-mail: jhuang@yorku.ca

D. Song
School of Computing, The Robert Gordon University, Aberdeen AB25 1HG, UK
e-mail: d.song@rgu.ac.uk

S. Rüger
Knowledge Media Institute, The Open University, Milton Keynes MK7 6AA, UK
e-mail: s.rueger@open.ac.uk

finding, and conduct a systematic investigation of their effects. It is worth noting that our novel approach achieves better results in terms of MAP than previous language model based approaches and the best automatic runs in both the TREC2006 and TREC2007 expert search tasks, respectively.

Keywords Expert finding · Language models · Enterprise search

1 Introduction and motivation

Expert finding is a key task in enterprise search and has recently attracted lots of attention from both research and industry communities as evidenced by the organization of expert search tasks in the Text REtrieval Conference (TREC) in 2005, 2006 and 2007 [1,12,35], and the SIGIR 2008 Future Challenges in Expertise Retrieval Workshop.[1] In particular, for large national and global corporations, which are often distributed over different sites, it is a real challenge to automatically identify people with the necessary up-to-date expertise. A typical user scenario is one in which users need to learn about a subject and want to talk to someone who knows about it as the first step. Another use case is when a project manager wishes to assemble a project team made of people with a range of skills. Accordingly, Yimam-Seid and Kobsa [40] identified two main motives for expert finding, namely, as a source of information to answer the question "who knows about topic x?" (i.e., to find experts for a particular topic such as "Java programming" or "climate change", etc.) and also to answer questions such as "does person y know about topic x?" or "what else does y know?" They argued that manually developed expertise databases are labor-intensive and often quickly out-of-date. For example, in a large organization with lots of employees, it is a challenging task to organize a team of experts with different skills or chart the expertise of all its employees. It is hard to maintain an expertise database since there are both employees leaving the organization or joining the organization, and existing employees can gain new skills. On the other hand, much valuable and up-to-date expertise information often exists implicitly or explicitly in documents produced within the organization, for example, emails, blogs, wikis and web pages of individuals or groups, etc. For example, a person with expertise in "Java programming" may list "Java programming" on his/her homepage or blog, his/her email communications may be often associated with "Java programming", and the projects or groups associated with him/her may be related to "Java programming", etc.

The TREC enterprise track [1,12,35] has been the major forum for empirically comparing expertise modeling techniques. Since 2005, tremendous progress has been made in terms of expertise modeling, algorithms, and evaluation strategies. The goal of expert finding is to identify a list of people who are knowledgeable about a given topic. This task is usually addressed by uncovering associations between people and topics [12]; commonly, co-occurrences of a person's name with topic terms in the same context are assumed to be evidence of expertise. One example is that a person frequently associated with "Java programming" may have expertise on the topic. Furthermore, a ranked list of experts is preferable for the TREC expert finding task. The reason is that there may be many people with a particular expertise, and the ranked list based on a certain utility function can be more helpful to users, such as that the ranking is based on the level of expertise or accessibility, etc.

A prominent language modeling approach has been proposed by Balog et al. [2]. They distinguish between "Model 1", which directly represents the knowledge of an expert from

[1] http://ilps.science.uva.nl/fCHER/.

associated documents, and "Model 2", which first locates documents on the topic and then finds the associated experts. [31] have further improved their models by proposing a proximity-based document representation for incorporating sequential information in text. Serdyukov and Hiemstra [33] propose a novel expert-centric language model for expert search.

However, all these language modeling approaches have not sufficiently considered the effect of document features in expert finding. As rich document features exist in an organizational intranet environment and are shown to be effective for document retrieval [10], it is timely to study the effect of document features in expert finding. We discuss the following document features that expert finding is potentially sensitive to.

1. Internal document structure. Many organizational documents follow a certain template in formatting their contents. We argue that a document's internal structure can often be helpful in determining whether a person mentioned in the document is an expert on a topic that is also mentioned. For example, the occurrence of a person's name in the author, content, reference, or acknowledgement section of a technical paper on "climate change" may have different implications of the person's expertise on the "climate change" topic. If the person is the author or co-author of the paper, we are very certain that he/she has expertise on "climate change". And if the person's work is referenced in the paper, we need to check whether the person's work is on "climate change" in order to evaluate his/her expertise.
2. Document URLs. A URL (Uniform Resource Locator) often reflects the position of the document in the hierarchy of a website. We define the length of a URL as the number of sections divided by the "/" separator. Given a topic, entry documents on the topic often have shorter URLs, i.e., close to the root, while more detailed documents on the topic have longer URLs. Typically, each entry document links to these more detailed documents on a topic. We will study the effect of URL length in expert finding. Consider for example that one person is mentioned on an entry page about "climate change", and another person is mentioned on a more detailed page about "climate change", what is the two pages' effect on the two persons' expertise on "climate change"?
3. PageRank and indegree. The number of incoming links of a document (indegree) correlates with the document's PageRank [37]. Craswell et al. [11] integrate PageRank and indegree with a BM25 baseline model for more effective document retrieval than the BM25 baseline model. Cheng et al. [8] propose the use of PageRank for entity retrieval. We hypothesize that more authoritative documents are typically linked to more often by other documents. Based on the assumption that people mentioned in authoritative documents are more likely to be experts on a topic, we will investigate the effect of PageRank and indegree in expert finding, respectively.
4. Anchor texts. Anchor texts of a document often highlight its key topic. Sometimes, keywords for identifying a document's topic may even be missing in the document itself but exist in its anchor texts, e.g., the BMW homepage does not mention "car", but anchor texts pointing to the page often do. Anchor texts have been shown to be helpful in document retrieval on the Web [11,15]. We will study whether the effectiveness of anchor texts in document retrieval can be converted into their effectiveness in expert finding. For example, the anchor text pointing to a person's homepage may contain the keyword "climate change".
5. Multiple levels of associations between experts and topics. The proximity between occurrences of an expert and topic terms is a strong indicator of the expert's relevance to the topic. In traditional window-based association methods, a text window is set to measure the co-occurrences of the expert and query terms. Once the window size is set, it is fixed.

However, in expert finding, there are associations between an expert and query terms on multiple levels, i.e., from phrase, sentence, paragraph up to document levels. All these levels of associations need to be considered. For example, a person's name may co-occur with "climate change" within a text window of phrase, sentence, paragraph, or up to a document size, respectively. We will study the effect of multiple levels of associations in expert finding. Multiple levels of associations are further integrated with internal document structure in expert finding.

In this paper, we propose integrating the above five aspects in a unified language modeling approach for more effective expert finding. To the best of our knowledge, this is the first attempt to integrate a number of document features in a language model for expert finding [44]. Another vital contribution of this paper is to conduct a systematic investigation into the effects of multiple document features in expert finding on different TREC test collections.

The remainder of the paper is organized as follows. In Sect. 2, we review the related work. Our novel language model that integrates multiple document features is presented in Sect. 3. The experimental results on two large scale organizational intranet datasets, namely, the W3C (http://www.w3.org) and CSIRO (http://www.csiro.au) datasets, which represent real world expert finding scenarios, are reported in Sect. 4.

2 Related work

Early expert finding approaches have used corpus-wide statistical data. Expert Finder [27] works on such evidence as frequency of documents published by an expert on the topic, contents of resumes, and co-occurrence of the expert and query terms in documents. Conrad and Utt [9] and we [43] used corpus-wide statistical metrics such as mutual information, phi-squared, and CORDER measures to discover associations between named entities. Ohsawa et al. [29] used word co-occurrences for classifying Web communities. However, these approaches are solely based on co-occurrences or corpus statistics, and so do not consider document relevance with respect to the query. Although they are effective to confined domains such as community message boards in Ohsawa et al. [29], they are susceptible to noise in large-scale Web collections.

Campbell et al. [6] used email content to find related emails to a given topic, from which they constructed a graph consisting of email senders and receivers. They applied the HITS algorithm to the graph in order to identify experts with high authority. In a similar fashion, Tyler et al. [36] applied a betweenness centrality algorithm for finding communities in the link networks consisting of email senders and receivers, Guimera et al. [19]'s analysis of the email sender-receiver networks revealed the self-organization of the networks into a state of self-similarity, and Bar-Yossef et al. [4] proposed a strength measure called the integrated cohesion to clustering the link networks. However, these approaches are limited only to datasets with explicit linkage information.

Semantic web technologies have also been applied to expertise matching and search including the application of ontologies to peer-to-peer networks [20], and expertise matching based on published RDF files about experts' expertise [24]. Our text based expert finding approach can be integrated with the above link-analysis and semantic web based expert finding approaches. Our approach can also serve as a bootstrapping process for these link-analysis and semantic web based approaches. Furthermore, our approach can help alleviate the quality of expert finding results in these link-analysis and semantic web based approaches by providing explicit rankings of experts in response to a query.

The major forum for research in expert finding has been in the TREC Enterprise track [1,12,35]. Two real world large scale organizational intranets, i.e., W3C and CSIRO, have been used for experiments in 2005, 2006, and 2007.

Essentially, the two most popular and well-performing types of approaches in TREC expert search task are profile-centric and document-centric approaches [1,12,35].

Expert-profile-centric approaches build an expert profile as a pseudo document by aggregating text segments relevant to the expert, e.g., context text windows of the expert in documents [18]. Profiles of experts are indexed and searched for experts on a topic. Profiles can be significantly smaller than the original corpus, making the retrieval of experts efficient.

Document-centric approaches are typically based on traditional document retrieval techniques. Firstly, we estimate the conditional probability $p(q|d)$, of the query topic q given a document d. Based on the assumption that terms co-occurring with an expert in the same context describe the expert, $p(q|d)$ is used to weight the evidence of co-occurrence of experts with q in documents. The conditional probability $p(c|q)$ of an expert candidate c given a query q can be estimated by aggregating all the evidences in all the documents where c and q co-occur.

Document-centric approaches normally outperform profile-centric approaches [35] as the latter achieve efficiency at the expense of useful information in terms of internal document structure and high-level language features [30].

In contrast to the models by [2,31,33], which were discussed in the introduction, Cao et al. [7] propose a two-stage language model combining a document relevance and co-occurrence model. We [44] propose a unified language model integrating document features for expert finding. Fang and Zhai [17] derive a generative probabilistic model from the probabilistic ranking principle and extend it with query expansion and non-uniform candidate priors. We [42] propose a novel multiple window based approach for integrating multiple levels of associations between experts and query topic in expert finding. A number of query expansion techniques are also applied to expert finding [3,25,30]. The data fusion voting based expert finding approach by [25] can be seen as a combination of the document-centric and profile-centric approaches, where a voting model consists of votes from documents associated with each expert, and the rankings of these documents based on ad hoc retrieval techniques are used to determine the significance of the votes.

In our previous work, a multiple window based expert finding approach is proposed in [42], we improve the approach in [42] and present a generic language modelling framework for integrating document features in [44], we explore the relationship between expert finding and ad-hoc document retrieval in [45], and we proposed a generic two-stage approach for expert finding in [46].

The two differences between this paper and our work in [46] are as follows. Firstly, term independence in the co-occurrence model is assumed in [46], while term dependence in the co-occurrence model is assumed in this paper. Documents that partially match the query topic may be mistakenly taken into account under the term independence assumption, while the term dependence assumption may miss some relevant documents that only partially match the query topic. This is also noted in [31]. Since the title of a query generally consists of one to four terms, we think that the term dependence assumption in this paper is superior to the term independence assumption in [46]. Secondly, a generic two-stage approach is proposed in [46], while this paper focuses on a language modelling two-stage approach.

In summary, the new contributions in this paper are as follows. A systematic investigation of multiple document features and their effects on expert finding is carried out on large-scale TREC test collections in order to test the language modelling framework in [44]. We also explore the relationships between document features and two expert finding sub-tasks, i.e.,

knowledgeable-person and key-contact search, which are defined later in the paper, for two different TREC test collections. Our work in this paper on expert finding complements the findings in [45] about the relationship between expert finding and document retrieval.

Expert finding can be generalized to any type of entity search. The introduction of Entity Ranking Track in INEX 2007 on the Wikipedia dataset provides a platform for entity search evaluation [13]. Cheng et al. [8] propose the EntityRank algorithm which integrates local co-occurrence and global access information for entity search into a probabilistic estimation of entity and query association, which is quite similar to the above document-centric approaches used in expert finding.

3 Modeling document features

We first present our overall language modeling approach for expert finding. Secondly, we present our approach for integrating three query independent features, namely, PageRank, indegree, and URL length, in estimating document priors. Finally, we describe our approach for integrating multiple levels of associations and internal document structure in our co-occurrence model.

3.1 Language model for expert finding

Our models are instances of document-centric generative language modeling approaches to rank experts. Formally, given a set D of documents, a query topic q, and a set C of candidates, we state the problem of finding experts on q as "what is the probability of a candidate c in C being an expert given a query topic q?" The aim is to determine $p(c|q)$ and rank the set of candidates according to this probability:

$$p(c|q) = \frac{p(c,q)}{p(q)} \tag{1}$$

Here $p(c,q)$ is the joint probability of the candidate and query, and $p(q)$ is the probability of the query q. When evaluating $p(c|q)$, q is fixed, therefore, $p(c|q)$ is proportional to $p(c,q)$. To determine $p(c,q)$, we adopt a document-centric generative language modeling approach. We randomly draw independent samples of documents from $p(c,q)$ and represent the joint as a weighted average of the document models.

$$p(c,q) = \sum_{d \in D} p(c,q|d)p(d), \tag{2}$$

where $p(c,q|d)$ is the conditional probability of c and q given d, and $p(d)$ is the probability of d.

We can decompose $p(c,q|d)$ as:

$$p(c,q|d) = p(c|q,d)p(q|d) \tag{3}$$

By substituting Eq. 3 into Eq. 2, we obtain our expert finding model as:

$$p(c,q) = \sum_{d \in D} p(c|q,d)p(q|d)p(d), \tag{4}$$

The first term $p(c|q,d)$ of our expert finding model in Eq. 4 models the proximity between the topic and candidates. $p(c|q,d)$ also denotes a co-occurrence model as noted by Cao et al. [7]. We illustrate our approach for estimating $p(c|q,d)$ in Sect. 3.2.

The second term $p(q|d)$ of our model is the traditional language model for document retrieval for estimating the probability that d generates q. We will integrate anchor texts with document contents in the document retrieval in Sect. 3.3.

Finally, the third term $p(d)$ of our model incorporates the document priors. Most previous approaches ignore $p(d)$ by assuming that it is uniform for all documents. Therefore, there is no systematic study of the effect of $p(d)$ in expert finding. However, we argue that the estimation of $p(d)$ based on multiple features of d such as URL length, indegree, and PageRank, etc. can influence the performance of expert finding. We detail our approach for estimating $p(d)$ in Sect. 3.4.

3.2 Co-occurrence model

Our co-occurrence model in Eq. 4 is based on our previous work [42,44]. In 2006, we first proposed a multiple window based co-occurrence model in [42], and successfully applied the model to TREC 2006 expert search collection. Following our approach, Petkova and Croft [31] proposed a proximity model for expert finding. We proposed a language model based approach [44] based on our multiple window based approach in [42]. Our approach has two advantages over Petkova and Croft's [31] approach by taking into account multiple document features. Firstly, document internal structures as a document feature are considered in our co-occurrence model as discussed in this section. Secondly, our co-occurrence model is integrated with other document features such as anchor texts and document priors including PageRank, indegree, and URLs in our overall expert finding approach in Eq. 4.

In Eq. 4, by making a strong assumption that query terms and candidates are independent given a document, the probability $p(c|d, q)$ can be reduced to $p(c|d)$. This is true when explicit relationships between the candidate and the document can be established for cases such as that c is the author of d, or c is listed as the lead researcher on a project page, e.g., Sandra Eady, an expert on "meat rabbit production", is listed as a lead researcher on the project page of "Crusader meat rabbits" at http://www.csiro.au/science/psxo.html, etc. However, this assumption often does not hold due to two main reasons. Firstly, topic drift especially in long documents is very common, e.g., a research group's home page may include introductions of group members working in different research areas, and we cannot assume that all of them have expertise on a research area mentioned on the page. Secondly, people can be mentioned on a document due to reasons other than expertise associations such as that she is a contact point for a project, she is acknowledged by the author, or her paper is referenced by the document, etc. In all these cases, it is risky to say that the person is an expert on a topic mentioned in the document.

Therefore, it is a challenge to establish candidate and document associations. However, domain-specific features, such as the templates used for formatting W3C technical reports, can help disambiguate true and false associations between people and topics in documents. We will incorporate this kind of internal document structure feature in the co-occurrence model in the latter part of this section.

The co-occurrence model is constructed as a linear interpolation of $p(c|d, q)$ and the background model $p(c)$ to ensure there are no zero probabilities, we get

$$p(c|\theta_d, \theta_q) = (1 - \mu)p(c|d, q) + \mu p(c), \tag{5}$$

where $p(c)$ is the probability of candidate c. We estimate $p(c)$ as

$$p(c) = \frac{1}{df_c} \sum_{d' \in D} \frac{f(c, d')}{\sum_{c' \in C} f(c', d')}, \tag{6}$$

where $f(c, d')$ is the frequency of candidate c in document d' and df_c is the document frequency of c.

We use a Dirichlet prior for the smoothing parameter μ

$$\mu = \frac{\kappa}{\sum_{c' \in C} f(c', d') + \kappa}, \tag{7}$$

where κ is the average term frequency of all candidates in the corpus.

Based on the aforementioned reason that query terms and candidates are often not independent given a document, we use a proximity based document representation to estimate $p(c|d, q)$.

Since there are associations between a candidate and query terms on multiple levels, i.e., from phrase up to document level, we use a multiple window based approach in estimating $p(c|d, q)$. Based on an assumption that small windows often lead to more probable associations, and large windows may introduce noise resulting in nosier associations, we weight the contributions of smaller windows higher than larger windows. We will validate this assumption in Sect. 4.2.1. Vechtomova et al. [38] use long span associations between terms for query expansion. Metzler and Croft [28] use text windows of different sizes to model sequential dependence and full dependence of terms in a Markov random field model for document retrieval. Petkova and Croft [31] also report that a step function, which is equivalent to our multiple window based approach, results in better expert finding results than both Gaussian and triangle kernels.

Given a list W consisting of N windows w_i, $(i = 1, \ldots, N)$ of different sizes, we estimate $p(c|d, q)$ as

$$p(c|d, q) = \sum_{w} p(w) p(c|d, q, w), \tag{8}$$

where $p(w)$ is the probability for each of the window-based co-occurrence models.

Based on the nature of the section where c is mentioned in a document, we combine the internal document structure information with the window-based co-occurrence model. Given a number of text windows where c co-occurs with q as w_i, we estimate $p(c|q, d, w)$ as follows

$$p(c|d, q, w) = \sum_{w_i} \frac{f(c, d, q, w_i)}{\sum_{c' \text{in } C} f(c', d, q, w_i)}, \tag{9}$$

where $\sum_{c' in C} f(c', d, q, w_i)$ is the total frequency of candidates in w_i. Given a number of occurrences of c in w_i as c_j, $f(c, d, q, w_i)$ is estimated by combining internal document structure as

$$f(c, d, q, w_i) = \sum_{c_j} \delta[\text{Section}(c_j)], \tag{10}$$

where $\delta[\text{Section}(c_j)]$ is a weighting function given to the section where c_j occurs, e.g., higher weight to occurrences of c in the author section of a technical paper, and lower weight to occurrences of c in the acknowledgement section of the paper, etc. We train the weighting function in Sect. 4.2.4.

3.3 Document retrieval model

$p(q|d)$ in Eq. 4 is the probability that d generates q, and can be estimated by inferring a document language model θ_d for each document d such that

$$p(q|\theta_d) = \prod_{t \in q} p(t|\theta_d)^{n(t,q)}, \tag{11}$$

where t is a query term and $n(t, q)$ is the number of times it is used in q. We propose using a mixture of components to represent each document, where each component corresponds to certain fields or parts of the document. These components can be document body, title, anchor texts, and metadata, etc. We have focused on the effect of anchor text in expert finding, therefore

$$p(T|\theta_d) = (1 - \lambda_c)[\lambda_t\, p(t|d_{\text{text}}) + \lambda_a\, p(t|d_{\text{anchor}})] + \lambda_c\, p(t), \tag{12}$$

where the document content part is weighted with $(1 - \lambda_c)\lambda_t$, anchor text part is weighted with $(1 - \lambda_c)\lambda_a$, $\lambda_t + \lambda_a = 1.0$, and $p(t)$ is the maximum likelihood estimate (MLE) of the term t given the background model, weighted with λ_c. We carry out a systematic investigation of the effect of the settings of λ_1 in expert finding in Sect. 4.

3.4 Estimating document priors

In typical language modeling approaches, the prior probability of each document is assumed to be uniform. This is often approximately true for a static text collection such as the Wall street journal dataset. However, for an organizational intranet with rich query independent features, we assume that these features may help us estimate document priors in the language model.

Craswell et al. [11] use a number of query independent features including PageRank, indegree, URL Length and ClickDistance for document retrieval. They propose sigmoid transformations of these features in combination with a BM25 baseline. Their experiments on the TREC Web Track dataset show that the BM25 model integrating these features outperforms the BM25 baseline in document retrieval.

We study the effect of three query independent features, namely, PageRank, indegree, and URL length, in expert finding. Assuming PageRank, indegree, and URL length are independent features, we estimate $p(d)$ as

$$\begin{aligned} p(d) &\propto f_{\text{PR}}(d)\, f_{\text{URL}}(d), \text{ or} \\ p(d) &\propto f_{\text{indegree}}(d)\, f_{\text{URL}}(d), \end{aligned} \tag{13}$$

where $f_{\text{PR}}(d)$, $f_{\text{URL}}(d)$, and $f_{\text{indegree}}(d)$ are the transformation functions proposed by Craswell et al. [11] for PageRank, URL length, and indegree, respectively. Craswell et al. [11]'s experiments show that integrating PageRank via a sigmoid transformation function can greatly improve the MAP (mean average precision) of document retrieval on the TREC2004 Web Track dataset, and integrating both URL length and indegree via the same function also shows effectiveness in document retrieval, respectively. We use the sigmoid transformation

function proposed by them for estimating $f_{\text{PR}}(d)$, $f_{\text{URL}}(d)$, and $f_{\text{indegree}}(d)$, respectively:

$$f_{\text{PR}}(d) \propto w \frac{\text{PR}(d)^a}{k^a + \text{PR}(d)^a} \tag{14}$$

$$f_{\text{indegree}}(d) \propto w \frac{\text{indegree}(d)^a}{k^a + \text{indegree}(d)^a} \tag{15}$$

$$f_{\text{URL}}(d) \propto w \frac{\text{URL}_{\text{length}}(d)^a}{k^a + \text{URL}_{\text{length}}(d)^a} \tag{16}$$

Here w, a and k are parameters, and $\text{PR}(d)$, $\text{indegree}(d)$ and $\text{URL}_{\text{length}}(d)$ are the PageRank, indegree, and URL length of d, respectively. We followed the parameter settings[2] used by Craswell et al. [11] for PageRank, indegree, and URL length transformations by setting the values of w, a, and k as 1.8, 0.6, and 1.0, respectively, in Eq. 7; 3.6, 0.2, and 5, respectively, in Eq. 8; and 4.5, 0.5, and 4, respectively, in Eq. 9.

4 Experimental evaluation

The aim of our evaluation is to study the effects of these document features including internal document structure, URL length, PageRank, indegree, anchor texts and multiple sized windows in expert finding. We conduct a number of experiments on two large scale TREC datasets, i.e., the W3C and CSIRO datasets. All the 49 topics on the W3C dataset in the TREC2006 Expert Search task [35] and 50 topics on the CSIRO dataset in the TREC2007 Expert Search task [1] are used in our experiments.

4.1 Finding experts in documents

The W3C dataset consists of email lists, development code, web pages, wiki pages, other pages, and personal web pages. There are 331,037 documents in total. After excluding 62,509 documents containing development code, the average document length in the dataset is 699.85 terms, and there are 1,031,317 unique terms.

The CSIRO dataset is a crawl of the publicly available web pages from the *.csiro.au domain, known as the CSIRO Enterprise Research Collection (http://es.csiro.au/cerc/). The dataset consists of 370,715 documents with an average document length of 457.01 terms and 1,549,127 unique terms altogether.

The CSIRO dataset has a much smaller average document length than the W3C dataset, whose effects on our window based language model will be discussed in Sect. 4.2.1.

For the W3C dataset there is a pre-defined list of 1,092 W3C related people with their names and email addresses, which simplifies the problem of identifying people in text. However, as noted by Petkova and Croft [31], the quality of expert name extraction influences the performance of expert finding. We [42] created annotations of candidate occurrences, where advanced named entity recognition techniques are used, e.g., people's full names, name variations, email addresses, user IDs, etc. are matched using the Aho–Corasick algorithm. There are in total 1,662,024 occurrences of candidates in the W3C dataset. Our annotations have been widely used by other researchers in expert finding [31,39], and are therefore also used in our experiments here.

[2] Note that these parameter settings are only used as a guideline, and our experiments on the two datasets showed that our expert finding results are not sensitive to these parameter settings, i.e., for parameter values of a wide range, we have similar findings as those reported in Sect. 4.

However, expert name extraction on the CSIRO dataset is more like a real world people name identification problem since there is not a pre-defined list of candidates. Based on the observation that most CSIRO employees have a CSIRO email address following the pattern "firstname.lastname@csiro.au", we extract a list of candidates with email addresses matching this pattern from text. The candidates' full names, other names, and other email addresses are also extracted from text using regular expression patterns, and grouped with their CSIRO email addresses using identity matching techniques. Advanced named entity recognition techniques are used for generating variations of people's names. People's full names, name variations, email addresses, user IDs, etc. are matched using the Aho–Corasick algorithm. The total number of candidates is 3,483 with 808,148 occurrences in the dataset.

Our experiments on the two datasets show that while finding candidate occurrences is essential for good performance, when a large proportion of name occurrences have been recognized from text, the retrieval performance gains little despite more name occurrences being recognized. Furthermore, the performance of expert finding is robust since a small number of errors in name recognition do not hurt MAP significantly. Our findings are consistent with [31]'s.

Interestingly, the occurrences of candidates on both the W3C and CSIRO datasets follow the power law distribution as shown in Fig. 1a and b, respectively. A small number of candidates have a very large number of occurrences, and a majority of candidates have a small number of occurrences. We can see that the candidate occurrences on the CSIRO dataset conform more to a power law distribution than those on the W3C dataset. The reason is that the pre-defined list of the W3C experts excludes some W3C related people while a more complete list of CSIRO related people is extracted.

4.2 Experimental results and discussions

We pre-processed the two datasets by removing HTML tags, and used regular expression patterns to segment the documents into multiple sections. We indexed and searched the datasets with Lemur (http://www.lemurproject.org/). We report MAP, the main performance measure in TREC Expert Search task, of our expert finding approach, and study the effects of window size, anchor text, internal document structures, multiple windows, and document priors including URL length, PageRank and indegree in the following five subsections. Where stated, we tested statistical significance with t tests (one-tail critical values for significance levels $\alpha = 0.05$).

4.2.1 Effects of the window size

In Eq. 5, we smooth the mixture of document and anchor text models with the background model by setting λ_c as 0.05.

Our baseline is a basic single window based language model for the two datasets where anchor texts are not used, i.e., $\lambda_t = 1.0$, $\lambda_a = 0.0$ and $\lambda_c = 0.05$, as shown in Fig. 2.

In Fig. 2, the two curves show the MAP with respect to different window sizes for the W3C and CSIRO datasets, respectively. The two curves are similar in that when the window size is under 170 for the W3C and 110 for the CSIRO datasets, the MAP increases rapidly when the window size increases. When the window size increases further beyond 170 for the W3C and 110 for the CSIRO datasets, the MAP does not increase significantly and in the case of the CSIRO dataset drops to a significantly lower level at around 400 terms.

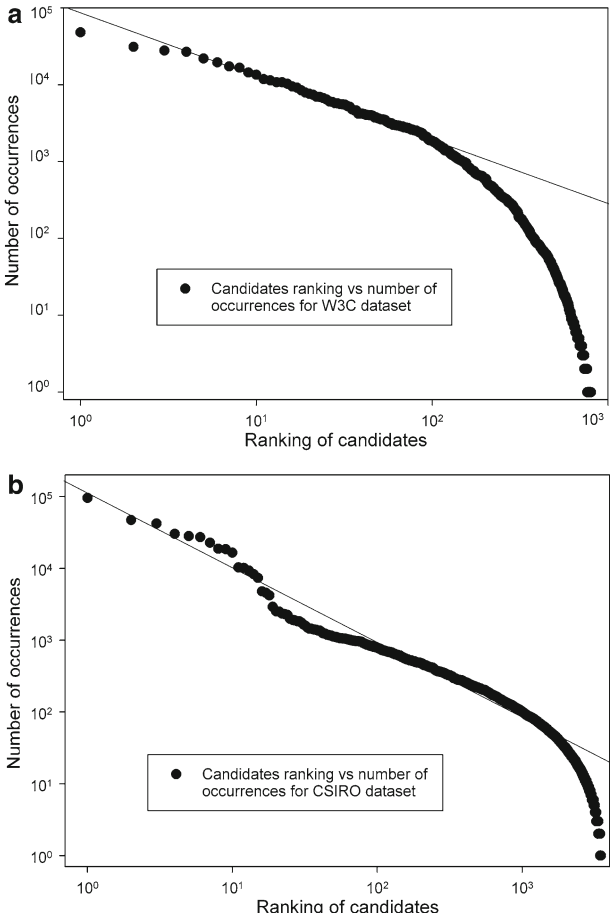


Fig. 1 Candidate occurrences on the two datasets follow approximately the Zipf's law. **a** W3C dataset, and **b** CSIRO dataset

The window based approach on the W3C dataset shows robustness in terms of the observation that the MAP reaches a rather stable level at the window size of around 260, and further increases in window size only result in statistically insignificant changes in MAP. The W3C curve reflects that there are many levels of associations between candidates and query topics, e.g., sentence, paragraph, section levels, etc. The increase of a small window size leads to many novel associations discovered with little noise, resulting in rapidly increasing MAPs. When the window size exceeds 170, there are less novel associations discovered and more noise is introduced, leading to slow increase.

For the CSIRO dataset, a window size of between 100 and 200 produces a relatively high MAP, with higher values tailing slightly off.

Dissimilarities of the two curves can be understood in terms of a key difference in the characteristics of the two datasets:

Expert finding on the CSIRO dataset is essentially a key-contact search, while knowledgeable-person search is carried out on the W3C dataset. We define key-contact search as a search task where only the main contacts or project leaders of a particular query topic need

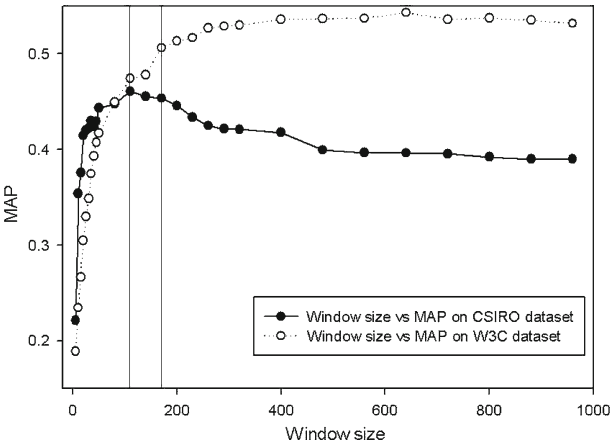


Fig. 2 MAP of baseline single window language model for different window sizes; $\lambda_t = 1.0$, $\lambda_a = 0.0$ and $\lambda_c = 0.05$

to be identified. Since the key-contacts are usually only one or two in the CSIRO collection, the users can easily identify the experts to contact with regarding a query. On the other hand, we define knowledgeable-person search as a search task where people with general knowledge on a query topic can all be recognized as experts. Knowledgeable-person search is more helpful to cases such as charting the expertise inside an organization and enterprise knowledge management, etc.

Experts for the CSIRO dataset were provided by the CSIRO science communicators, who only selected a few key contacts on these topics resulting in only 2.76 experts per topic on average. In contrast, experts on the TREC2006 W3C test collection were manually judged by the participants, who identified 28.43 experts per topic on average, which is much more than that of the CSIRO dataset.

Furthermore, the average document length of the W3C dataset is also significantly longer than that of the CSIRO dataset, probably resulting in more long range associations between experts and topics.

Hence, it is natural that the associations between key contacts and topics on the CSIRO dataset are more concentrated in close or medium range, i.e., sentence or paragraph levels rather than document levels. Therefore, the MAP for the CSIRO dataset can be expected to initially increase more quickly than the MAP for the W3C dataset. In addition, associations between experts and topics on the W3C dataset are more evenly distributed across multiple levels owing to the larger number of experts, and longer documents exemplified by long technical reports and technical papers. Therefore, the MAP for the W3C dataset keeps increasing much longer with the window size.

To study the effects of different levels of candidate and query topic associations in expert finding, we consider performance measures such as MAP and Precision at 5 (P@5) for gap windows of equal length of 20, i.e., 0–20, 20–40, 40–60, 60–80, and so on. We count the total number of co-occurrences of candidates and query topics for each gap window, and get MAP and P@5 for each gap window. If we divide these performance measure scores by their respective total number of co-occurrences of candidates and query topics, the results can give us an idea of how much each co-occurrence for a gap window contributes to the effectiveness of expert finding on average. The higher the contribution, the more useful each co-occurrence is in expert finding, and vice versa. The results are shown in Table 1.

Table 1 MAP and P@5 divided by the total number of co-occurrences of candidates and query terms for each gap window, and MAP and P@5 gains for each gap window

Gaps	W3C dataset		CSIRO dataset	
	Avg MAP for each co-occurrence ($\times 10^{-7}$)	Avg P@5 for each co-occurrence ($\times 10^{-7}$)	Avg MAP for each co-occurrence ($\times 10^{-7}$)	Avg P@5 for each co-occurrence ($\times 10^{-7}$)
0–20	65.64	125.72	89.36	173.48
20–40	55.24 (−15.84%)	106.14 (−15.58%)	71.68 (−19.79%)	141.25 (−18.58%)
40–60	54.70 (−16.67%)	105.73 (−15.90%)	70.84 (−20.72%)	137.01 (−21.02%)
60–80	52.21 (−20.46%)	101.79 (−19.03%)	63.46 (−28.98%)	127.06 (−26.76%)
80–100	47.22 (−28.06%)	94.76 (−24.63%)	52.27 (−41.51%)	110.83 (−36.11%)
100–120	45.63 (−30.48%)	93.04 (−26.00%)	48.94 (−45.23%)	111.01 (−36.01%)
120–140	44.13 (−32.77%)	91.51 (−27.21%)	43.07 (−51.80%)	104.63 (−39.69%)
140–160	42.89 (−34.67%)	85.62 (−31.89%)	40.27 (−54.93%)	95.12 (−45.17%)
160–180	41.63 (−36.59%)	80.97 (−35.60%)	38.34 (−57.09%)	94.88 (−45.31%)
180–200	36.26 (−44.77%)	73.71 (−41.37%)	37.28 (−58.28%)	81.02 (−53.30%)
200–220	35.85 (−45.39%)	73.92 (−41.20%)	32.20 (−63.97%)	74.49 (−57.06%)
200–240	33.16 (−49.48%)	67.99 (−45.92%)	28.09 (−68.57%)	75.29 (−56.60%)
240–260	34.55 (−47.37%)	71.51 (−43.12%)	23.95 (−73.20%)	69.88 (−59.72%)
260–280	32.86 (−49.94%)	70.77 (−43.71%)	22.27 (−75.08%)	66.72 (−61.54%)
280–300	30.00 (−54.30%)	66.19 (−47.35%)	20.57 (−76.98%)	57.98 (−66.58%)
300–320	28.41 (−56.72%)	65.76 (−47.69%)	19.93 (−77.70%)	57.16 (−67.05%)
320–340	27.90 (−57.50%)	65.46 (−47.93%)	19.15 (−78.57%)	56.58 (−67.38%)

We can clearly see from Table 1 that the average MAP or P@5 score for each co-occurrence on both datasets consistently decreases when the distances between candidates and query topic terms increase. When the gap window is beyond 340, the average MAP or P@5 score for each co-occurrence keeps decreasing in a similar trend as illustrated in Table 1 for both datasets, respectively. In particular, for the W3C dataset, the average MAP and P@5 for each co-occurrence of the gap window 320–340 decreases by 57.5 and 47.93% from those of the gap window 0-20, respectively. For the CSIRO dataset, the average MAP and P@5 for each co-occurrence of the gap window 320–340 decreases by 78.57 and 67.38% from those of the gap window 0–20, respectively.

Results in Table 1 support our assumption that close range co-occurrences often indicate more probable expertise associations than longer range co-occurrences. Longer range co-occurrences introduce more associations at the expense of more noise, therefore, the average MAP or P@5 score for each co-occurrence decreases as the distances between candidates and query terms increase.

Furthermore, the average MAP and P@5 for each co-occurrence on the CSIRO dataset decrease more quickly than those on the W3C dataset, respectively. This supports our findings on Fig. 2 that expertise associations on the CSIRO dataset are more concentrated on short ranges than those on the W3C dataset. Therefore, expert finding on the W3C dataset will benefit more from our multiple-window-based approach than that on the CSIRO dataset. Our experimental results in Sect. 4.2.5 support this finding.

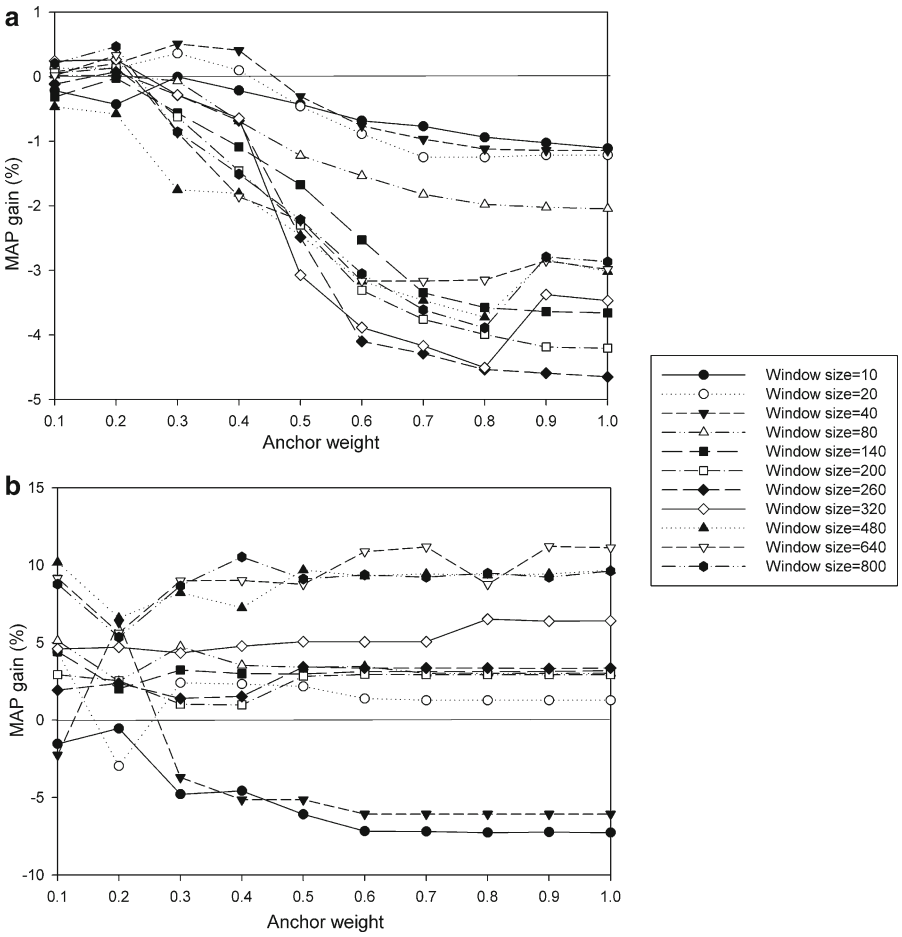


Fig. 3 When anchor text takes different weights in the model, percentage of gain on MAP (pgMAP) for (**a**) W3C and (**b**) CSIRO datasets

4.2.2 Effects of anchor text

We experimented with different configurations of λ_t and λ_a in Eq. 5 using a number of window sizes. The aim was to see whether the effectiveness of anchor texts in document retrieval on the Web [10,15] can be carried over to expert finding on the two Website datasets.

We varied λ_t from 0.9 to 0.0 in steps of 0.1. The curves of the MAPs for these ten configurations were very similar to the two curves in Fig. 2 for the CSIRO and W3C datasets, respectively. Therefore, we used the two curves in Fig. 2 as the baselines to study the effect of the other parameter λ_a, and plot the percentage of gain on MAP (pgMAP), which is defined as the difference between the new and old MAPs (gMAP) divided by the old MAP, of these configurations for different window sizes in Fig. 3a and b, respectively.

In Fig. 3a, when the contribution of anchor text is small, i.e., $\lambda_a = 0.1$ and 0.2, the baseline MAP values for most window sizes are improved. However, these increases are not statistically significant. When the contribution of anchor text increases further, i.e., λ_a is 0.3 and

higher, MAP will be hurt. When λ_a is between 0.3 and 0.5, the decreases from the baseline are not statistically significantly, but when λ_a is between 0.6 and 1.0, the performance is statistically significantly worse than the baseline.

However, in Fig. 3b, when the window size is 480 or above, all the anchor text enhanced models perform statistically significantly better than the baseline.

We think that the difference of the results on the two datasets is again due to the different nature of the two test collections. The CSIRO communicators create topics and provide key contacts as experts on these topics. These topics are generally well known research areas inside the CSIRO, and these key contacts are often mentioned in authoritative documents on these topics. These authoritative documents typically have more links from other pages and therefore keywords on the topic often occur in anchor texts. Therefore, anchor texts are helpful in expert finding on the CSIRO dataset. This is reinforced by our findings of the effect of PageRank, indegree, and URL length in the next section.

On the other hand, because there are many more experts per topic on the W3C dataset, over-stressing the importance of authoritative documents will introduce more noise than useful information, e.g., some people appearing on the authoritative documents are not experts while some true experts may not appear on documents with lots of incoming links.

4.2.3 Effects of PageRank, indegree, and URL length

We used the model enhanced by anchor text where $\lambda_a = 0.2$ as the baseline for integrating PageRank, indegree, and URL length.

In both Fig. 4a and b, for both datasets, we can see that the three models enhanced by indegree, PageRank, and indegree + URL length, respectively, improve the baseline. The two models enhanced by indegree and PageRank, respectively, have very similar curves, showing a strong correlation between indegree and PageRank, and therefore their effect in expert finding. For both datasets the two models enhanced by indegree and PageRank, respectively, perform better than the model enhanced by indegree + URL length. The model enhanced by URL length alone performs the worst. This coincides with previous research that PageRank and indegree are better measures for document authority than URL length in document retrieval [11].

Since PageRank, indegree, and URL length are all indicators of document authority, our results show the effect of differentiating authoritative documents from ordinary documents in expert finding. On the W3C dataset, although the three models seem to improve the baseline, these increases are not statistically significant. On the other hand, on the CSIRO dataset, when the window size is above 400, all four models improve the baseline with statistical significance. The models in Fig. 4b exhibit a strong similarity to the models in Fig. 3b. We think this similarity is due to the fact that anchor texts, PageRank, indegree, and URL length all incorporate page authority information and so they have similar effect in expert finding.

On the other hand, for the W3C dataset, there are many more experts per topic, and many of them may not often appear on authoritative pages but rather technical reports and papers, therefore, the incorporation of page authority is helpful but less effective than for the CSIRO dataset.

4.2.4 Effects of internal document structure

In the W3C dataset, candidates often appear in the context of long technical reports, papers, and emails. We study how this internal structure of these documents can be helpful in determining a candidate's expertise on a topic.

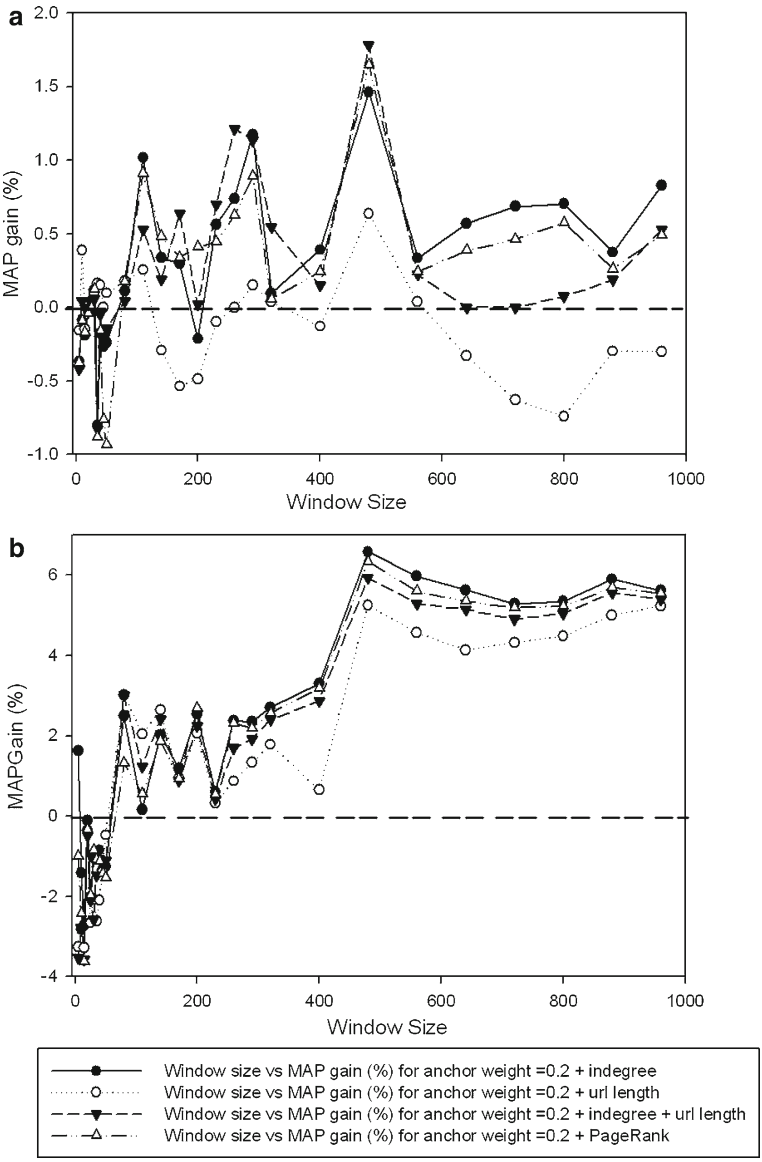


Fig. 4 When integrated with indegree, URL length, and/or PageRank, percentage of gain on MAP (pgMAP) for (**a**) W3C and (**b**) CSIRO datasets

We used the 50 topics on the W3C dataset in the TREC2005 Expert Search Task to train the weighting function in Eq. 15. After training, $\delta[(\text{Section}(c_j)]$ is set as 1.0, 7.5, 0.6, 0.2, 5.2, 1.2, 0.7, and 0.5 for candidate occurrences in the document body, author, acknowledgements, references, email sender, email receiver, email CC, BCC sections, respectively.[3] We

[3] We used fivefold cross-validation for training the weights. Based on the assumption that these weights are independent, i.e., these weights do not influence each other, we trained the weights one after another, e.g., first train the weight for the acknowledgements section, then the weight for the references section, and so on.

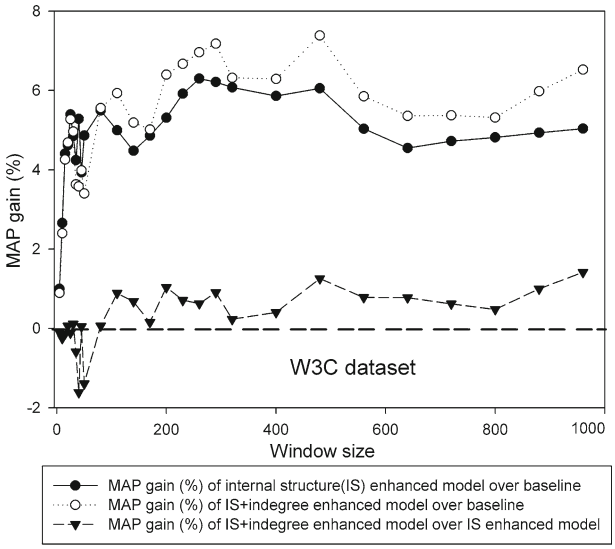


Fig. 5 When integrated with internal structure and/or indegree, percentage of gain on MAP (pgMAP) for W3C dataset

used the model enhanced by anchor text, where $\lambda_a = 0.2$ as the baseline for integrating internal document structure and indegree.

In Fig. 5, both models enhanced by internal structure (IS), and IS + indegree, respectively, outperform the baseline model with statistical significance. Since internal structure and indegree are independent features, i.e., they describe different aspects of documents, the IS + indegree enhanced model further improves the IS alone enhanced model, however, the increased MAPs are not statistically significant.

Since the CSIRO dataset does not contain many academic papers and emails with internal structure information useful for expert finding, internal document structure is not considered for the CSIRO dataset.

4.2.5 Effects of window combination

Our results in Sect. 4.2.1 show that shorter range associations between candidates and query topic can often provide more probable expertise evidences than longer range associations. In single window based approach, once the window size is set, the co-occurrence model does not distinguish the range in which a candidate and the query topic co-occur. Therefore, it is hard to set the optimal window size. To overcome the limitation of single window based approach, we propose a multiple window based approach which takes into account proximity of candidates and query topic terms.

We approximate sentence, paragraph, section, and document level expertise associations by window size under 20, between 20 and 100, between 100 and 350, and above 350, respectively. In Eq. 13, we assume that $p(w)$ follows a Gaussian distribution function as used in [31] for combining co-occurrence models. We selected three window sizes, i.e., 100, 300, and 640, based models as the baselines for combining with another window. In order to study the effects of document features in window combinations, in Fig. 6a and b, we plot the MAP gains of these window combination models, and indegree and/or internal structure enhanced window combination models on the W3C and CSIRO datasets, respectively.

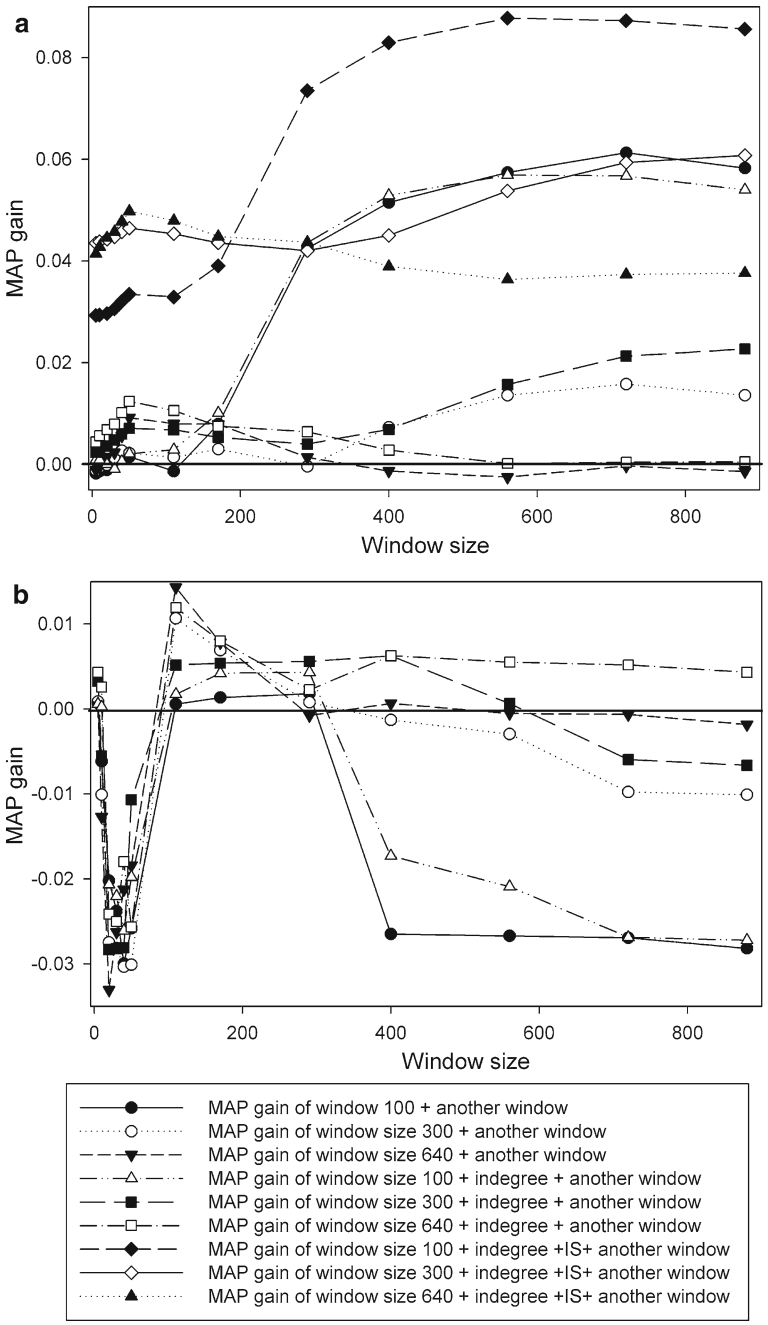


Fig. 6 When integrated with another window, internal structure, and/or indegree, gain on MAP (gMAP) for (**a**) W3C and (**b**) CSIRO datasets

In Fig. 6a, we can see that by combining two windows, the MAPs of two baselines, i.e., window size 100 and 300, are largely improved. In particular, MAPs of the window size 100 baseline are statistically significantly improved by combining with a window above 280,

while MAPs of the window size 640 baseline are not improved. Introduction of internal structure further improves all three combination models, but not significantly so. Internal structure greatly improves all three combination models with statistical significance.

The model consisting of window sizes 50 and 640, indegree, and internal structure achieves the highest MAP of 0.5948, which is even better than the best automatic run in the TREC2006 expert search task [35] with the MAP value of 0.5947, and significantly better than the results reported in previous language models such as [3] with the highest MAP of 0.4728. Furthermore, for the two window combination, i.e., 50 and 640, indegree, and internal structure, we found that adding a third window with size between 100 and 300 will increase the MAP further. The window combination of 50, 200, and 640 achieves the MAP value of 0.6087, which largely improves previous reported results and is still not the optimal value for our window combination approach. Our window combination approach has good potential for improving the single window based approach. In future work, we will investigate a systematic approach for window combinations.

However, in Fig. 6b, we can see that window combination does not help improve the three baselines and hurts the performance sometimes, i.e, window combination only results in statistically insignificant changes in MAP. This matches our finding in Sect. 4.2.1 that expertise associations on the CSIRO dataset concentrates more on short range than those on the W3C dataset.

In Fig. 2, our single window based approach can significantly outperform the best two stage model based approach in the TREC2007 expert search task [1,14] with the best MAP value of 0.4427 and all the other language model based approaches in the task [1]. Considering that [14] used query expansion, the improvement made by our approach is more significant. Our single window approach achieves the best MAP of 0.4609, 0.4553, and 0.4535 for the window size of 100, 140, and 170, respectively. By combining with anchor text in Fig. 3b, our best results in terms of MAP are further improved. The window combination model in Table 2 consisting of window sizes 100 and 280, and indegree achieves an even higher MAP of 0.4688.

The improvements of MAP by document features and window combination for three window sizes, i.e., 100, 300, and 640, on the two datasets, are summarized in Table 2, where both positive and statistically significant improvements with t tests (one-tail critical values for significance levels $\alpha = 0.05$) over the two baselines are highlighted, respectively. A comparison of our approach with competing systems on the TREC task is summarized in Table 3.

5 Conclusions

In order to develop generic expert finding approaches applicable to different scenarios, we have demonstrated that it is important and beneficial to study the effect of multiple document features. We proposed a novel approach of integrating document features in a language model for expert finding, and carried out a systematic investigation of the effects of document features in expert finding. Based on our experiments on the two TREC datasets, i.e., W3C and CSIRO datasets, we have the following findings.

We found that in order to achieve good MAP, the window size used for association discovery should be sufficiently large, e.g., above 100 terms. Small window sizes, e.g., under 50 terms, are certain to miss useful associations.

Expert finding on the CSIRO dataset is a key contact search where very few experts per topic are defined, while expert finding on the W3C dataset is a knowledgeable-person search

Table 2 Effect of document features and window combination on MAP, where MAPs with positive improvements are in bold, and statistically significant improvements are in bold and marked with '*'

	$W = 100$, $\lambda_a = 0$ (Baseline)	$W = 100$, $\lambda_a = 0.2$	$W = 100$, $\lambda_a = 0.2$ indegree	$W = 100$, and 280, λ_a=0.2	$W = 100$, and 280, $\lambda_a = 0.2$ indegree	$W = 100$, $\lambda_a = 0.2$, IS	$W = 100$, $\lambda_a = 0.2$, indegree, IS	$W = 100$ and 280, $\lambda_a = 0.2$, indegree, IS
MAP (W3C)	0.4741	0.4723 (−0.38%)	**0.4771** (+0.63%)	**0.5189*** (+9.45%)	**0.5194*** (+9.55%)	**0.4961** (+4.64%)	**0.5003** (+5.53%)	**0.5503*** (+16.07%)
MAP (CSIRO)	0.4609	**0.4643** (+0.74%)	**0.4685** (+1.65%)	**0.4651** (+0.91%)	**0.4688** (+1.71%)			
	$W = 300$, $\lambda_a = 0$ (Baseline)	$W = 300$, $\lambda_a = 0.2$	$W = 300$, $\lambda_a = 0.2$ indegree	$W = 300$, and 110, λ_a=0.2	$W = 300$, and 110, $\lambda_a = 0.2$ indegree	$W = 300$, $\lambda_a = 0.2$, IS	$W = 300$, $\lambda_a = 0.2$, indegree, IS	$W = 300$ and 110, $\lambda_a = 0.2$, indegree, IS
MAP (W3C)	0.529	0.5281 (-0.17%)	**0.5343** (+1.00%)	**0.5348** (+1.10%)	**0.5391** (+1.91%)	**0.5608*** (+6.01%)	**0.5660*** (+6.99%)	**0.5784*** (+9.34%)
MAP (CSIRO)	0.4215	**0.4286** (+1.68%)	**0.4314** (+2.35%)	**0.4332** (+2.78%)	**0.4305** (+2.14%)			
	$W = 640$, $\lambda_a = 0$ (Baseline)	$W = 640$, $\lambda_a = 0.2$	$W = 640$, $\lambda_a = 0.2$ indegree	$W = 640$, and 50, λ_a=0.2	$W = 640$, and 50, $\lambda_a = 0.2$ indegree	$W = 640$, $\lambda_a = 0.2$, IS	$W = 640$, $\lambda_a = 0.2$, indegree, IS	$W = 640$ and 50, $\lambda_a = 0.2$, indegree, IS
MAP (W3C)	0.5432	**0.5450** (+0.33%)	**0.5481** (+0.90%)	**0.5542** (+2.03%)	**0.5574** (+2.61%)	**0.5698** (+4.90%)	**0.5742*** (+5.71%)	**0.5948*** (+9.50%)
MAP (CSIRO)	0.3964	**0.4152** (+4.74%)	**0.4187*** (+5.63%)	**0.4208*** (+6.16%)	**0.4176** (+5.35%)			

Table 3 Comparing our results with competing systems on the TREC task, where our results that are better than all of the previous competing systems' results are in bold

	$W = 50$, 640, indegree IS, $\lambda_a = 0$	$W = 50$, 200, 640, indegree IS, $\lambda_a = 0$	Best TREC 2006 run	Balog et al. [3]	
MAP (W3C)	**0.5948**	**0.6087**	0.5947	0.4728	
	W=100	W=140	W=170	W=100, 280 indegree, λ_a=0	Duan et al. [14]
MAP (CSIRO)	**0.4609**	**0.4553**	**0.4535**	**0.4688**	0.4427

where dozens of experts per topic are typical. Based on this difference, medium sized window should be used for key contact search since these key contact associations with the topic are more focused within medium range, and large windows introduce more noise than useful information. Associations of knowledgeable people and a topic tend to distribute more evenly across multiple windows, therefore, large window sizes should be used.

Table 4 Effects of document features in two types of expert finding tasks

	Key-contact search	Knowledgeable-person search	Both search tasks
Single-window size	Medium size	Large size	Above 100 terms
Anchor text	Effective	Less effective	
URL length	Effective	Less effective	Correlate with PageRank and indegree, but less effective than them
Document internal	(Not tested)	Effective with statistical significance, and complement indegree	
Windows combination	Less effective	Effective	

Anchor texts are more useful in key contact search since key contacts often appear in authoritative documents which attract inlinks, therefore anchor texts. We found that an increased weight of anchor text in expert finding leads to better performance than a pure document content based approach for large window sizes. However, anchor texts are less effective for knowledgeable-person search since many experts may not appear in authoritative documents.

URL length is less effective than PageRank and indegree for both datasets in expert finding, which is also the case in document retrieval [11]. Due to the strong correlations between PageRank/indegree and document authority, they are both effective for key contact search, but less effective for all knowledgeable person search. PageRank/Indegree and URL length have duplicate effect in expert finding.

The rich internal structures of documents in the W3C dataset help improve expert finding with statistical significance, signifying its importance in expert finding on structurally rich datasets. Internal structures and indegree are complementary in expert finding since they describe different aspects of documents.

Window combination is effective for expert finding on the W3C dataset showing the wide distribution of expertise associations on different ranges, while less effective on the CSIRO dataset due to the concentration of expertise associations in small and medium ranges. Indegree and internal structures are both effective for combination windows on both datasets, especially, internal structure help improve combination models with statistical significance.

We summarize the effect of different document features on expert finding in Table 4.

Our expert finding approach has achieved superior results in terms of our best MAPs on the two TREC datasets that are both better than previous language model-based approaches [1,3] and those of the best automatic two-stage model runs in the TREC2006 and TREC2007 expert search tasks [1,14,35], respectively, even without using other techniques such as query processing and query expansion. We believe that our expert search performance can be further improved by using some efficient query expansion techniques, and our window combination approach has the potential for further improvement in terms of three or more window combinations and methods for window combination optimization.

In our future work, we plan to study the effect of query expansion and its relationships with multiple document features in expert finding. In integrating PageRank, URL length,

and indegree, we will investigate different transformation functions and explore the effect of parameters in the transformation functions in expert finding. It would also be useful to study the effect of other document features, such as document types, and clickthrough data. We will also apply our approach to other datasets and generic entity search.

The layouts of web sites even within an organization can vary. To tackle this challenge of extracting document internal structures, we will integrate our approach with the wrapper induction for information extraction approach proposed by Kushmerick et al. [23]. The approach has been successfully applied to web site data extraction such as the Lixto system [5]. Since such wrapper induction approaches can dynamically and automatically extract structured knowledge from semi-structured information sources [5,23], there will be little manual labor involved when applying our expert finding approach to web sites of different organizations.

URL length might have been indicative of document hierarchy when HTML was prevalent. However, many Web sites now generate content dynamically from SQL databases. The organization of the SQL database may not be typically reflected in the URL. We will explore how to apply our approach to such dynamic Web sites. One possible approach is to use the URL rewriting technique to convert URLs of dynamic web pages into more informative URLs with structures [16]. For example, a URL which contains query string parameters to encode the date of a blog posting[4] can be automatically converted to a new URL as http://www.example.com/Blogs/2006/12/10/. Therefore, the URL length can be taken into account in our approach.

When applying our approach to a new domain for expert finding, our language modelling approach has the advantage of providing a range of parameters for tuning in order to adapt to the new domain. These parameters control the effect of different document features such as multiple windows, anchor texts, PageRank, and URL length, etc., as illustrated in Sect. 3. Using machine learning and data mining techniques to automatically tune parameters of an information system (such as an IR or database systems) is a well-established research topic [21,22,32,34,41]. In our future work, we will study how to use these techniques in relevance feedback, such as that a user has given us one or two true experts for a topic, for automatically tuning our expert finding model parameters.

Acknowledgments We would like to thank the anonymous reviewers for their valuable and constructive comments.

References

1. Bailey P, Craswell N, de Vries AP, Soboroff I (2008) Overview of the TREC 2007 enterprise track. In: Proceedings of TREC 2007
2. Balog K, Azzopardi L, de Rijke M (2006) Formal models for expert finding in enterprise corpora. In: Proceedings of SIGIR 2006, pp 43–50
3. Balog K, Bogers T, Azzopardi L, de Rijke M, van den Bosch A (2007) Broad expertise retrieval in sparse data environments. In: Proceedings of SIGIR 2007, pp 551–558
4. Bar-Yossef Z, Guy I, Lempel R, Maarek YS, Soroka V (2008) Cluster ranking with an application to mining mailbox networks. Knowl Inf Syst 14(1):101–139
5. Baumgartner R, Frölich O, Gottlob G (2007) The Lixto systems applications in business intelligence and semantic Web. In: Proceedings of ESWC (European Semantic Web Conference), pp 16–26
6. Campbell CS, Maglio PP, Cozzi A, Dom B (2003) Expertise identification using email communications. In: Proceedings of CIKM 2003

[4] http://www.example.com/Blogs/Posts.php?Year=2006&Month=12&Day=10.

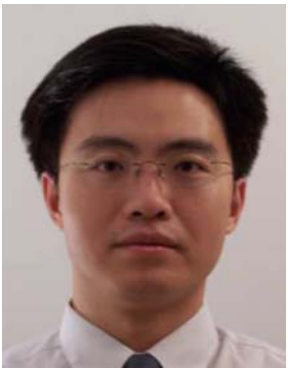
7. Cao Y, Liu J, Bao S, Li H (2006) Research on expert search at enterprise track of TREC 2005. In: Proceedings of TREC 2005
8. Cheng T, Yan X, Chang KC-C (2007) EntityRank: searching entities directly and holistically. In: Proceedings of VLDB 2007, pp 387–398
9. Conrad JG, Utt MH (1994) A system for discovering relationships by feature extraction from text databases. In: Proceedings of SIGIR 1994, pp 260–270
10. Craswell N, Hawking D (2005) Overview of the TREC-2004 Web track. In: Proceedings of TREC 2004
11. Craswell N, Robertson SE, Zaragoza H, Taylor MJ (2005) Relevance weighting for query independent evidence. In: Proceedings of SIGIR 2005, pp 416–423
12. Craswell N, de Vries AP, Soboroff I (2006) Overview of the TREC-2005 enterprise track. In: Proceedings of TREC 2005
13. de Vries AP, Thom JA, Vercoustre A-M, Craswell N, Lalmas M (2008) Overview of the INEX 2007 entity ranking track. In: Proceedings of INitiative for the Evaluation of XML Retrieval
14. Duan H, Zhou Q, Lu Z, Jin O, Bao S, Cao Y, Yu Y (2008) Research on enterprise track of TREC 2007 at SJTU APEX lab. In: Proceedings of TREC 2007
15. Eiron N, McCurley KS (2003) Analysis of anchor text for Web search. In: Proceedings of SIGIR 2003, pp 459–460
16. Engelschall R (1999) A users guide to URL rewriting with the Apache Webserver. http://httpd.apache.org/docs/2.0/misc/rewriteguide.html
17. Fang H, Zhai C (2007) Probabilistic models for expert finding. In: Proceedings of ECIR 2007, pp 418–430
18. Fu Y, Xiang R, Zhang M, Liu Y, Ma S (2006) A PDD-based searching approach for expert finding in intranet information management. In: Proceedings of AIRS 2006, pp 43–53
19. Guimera R, Danon L, Diaz-Guilera A, Giralt F, Arenas A (2003) Self-similar community structure in a network of human interactions. Phys Rev Ser E 68(6):065103-1–065103-4
20. Haase P, Siebes R, van Harmelen F (2008) Expertise-based peer selection in Peer-to-Peer networks. Knowl Inf Syst 15(1):75–107
21. Huang X, Huang Y, Wen M, An A, Liu Y, Poon J (2006a) Applying data mining to pseudo-relevance feedback for high performance text retrieval. In: Proceedings of the 2006 IEEE International Conference on Data Mining (ICDM'06), pp 295–306
22. Huang X, Yao Q, An A (2006b) Applying language modeling to session identification from database trace logs. Knowl Inf Syst Int J (KAIS) 10(4):473–504
23. Kushmerick N, Weld D S and Doorenbos R B (1997) Wrapper induction for information extraction. In: Proceedings of IJCAI, pp 729–737
24. Liu P, Curson J, Dew PM (2008) Use of RDF for expertise matching within academia. Knowl Inf Syst 8(1):103–130
25. Macdonald C, Ounis I (2007) Expertise drift and query expansion in expert search. In: Proceedings of CIKM 2007, pp 341–350
26. Macdonald C, Ounis I (2008) Voting techniques for expert search. Knowl Inf Syst 16(3):259–280
27. Maybury M, D'Amore R, House D (2001) Expert finding for collaborative virtual environments. Commun ACM 44(12):55–56
28. Metzler D, Croft WB (2005) A Markov random field model for term dependencies. In: Proceedings of SIGIR, pp 472–479
29. Ohsawa Y, Soma H, Matsuo Y, Matsumura N, Usui M (2002) Featuring web communities based on word co-occurrence structure of communications. In: Proceedings of WWW2002, pp 436–442
30. Petkova D, Croft WB (2006) Hierarchical language models for expert finding in enterprise corpora. In: Proceedings of IEEE international conference on tools with artificial intelligence, pp 599–608
31. Petkova D, Croft WB (2007) Proximity-based document representation for named entity retrieval. In: Proceedings of CIKM, pp 731–740
32. Salton G, Buckley C (1990) Improving retrieval performance by relevance feedback. J Am Soc Inf Sci (JASIS) 41(4):288–297
33. Serdyukov P, Hiemstra D (2008) Modeling documents as mixtures of persons for expert finding. In: Proceedings of ECIR 2008
34. Shen X, Tan B, Zhai C (2005) Context-sensitive information retrieval using implicit feedback. In: Proceedings of SIGIR2005, pp 43–50
35. Soboroff I, de Vries AP, Craswell N (2007) Overview of the TREC 2006 enterprise track. In: Proceedings of TREC 2006
36. Tyler JR, Wilkinson DM, Huberman BA (2003) Email as spectroscopy: automated discovery of community structure within organizations. Communities and Technologies, pp 81–96
37. Upstill T, Craswell N, Hawking D (2003) Predicting fame and fortune: PageRank or indegree? In: Proceedings of the Australasian Document Computing Symposium ADCS 2003

38. Vechtomova O, Robertson SE, Jones S (2003) Query expansion with long-span collocates. Inf Retr 6(2):251–273
39. Westerveld T (2007) Correlating topic rankings and person rankings to find experts. In: Proceedings of TREC2006
40. Yimam-Seid D, Kobsa A (2003) Expert finding systems for organizations: problem and domain analysis and the DEMOIR approach. Journal of Organizational Computing and Electronic Commerce 13(1)
41. Zhai C, Lafferty JD (2001) Model-based feedback in the language modeling approach to information retrieval. In: Proceedings of CIKM2001, pp 403–410
42. Zhu J, Song D, Rüger S, Eisenstadt M, Motta E (2007) The Open University at TREC 2006 enterprise track expert search task. In: Proceedings of TREC 2006
43. Zhu J, Gonçalves AL, Uren VS, Motta E, Pacheco R, Eisenstadt M, Song D (2007) Relation discovery from web data for competency management. Web Intell Agent Syst Int J 5(4):405–417
44. Zhu J, Song D, Rüger S, Huang J (2008) Modeling document features for expert finding. In: Proceedings of CIKM 2008, pp 1421–1422
45. Zhu J (2008) A study of the relationship between ad hoc retrieval and expert finding in enterprise environment. In: Proceedings of CIKM tenth international workshop on web information and data management (WIDM), pp 25–30
46. Zhu J, Song D, Rüger S (2009) Integrating Multiple Windows and Document Features for Expert Finding. J Am Soc Inf Sci Technol (JASIST) (in press)

Author Biographies

Jianhan Zhu is a postdoctoral research fellow at the University College London, UK. Before this, he worked as a research fellow at the Knowledge Media Institute of the Open University, UK, between 2003 and 2008. He received a PhD in Computer Science working on Web data mining from the University of Ulster, UK, in 2003. His Bachelor's degree in automatic control was obtained from the Tsinghua University, China, in 1998. He has published extensively in internationally recognized journals and conferences. He has been the co-investigator and manager of a number of research projects. His research interests include statistical information retrieval models, expert finding, semantic web, and data mining.

Xiangji Huang joined York University as an Assistant Professor in 2003 and became an Associate Professor in 2006. Previously, he was a Post Doctoral Fellow at the School of Computer Science, University of Waterloo. He did his PhD in Information Science at City University in London. He also worked in the financial industry in Canada doing E-business, where he was awarded a CIO Achievement Award. He has published more than 80 refereed papers in journals, book chapters and international conference proceedings. His Master and Bachelor degrees were in Computer Organization & Architecture and Computer Engineering, respectively. His research interests include information retrieval, health informatics and data mining. In April 2006, he was awarded tenure at York University. He received the Dean's Award for Outstanding Research in 2006, an Early Researcher Award, formerly the Premier's Research Excellence Awards in 2007 and the Petro Canada Young Innovators Award in 2008.



Dawei Song is a Professor of Computing at The Robert Gordon University, UK. His major research interests are information retrieval models with special focus on context-sensitive IR; enterprise and domain specific search systems; and knowledge discovery from textual and multimedia data repositories. He received his PhD from Chinese University of Hong Kong in 2000 and has worked at the Distributed Systems Technology Centre, Australia, during 2000–2005 and The Open University, UK, during 2005–2008. He has been the Principal Investigator of a number of UK and Australia research council projects, and has published regularly at various highly respected conferences and journals.

Stefan Rüger joined The Open University's Knowledge Media Institute in 2006 to take up a chair in Knowledge Media and head a research group working on Multimedia Information Retrieval. Before that he was a Reader in Multimedia and Information Systems at the Department of Computing, Imperial College London, where he also held an EPSRC Advanced Research Fellowship (1999–2004). Rüger is a theoretical physicist by training (FU Berlin) and received his PhD in the field of computing for his work on artificial intelligence and, in particular, the theory of neural networks from TU Berlin in 1996. For further information see http://kmi.open.ac.uk/mmis.