Modeling Multi-step Relevance Propagation for Expert Finding

Pavel Serdyukov
Database Group
University of Twente
Enschede, The Netherlands
serdyukovpv@cs.utwente.nl

Henning Rode
CWI
Amsterdam, The Netherlands
h.rode@cwi.nl

Djoerd Hiemstra
Database Group
University of Twente
Enschede, The Netherlands
hiemstra@cs.utwente.nl

ABSTRACT

An expert finding system allows a user to type a simple text query and retrieve names and contact information of individuals that possess the expertise expressed in the query. This paper proposes a novel approach to expert finding in large enterprises or intranets by modeling candidate experts (persons), web documents and various relations among them with so-called *expertise graphs*. As distinct from the state-of-the-art approaches estimating personal expertise through one-step propagation of relevance probability from documents to the related candidates, our methods are based on the principle of multi-step relevance propagation in topic-specific expertise graphs. We model the process of expert finding by probabilistic random walks of three kinds: finite, infinite and absorbing. Experiments on TREC Enterprise Track data originating from two large organizations show that our methods using multi-step relevance propagation improve over the baseline one-step propagation based method in almost all cases.

Categories and Subject Descriptors

H.3 [**Information Storage and Retrieval**]: ;; H.3.3 [**Information Search and Retrieval**]: ;

General Terms

Algorithms, Measurement, Performance, Experimentation

Keywords

Enterprise search, expertise, expert finding, language models, random walks, Markov processes

1. INTRODUCTION

In large organizations or in the scope of the global web users often search for persons rather than for relevant documents. The demanded information may be just not published: not considered important to be released, not available in electronic format, or just hardly expressible in written language. In these cases asking people becomes the only way to find an answer [15]. Besides being sources of unpublished knowledge, the experts on the search topic are often able to explain the details and guide the user further. Experts can be in demand not only for asking them questions, but also for assigning them to some role or a job. Conference organizers may search for reviewers, company recruiters for talented employees, even consultants for other consultants to redirect inquiries and not to lose clients [26].

The need in finding a well-informed person may be critical for any kind of project. However, any attempts to identify experts by manual browsing through organizational documents or via informal social connections may fail in large enterprises, especially when they are geographically distributed. A standard text search engine may be of great help, but still is not able to automate this task. Usually, a specialized *expert finding system* is developed to assist in the search for individuals or departments that possess certain knowledge and skills within the enterprise and outside [36]. It allows either to save time and money on hiring a consultant when company's own human resources are sufficient, or helps to find an expert at affordable cost and convenient location in another organization. Similarly to a text search engine, an automatic expert finder uses a short user query as an input and returns a list of persons sorted by their level of knowledge on the query topic.

It is common to consider that the more often a person is related to the documents containing many words describing the topic, the more likely we may rely on this person as on an expert. The proof of the relation between a person and a document can be an authorship (e.g. we may consider external publications, descriptions of personal projects, sent emails or answers in message boards), or just the occurrence of personal identifiers (names, email addresses etc.) in the text of a document. Thus, the most successful approaches to expert finding obtain their estimator of personal expertise by summing the relevance scores of documents directly related to a person [5, 35]. Alternative approaches find experts by measuring network centrality of persons in specialized professional communities [10].

Both aggregated relevance and centrality based expert finding methods still ignore some properties of the data. Methods using aggregated relevance do not reflect relations between experts and do not consider documents that are indirectly related to persons. Whereas the centrality based methods simply model documents as unweighted links between candidates, neglecting their relevance to the query.

CIKM'08, October 26–30, 2008, Napa Valley, California, USA.

Our contributions. In this paper, we advance previous work by combining features of both above-mentioned approaches. We demonstrate that it is beneficial to continue the propagation of document relevance after the first step of its aggregation on the level of directly related persons. Following the principles of spreading activation algorithms [17], we allow the probability of document relevance to flow further through reciprocal connections between persons and documents. Specifically, our contributions are:

- We propose several ways to model the multi-step relevance dissemination in topic-specific *expertise graphs* consisting of persons and top retrieved documents. The introduced expertise graphs form the background for three different expert finding methods: based on a finite, an infinite and a specialized parameter-free absorbing random walk. As a result, we allow persons to receive expertise evidence from documents even not being in immediate proximity with them. At the same time, documents in turn get evidence of relevance not solely from their own content but also partly from the content of directly and indirectly linked documents.
- Since we model the expert finding as a walking process in a graph of topical documents and related persons, our approach has the advantage that it naturally utilizes hyperlinks between documents and professional connections between people.
- Experiments demonstrate that the principle of multi-step relevance propagation not only represents a more generalized view on modeling of expert finding, but also leads to noticeable improvements over the baseline one-step propagation. These improvements are observed over almost all points of the parameter space and are statistically significant.

The remainder of this paper is organized as follows. The related research on expert finding and link-based analysis is described in detail in the next section. In Section 3 we explain how the dynamics of an expertise domain can be modeled with graphs, as well as motivate and propose expert finding methods built upon random walks in these graphs. Our experiments with two real-world test collections supporting our assumptions are discussed in Section 4. Finally, Section 5 concludes the presented work and outlines the directions of future research.

2. RELATED WORK

2.1 Expert finding

Expert finding started to gain its popularity at the end of '90s, when Microsoft, Hewlett-Packard and NASA made their experiences in building such systems public [18, 19, 8]. They were not fully automated and represented repositories of manually created skill descriptions of their employees with simple search functionality. Nowadays these and other companies invest a lot into making their expert search engines commercially available and attractive [1, 2, 21]. However, apart from causing the new boom on the growing enterprise search systems market, expert finding systems also compelled close attention of the IR research community [23]. The expert search task is a part of the Enterprise track of the Text REtrieval Conference (TREC) since its first run in 2005 [14]. The TREC community created experimental data sets consisted of organizational document collections, lists of candidate experts and the set of search topics, each with a list of actual experts.

First, pioneering approaches to expert finding could be classified as *profile-centric* [15, 33]. This technology was the first step in full automation of expert finding in organizations and generally aimed to avoid manual maintenance of personal profiles. In these approaches, all documents related to a candidate expert are merged into a single personal profile prior to the actual retrieval process. The personal profiles are ranked like documents w.r.t a query using standard retrieval measures and corresponding best candidate experts are returned to the user. There is also a special class of effective profile-centric methods representing the relevance model as a mixture of personal language models and then measuring the probability that a person would generate the query given the probabilities to generate relevant terms in top documents [43, 42].

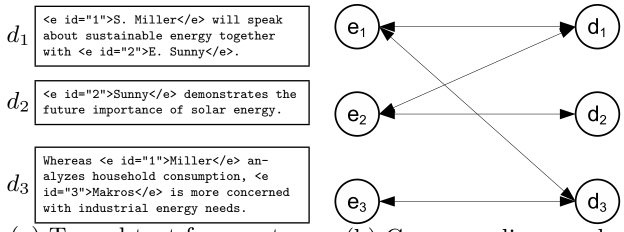
The follow-up *document-centric* approaches analyze the content of each document separately [11, 5, 35, 20]. All of them basically utilize a simple probabilistic model assuming that the probability of expertness of a person is a sum of relevance probabilities of all related documents (see Section 3.2 for details). Some recent works attempt to avoid propagation of the relevance of those documents or their parts that are not related strongly enough to the candidate expert. In one approach, only the score of the text window of a fixed size surrounding the person's mention is considered [34]. In the other approaches, either the overall pairwise distance between a candidate's mention and the query terms [40] or their order of occurrence in a document [44] are used to calculate the degree of association between the document and the candidate.

All the above-mentioned approaches share the principle claiming that the relevance of the textual context of a person adds up to the evidence of his/her expertness. Furthermore, each next class of approaches appears to be more effective than the previous one [6, 40], seemingly due to the estimation of relevance of the textual content related to a person on the lower and hence less ambiguous level. However, in their definition of personal context, these approaches restrict themselves to the scope of documents directly related to a person. They ignore the complexity of link structure among persons and documents and hence do not consider the expertness of directly and indirectly linked persons as well as the relevance of documents found not in immediate neighborhood of a candidate.

In this connection, it is important to mention another line of research that proposed finding experts by calculating their centrality in the organizational social network [10, 28, 51]. However, these approaches ignore the relevance of documents that serve just as the evidence of relation between persons. So, while being effective for query-independent tasks like finding the most authoritative experts in question/answering portals and forums [4], they are inferior in performance to the state-of-the-art query-dependent expert finding methods [12].

2.2 Link-based analysis

Random walk based models regularly appear in different IR research areas, but first of all known from web retrieval. Among them, Pagerank [39], HITS [29] (its random walk



```
<e id="1">S. Miller</e> will speak
about sustainable energy together
with <e id="2">E. Sunny</e>.
```

```
<e id="2">Sunny</e> demonstrates the
future importance of solar energy.
```

```
Whereas <e id="1">Miller</e> an-
alyzes household consumption, <e
id="3">Makros</e> is more concerned
with industrial energy needs.
```

(a) Tagged text fragments (b) Corresponding graph

Figure 1: Expertise Graphs

based version is described by Ng et. al. [38]) and SALSA [32] are probably the most popular. Several attempts have been made in the last years to make these models query and content dependent. The Intelligent Surfer [41] walks to linked pages biased by their relevance to the query. Personalized Pagerank allows to put preferences on certain webpages, so that the centrality of any document would depend on its proximity to preferred ones [39, 27]. Both ideas were further combined in a unified framework which considered also the bi-directional walk over hyperlinks [46]. Random walks on graphs containing queries and clicked links (or entire search trails) were recently utilized for web search result expansion [16, 9]. Searching with graph-based methods for typed entity classes on the Web was explored recently in several publications [50, 49].

There are applications of random walks beyond the bounds of hyperlinked corpora. Pagerank in graphs of terms, documents and document clusters was adapted for ad-hoc text retrieval [31, 30]. Finite random walk over terms through thesaural and syntactic relations is applied to query expansion [48, 13] and question answering [22] tasks. It also used as a model of preference flow between users in a recommendation system [47].

The presented work, to our knowledge, is the first extensive study of relevance propagation with random walks on query-dependent graphs in the field of expert finding.

3. EXPERTISE ESTIMATION BY RELEVANCE PROPAGATION

3.1 Expertise Graphs

This section proposes and discusses the modeling of appropriate graphs that represent the association between experts and documents in a certain domain of expertise. Suppose we have a set of documents associated with scores as the result of an initial standard document retrieval on the given topic. From the ranked documents, a second set of contained candidate experts is extracted. Their containment relations can be represented in an *expertise graph*, where both documents and candidate experts become vertices and directed edges symbolize the containment conditions (see Figure 1). The simplest form of expertise graphs is always bipartite, since all edges point from documents to experts only and back. Figure 2 shows a typical expertise graph computed for one of the TREC queries. Let us recall that we are interested in propagation of relevance through the graph network. So, it is further important to exploit all known connections between the entities of the graph.

Including Further Links. In many situations not only containment relations, but also links between documents or organizational connections between possible experts are known. Whereas inter-document links are represented by directed edges following the link, expert-to-expert edges are usually bidirectional. By including such additional edges, the graph gains a higher density and enables more intensive relevance propagation, however by losing its strict bipartite property.

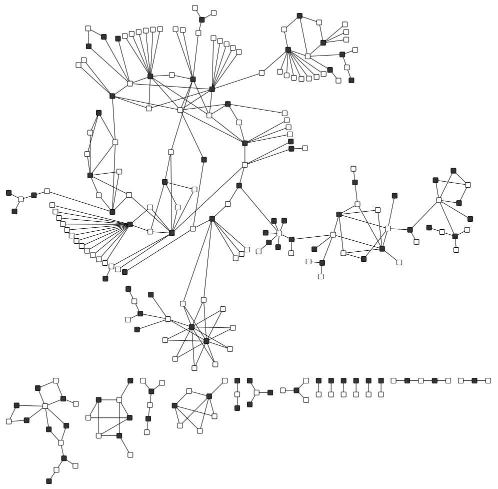


Figure 2: A fragment of the real expertise graph with links between documents (white nodes) and candidate experts (black nodes) for query "sustainable ecosystems"

Including Further Entity Types. Experts and documents do not need to be the only entities in the expertise graph. Although the expert finding task is only interested in the ranking of experts, it might still be useful for the relevance propagation to exploit additional connections via nodes of other types, such as dates, locations or events. Persons outside the company might reveal interesting connections as well, if they are mentioned in the documents together with the candidate experts. We may also incorporate relations and entities extracted from other external global professional networks (e.g. LinkedIn.com).

Controlling Graph Size and Topical Focus. Apart from the graph modeling itself, the most influential parameter on the graph size and density is the number of retrieved documents taken into account while building the graph. Notice that only the restriction to the top ranked documents makes the expertise graph model query dependent. We can include more lower ranked documents to increase the graph's density, but with the drawback of losing its topical focus.

3.2 Baseline: one-step relevance propagation

One of the most theoretically sound and effective representatives of document-centric expert finding methods (see Section 2), proposed by Balog et al. [5, 6], follows the probabilistic language modeling principle of IR [24] and defines the probability of expertness for the candidate expert e with respect to the query Q as:

$$P(Expert|e) = \sum_{D \in Top} P(R|D)P(e|D)P(D) \qquad (1)$$

$$P(R|D) \propto P(D|Q) = \frac{P(Q|D)P(D)}{\sum_{D' \in Top} P(Q|D')P(D')}, \quad (2)$$

where $P(R|D)$ is the probability that the document D is relevant and $P(Q|D)$ is the probability of the document D to generate the query Q. The latter probability is the measure of document relevance R according to LM-based IR [24]. $P(e|D)$ is the probability of association between the candidate and the document. Top is the set of documents retrieved and the prior probability $P(D)$ is distributed uniformly over the Top.

If we look at Equations 1 and 2 we may notice that they correspond to a probabilistic process, in which a user selects a document among the ones appearing in the initial ranking, looks through the document, enlists all candidate experts mentioned in it and refers with the current information need to one of them. The probability of selecting a document is its probabilistic relevance score since the user will most probably search for useful information and contacts of knowledgeable people in one of the top documents recommended by a search engine. The following selection of a candidate expert depends on the level of its responsibility to the content of the document: e.g. its author will most probably be selected first, but a person mentioned in the acknowledgments will be less likely considered useful. The described process can be interpreted as *one-step relevance probability propagation* from documents to related candidate experts.

We use the method described by Equation 1 as our baseline. The probability of the query to be generated by the document language model [24] is calculated as:

$$P(Q|D) = \prod_{q \in Q} P(q|D), \quad (3)$$

$$P(q|D) = (1 - \lambda_G)\frac{tf(q, D)}{|D|} + \lambda_G \frac{\sum_{D' \in C} tf(q, D')}{\sum_{D' \in C} |D'|} \quad (4)$$

where $tf(q, D)$ is the term frequency of q in the document D, $|D|$ is the document length and λ_G is a Jelinek-Mercer smoothing parameter - the probability of a term to be generated from the global language model calculated over the entire collection C. We set it to 0.8 which is the optimal value according to our preliminary experiments.

Here and further on we also use the probabilities of selecting a document given a person and of selecting a person given a document:

$$P(D|e) = \frac{a(e, D)}{\sum_{D'} a(e, D')}, P(e|D) = \frac{a(e, D)}{\sum_{e'} a(e', D)}, \quad (5)$$

where $a(e, D)$ is the non-normalized association score between the candidate e and the document D proportional to their strength of relation. Our way of distributing these scores over candidate experts in a document is described in the experimental part of the paper (see Section 4.1).

3.3 Motivating multi-step relevance propagation

If we want to automatically point the user to the most knowledgeable people on the topic, we should imagine how they could be found instead during manual search. The one-step probabilistic process is not quite a realistic model in this case. It is not likely that reading only one document and consulting only one person is enough to completely satisfy a personal information need in the enterprise. The real-world user should realize that the expertise needed is partly contained in several retrieved documents and partly in the personal memory of several experts.

We may imagine that the search for expertise may consist of the following repeating stages of gradual knowledge acquisition:

1. At any time: (a) randomly reading a document, or just picking a random candidate,
2. After reading a document: (a) consulting with a person mentioned in this document, or (b) checking for other linked documents and reading one of them, or
3. After consulting with a person: (a) reading other documents mentioning this person, or (b) consulting with another candidate expert which is recommended by this person.

Note that while modeling expertise gathering process, we apply different techniques to concentrate the random walk around the most relevant documents, since we rely on the assumption that all sources of the same knowledge are located close to each other in expertise graphs. In our methods described further in this section we try to overcome the limitations of the baseline one-step relevance propagation. We model the expert finding as a K-step, an infinite or an absorbing process of consulting with documents and people. First we present three models considering that expertise graphs are bipartite (graphs used for the infinite random walk are not strictly bipartite due to the probability of a jump to any node), and then we suggest the model taking links among same-type entities into account.

3.4 Finite random walk

In this approach, we consider that the user makes some predefined number of steps in his/her search for expertise. Since the user walks over a bipartite expertise graph with layers of document and candidate expert nodes, this walk becomes a process of moving to a node from an opposite layer at each step, starting from some node in a document layer.

In order to emphasize the importance of a candidate to be in close proximity to relevant documents, we utilize the probabilities of their relevance in two ways: (1) the probability of selection of the first document as a starting point for the walk is proportional to its probability of relevance, (2) the probability to stay at a document node at any step is also proportional to its probability of relevance. Actually, the non-zero self-transition probability is important for finite random walks, since it allows to diffuse the initial probability more slowly, smoothly and hence makes the algorithm less sensitive to the setting of number of steps.

Since we consider this walk as finite, we believe that at some point a user is tired/satisfied with some candidate and stops the search process. So, we iteratively calculate the probability that a random surfer will end up with a certain candidate after K steps of a walk started at one of the initially ranked documents:

$$P_0(D) = P(R|D), P_0(e) = 0, \quad (6)$$

$$P_i(D) = P(R|D)P_{i-1}(D) + \sum_{e \to D} P(D|e)P_{i-1}(e), \quad (7)$$

$$P_i(e) = \sum_{D \to e} (1 - P(R|D)) P(e|D) P_{i-1}(D) \quad (8)$$

The probabilities $P(e|D)$, $P(D|e)$ and $P(R|D)$ used in above equations are defined in Equations 2 and 5. Finally, we consider that $P(Expert|e) \propto P_K(e)$.

It is also possible to estimate the candidate's expertise using several finite walks of different lengths at once. For instance, it can be calculated as a weighted sum of probabilities $P_1(e) \ldots P_K(e)$. We could also smooth the current node probabilities with probabilities to appear in the same nodes in the past and future. However, all such approaches would significantly increase the size of our parameter set due to introduction of weight coefficients. So, despite that our method can be easily utilized in this way, we experiment only with its unsmoothed case.

3.5 Infinite random walk

In our second approach, we assume that the walk in search for expertise is a non-stop process. We may imagine that the user visits document and candidate nodes over and over again making a countless number of steps. By analyzing the statistics of this discrete Markov process we may conclude that persons visited more often during this infinite walk were more beneficial for the user. However, its stationary distribution does not depend on the initial probability distribution over states. In order to retain the importance for a candidate to stay in proximity to relevant documents and also to assure the existence of a stationary distribution, we introduce jump transitions to the nodes of a graph.

At first, we introduce the possibility to return regularly to the document nodes from any node of the expertise graph and to start the walk through mutual documents-candidates links again. We consider that the probability of jumping to the specific document $P_J(D)$ equals its probability to be relevant to the query. This assumption makes candidate experts which are situated closer to relevant documents visited more often in total during consecutive walk steps.

We also add a probability to jump to candidates $P_J(e)$. We consider that the more often the candidate appears in top documents, the more likely that it is known to the user sooner or later and hence can be selected for a random jump. So, we make it equal to the probability to find the candidate in a randomly selected document from the retrieved top. However, it can have other origins and may be proportional to the candidate's popularity/authority in the whole organization or inversely proportional to his/her occupancy level.

The following equations are used for iterations until convergence:

$$P_i(D) = \lambda P_J(D) + (1 - \lambda) \sum_{e \to D} P(D|e) P_{i-1}(e), \quad (9)$$

$$P_i(e) = \lambda P_J(e) + (1 - \lambda) \sum_{D \to e} P(e|D) P_{i-1}(D) \quad (10)$$

$$P_J(D) = P(R|D), P_J(e) = \frac{cf(e, Top)}{|Top|}, \quad (11)$$

where λ is the probability that at any step the user decides to make a jump and not to follow outgoing links anymore, $cf(e, Top)$ is the number of top documents where the candidate e appears, $|Top|$ is the size of a result set. The described Markov process is aperiodic and irreducible (due to introduced jump probabilities), and hence has a stationary distribution. Consequently, we consider that $P(Expert|e)$ is proportional to the stationary probability $P_\infty(e)$. Although our method is computationally intensive, it converges very fast (after 200-300 iterations) for typical expertise graphs containing at most 2000 nodes according to our experiments.

3.6 Absorbing random walk

In our next approach we represent the search for an expert as *an absorbing random walk* in a document-candidate graph [45]. We calculate the probability of finding a candidate if consider that this candidate is the required expert. The candidate node which we want to evaluate is only self-transient, since we assume it to be the final destination of the walk. This means that in contrast to the one-step approach, we calculate the probability of finding a certain candidate expert by making *any sufficient number of steps* in the graph. Formally speaking, we remove all outgoing edges from the measured candidate, add the self-transition edge to it and use the following equations iteratively:

$$P_0(D) = P(R|D), P_0(e) = 0, \quad (12)$$

$$P_i(D) = \sum_{e \to D} P(D|e) P_{i-1}(e), \quad (13)$$

$$P_i(e) = \sum_{D \to e} P(e|D) P_{i-1}(D) + P_{i-1}(e) P^{self}(e|e) \quad (14)$$

Finally, we consider that $P(Expert|e)$ is proportional to the probability $P_\infty(e)$. It should be mentioned that in strongly connected graphs with an absorbing node the probability of absorption in the infinity is effectively close to 1. However, the graphs we deal with are nearly uncoupled and, de facto, the probability of absorption for a certain node depends only on the connectivity of the region it belongs to and on the total probability of relevance of closely connected document nodes.

Making the full run of iterations for each candidate is unnecessary. If we rewrite the above equations in a matrix form, we get $\mathbf{p} = \mathbf{p}_0 \mathbf{A}^i$, where $\mathbf{p}_0$ is a vector of starting probabilities, the matrix $\mathbf{A}$ consists of one-step transition probabilities and $\mathbf{A}^i$ contains probabilities of transitioning from one node to another in i steps. In our calculations we use matrix $\mathbf{B}$ containing probabilities of transitioning from each node to another in the minimum number of steps. We get this matrix by filling it with those elements from $\mathbf{A}^i$, which become non-zero after some next iteration. When no new element in $\mathbf{A}^i$ becomes non-zero after some iteration, the filling of $\mathbf{B}$ is finished. The vector of probabilities $\mathbf{p}$ used for candidate ranking is calculated as $\mathbf{p} = \mathbf{p}_0 \mathbf{B}$.

The absorbing random walk based method has several theoretical advantages over the previously presented methods. First, it can be regarded as a generalization of the baseline method described by Equation 1, considering that we regard $P(e|D)$ as the probability to transfer from the document D to the candidate e not in one step, but in the *minimum sufficient number of steps*. Second, it does not need a training phase since it is parameter-free.

3.7 Using organizational and document links

Usually, for graph-based algorithms, the introduction of new information into the analysis often comes to discovering new links among analyzed entities. The scenario of searching for expertise in the enterprise may include not only moving from relevant documents to the candidate experts found in them and vice versa, but also along document-document and candidate-candidate connections. We may find it natural that a user goes over the ranked documents by following hyperlinks. The discovery of new experts may be possible not through documents only, but also with the help of candidate experts the user is in contact with already. For example, they can send the user to their colleagues in the same department who expectedly possess similar expertise. This "escalation phase" of expertise seeking, when people end up with experts not initially recommended by a system, but related to those, even crossing organizational boundaries, is common in enterprises according to recent user studies [3].

We experimented with adding these new transitions to our expertise graph and using them for the Infinite Random Walk method. The iterations specified in Equations 9 and 10 are updated in the following way:

$$P_i(D) = \lambda P_J(D) + (1-\lambda)((1-\mu_D)\sum_{e \to D} P(D|e)P_{i-1}(e) + \\ +\mu_D \sum_{D' \to D} P(D|D')P_{i-1}(D')), \quad (15)$$

$$P_i(e) = \lambda P_J(e) + (1-\lambda)((1-\mu_e)\sum_{D \to e} P(e|D)P_{i-1}(D) + \\ +\mu_e \sum_{e' \to e} P(e|e')P_{i-1}(e')), \quad (16)$$

where μ_D is the probability of following document-document connections, μ_e is the probability of following candidate-candidate connections. The new transition probabilities are calculated as:

$$P(D|D') = 1/N_{D'}, P(e|e') = 1/N_{e'}, \quad (17)$$

where $N_{D'}$ is the number of outgoing document links from the document D' and $N_{e'}$ is the number of outgoing candidate links from the candidate e'. It is, of course, reasonable to differentiate the strength of relation and directions of influence among co-workers or even include entire organizational hierarchy into the graphical model, but we do not have this information in our data sets.

4. EXPERIMENTS

4.1 Experimental setup

We conduct our experiments with two data sets provided by the TREC community. Although both testbeds allow to realistically simulate classic expert finding scenarios, they have some clear distinctions.

W3C data, TREC 2005, 2006. This collection represents the internal documentation of the World Wide Web Consortium (W3C) and was crawled from the public W3C (*.w3.org) sites in June 2004. The data consists of several sub-collections: web pages, source code, mailing lists etc. In our experiments we use the largest (1.85 GB, 198 000 documents), the most clean and structured part of the corpus, containing email discussions within the W3C. This part is rather homogeneous in format (each document is an email of average length 450 words) and hence its features would not significantly vary over different enterprises. It also allows the accurate detection of candidate experts in documents just using their unique email addresses. The W3C data is supplemented with the list of 1092 candidate experts represented by their full names and email addresses. We experiment with 49 queries and respective relevance (expertise) judgments used for TREC evaluations in 2006, which are more reliable comparing to the queries used for the pilot TREC evaluations in 2005, when candidate experts were not judged manually.

CSIRO data, TREC 2007. The data used in TREC 2007 is a crawl from publicly available pages of another organization - Australia's national science agency CSIRO. It includes about 370 000 web documents (4 GB) of various types: personal home pages, announcements of books and presentations, press releases, publications. Instead of a list of candidate experts, only the structure of candidates' email addresses was provided: *firstname.lastname@csiro.au*. Using this as a pattern we built our own candidates list by finding about 3500 candidates in the collection. 50 queries with judgments made by retired CSIRO employees were used for the evaluation.

At the collection preparation stage, we extract associations between candidate experts and documents. For both data sets we use simple recognition by searching for candidates email addresses and full names in the text of documents. For the CSIRO documents the association scores $a(e, D)$ between documents and found candidates are set uniformly to 1.0. In the case of W3C data, we may differentiate the type of a candidate-document relation, by looking at the email field where the candidate was detected: *from*, *to*, *cc* or *body*. We use the following association scores: $a(e, D^{from}) = 1.5$, $a(e, D^{to}) = 1.0$, $a(e, D^{cc}) = 2.5$ and $a(e, D^{body}) = 1.0$ respectively, which is the most realistic combination according to recent studies of W3C 'lists' sub-collection [7]. For the initial documents ranking as well as for the graph generation the open-source PF/Tijah retrieval system [25] was employed.

The results analysis is based on calculating popular IR performance measures also used in official TREC evaluations: Mean Average Precision (MAP), precision at top 5 ranked candidate experts (P@5) and Mean Reciprocal Rank (MRR). MAP shows the overall ability of a system to distinguish between experts and non-experts. P@5 is considered more significant than precisions at lower ranks since the cost of an incorrect expert detection is very high in an enterprise: the contact with a wrong person may require a mass of time. If we consider that the user can be satisfied with only one expert on the topic (considering that all experts are always available for requests), then the performance of MRR measure becomes crucial.

In our experiments discussed below we compare our methods with a baseline to study the effectiveness of the multistep relevance propagation approach. However, for the sake of a fair comparison, we also show the performance of the simplest of known methods, called Votes in [35], which ranks candidates just by the number of top documents where they appear.

The evaluation of the following methods is discussed further:

- **Votes**: the method, ranking candidates by the number of top documents where they appear [35],

- **Baseline**: the baseline one-step relevance propagation method [6] (see Section 3.2),
- **FRW**: the multi-step relevance propagation with the Finite Random Walk method (see Section 3.4),
- **IRW**: the multi-step relevance propagation with the Infinite Random Walk method (see Section 3.5).
- **ARW**: the multi-step relevance propagation with the Absorbing Random Walk method (see Section 3.6).

The first step in any document-based expert finding algorithm is a document retrieval run extracting relevant documents for analysis. Both datasets where indexed with the use of Snowball stemmer and standard English stopwords were removed. In our experiments we use the language model based approach to IR for scoring documents (see Equations 3, 4) and retrieve a predefined number of top ranked documents, which we consider sufficient to cover the topic of a query. We retrieved only documents with at least one mention of a candidate. The optimal number of retrieved documents varied considerably over the data sets. In our preliminary experiments we had to retrieve 1500 documents from the W3C collection and just 50 documents from the CSIRO collection for the maximum performance of the Baseline method. We believe that this difference is caused by two reasons. The average number of experts per query is very small for the CSIRO collection - 3, whereas it is 60 for the W3C collection. Since only very authoritative persons were considered experts in the CSIRO, they mostly appear in the top relevant documents on a topic. Moreover, the number of candidate experts is three times higher for the CSIRO data. This means that the number of persons competing with each other increases with each next retrieved document faster, what makes the task of finding experts among them harder.

In order to achieve a denser document-candidate graph we experimented not only with persons from the candidates list, but also with other persons found in the collection considering each found email address as an identifier of an individual. The additional person entities increased the graph-sizes by far, since also documents containing a person but no candidate experts were included into the graph network as well. This graph expansion allowed us to use the non-candidate persons that are not selected for the final ranking as mediators for the relevance transmission from candidate to candidate.

4.2 Experiments with multi-step relevance propagation

Both the FRW and the IRW methods depend on one parameter. In case of the FRW method this is the number of relevance propagation steps K to be done. Figures 3 and 4 compare the MAP performance of the Baseline method with the performance of the FRW method after from 1 to 43 propagation steps for both data sets.

We see that the maximum MAP is reached after making in average 13 steps for the W3C data and 33 steps for the CSIRO data. This is not very long walk in our graph, so the relevance will not be propagated too far. This is partly true because we have a probability to stay at document nodes, but also because a typical expertise graph (see Figure 2) is nearly uncoupled and the limited scope of its (almost) disjoint subsets makes a surfer to do a lot of steps to go out of the bounds of a subset. However, the most important observation is that for both collections the increasing of the number of propagation steps from one to more also leads to MAP increase for both datasets, making this noticeable already after a few steps.

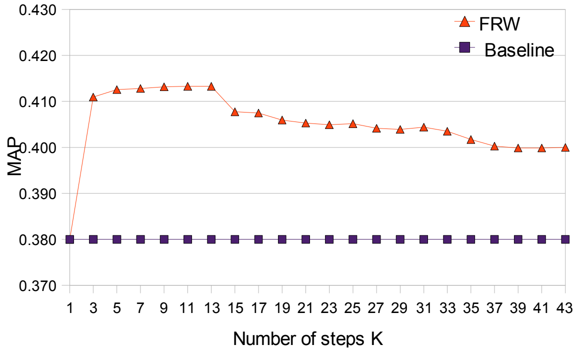


Figure 3: MAP for the Finite Random Walk method (FRW) with different numbers of propagations steps taken, W3C (2006) data

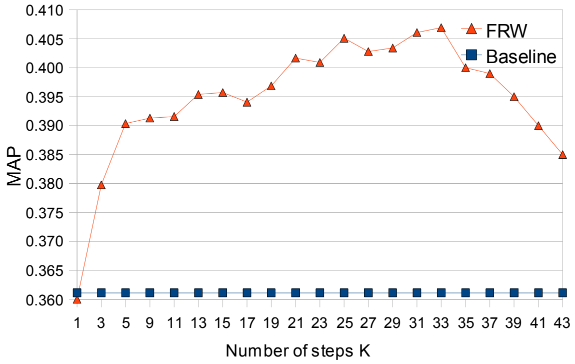


Figure 4: MAP for the Finite Random Walk method (FRW) with different numbers of propagations steps taken, CSIRO (2007) data

The only parameter to tune for the IRW method is λ - the probability of a walk restart. Empirically, in Figure 5 we see that any value starting from 0.03 and higher improves over the baseline for both datasets. However, for the W3C collection, 0.1 value gives the best result and 0.05 is the optimal setting for the CSIRO collection. These values actually mean that we restart our infinite random walk in average after 10 and 25 steps taken, what appears to be very close to the optimal numbers of steps for the finite random walk for the respective collections. Moreover, this setup of λ to the value between 0 and 0.15 is also typical for the use of random walks in web retrieval [37].

To avoid the effects of over-training for the IRW and the FRW methods in our final evaluation, we applied 5-fold cross-validation technique. We divided test queries for both collections in 5 parts and for each part trained our methods on the other 4 parts. We could also consider training on one TREC collection and testing on another. However, since the TREC data used in 2006 and 2007 is quite different (what actually allows us to conduct representative experiments), the structure, the size and the topical diversity of expertise graphs also differs considerably, so that they can not be used to tune parameters for each other. Since the ARW method is parameter-free, it could be directly applied to the entire query set without any training.

The performance of all methods for all measures is presented in Table 1. As hoped, we see that actually both Infinite and Finite Random Walk methods are equally effective and outperform the baseline method for all measures. The ARW method also shows the improvement over the baseline for all measures, but still seems inferior to the IRW and the FRW methods. To test the statistical significance of the obtained improvement with respect to the baseline, we calculated a paired t-test over both query sets for each method and each measure. Results indicated that the improvement for the MAP measure is significant for all three methods at the $p < 0.001$ level. For the MRR measure it is significant at the $p < 0.01$ level for the IRW method and at the $p < 0.05$ level for the FRW method. For the P@5 measure it is significant at the $p < 0.01$ level for the IRW and the ARW methods, and at the $p < 0.05$ level for the FRW method. The improvement we got is also comparable with the advantage of the baseline over the simplest Votes method. Having in mind that our baseline is the one of the most effective methods known, we may conclude that the improvements of our methods are significant in the area.

It is also important to mention that both methods that needed training phase showed the improvement over almost all regions of their parameter space (see Figures 3, 4 and 5) and our parameter-free method also showed the comparable improvement. Each full expert finding run for each method (including document retrieval and relevance propagation stages) can be performed in about 1 second on a desktop computer. This result suggests that the multi-step relevance propagation for expert finding is not only advantageous, but also practicable technique, which is easy to tune, resource-light and stable in performance.

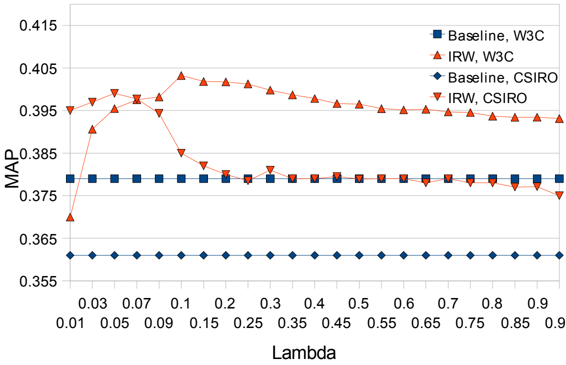


Figure 5: MAP for Infinite Random Walk method (IRW) over different values for jumping probability

4.3 Experiments with additional links

So far we considered only bipartite document-candidate graphs without links between nodes of the same type. However, the CSIRO collection allows to also include the additional information about relations among documents and candidate experts. Documents from **.csiro.au* are highly hyperlinked. Candidate experts are professionally interrelated, if they are employed in the same CSIRO department.

	W3C, 2006			CSIRO, 2007		
	MAP	MRR	P@5	MAP	MRR	P@5
Baseline	0.379	0.787	0.624	0.361	0.508	0.220
Votes	0.336	0.700	0.571	0.321	0.449	0.212
FRW	**0.413**	0.807	**0.660**	**0.407**	0.566	**0.236**
IRW	0.405	**0.810**	0.653	0.400	**0.582**	0.232
ARW	0.398	0.804	0.641	0.376	0.518	0.232

Table 1: Performance for all measures, both data sets and all tested methods

While inter-document links are easily extracted from documents since they are HTML tagged, a candidate's working department can be inferred only from the candidate's email address: the third level domain name is usually an abbreviation of a department's name. As an illustrative example, the candidate's email address *David.Dall@***ento***.csiro.au* shows that David Dall works at the CSIRO Entomology research department. We inter-link all candidates experts in the same department and also take into account the hyperlinks between documents. The experimental results shown in Figure 6 demonstrate only the benefit from adding organizational links. When we set the probability of relevance propagation between related candidate experts μ_e to 0.25, we get noticeable improvement (significant at the $p < 0.05$ level). Adding links between documents degrades the performance for almost all values of the inter-document propagation probability μ_D.

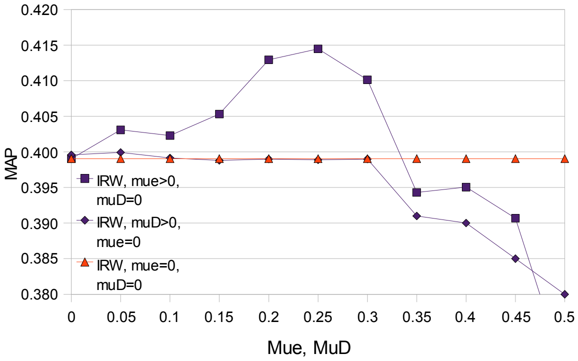


Figure 6: MAP for the Infinite Random Walk method (IRW) with additional links, CSIRO (2007) data

Intuitively, the inter-department links between candidate experts can help only within "functional" organizations[1], whose employees are highly specialized and separate units are divided by knowledge areas. In this case we may assume that people working in the same department are all experts on similar topics - otherwise, the evidence that the specific person is knowledgeable cannot be propagated to his/her co-workers. Unfortunately, the W3C data set contains neither any information about the distribution of candidate experts across organizational units nor any links between documents. There are few links due to "reply-to" relations between some emails, but in our preliminary experiments their use did not cause any change in performance. In order to prove our intuition, we make a simulation of the case described above. Using provided expertise judgments we in-

[1] `wikipedia.org/wiki/Organizational_structure`

terconnect all persons who are experts on the same topic, in order to see whether it helps to rank them higher. In other words, we test the situation when all experts on a specific topic work in the same department in the W3C. We see in Figure 7 that using simulated organizational links increases the performance of the Infinite Random Walk method for all values of μ_e with the maximum at 0.6 value (this result is significant at the p $<$ 0.01 level). This experiment shows the potential advantage of modeling professional connections among employees in the enterprises with the structure similar to the simulated.

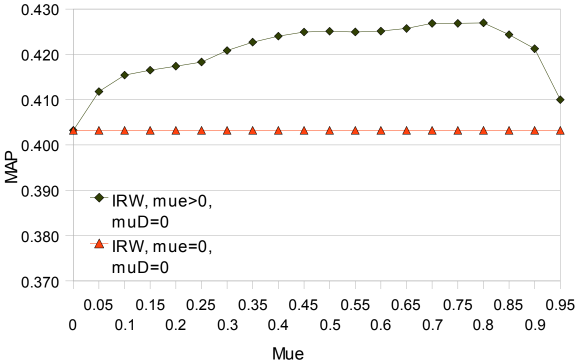


Figure 7: MAP for the Infinite Random Walk method (IRW) with simulated organizational links, W3C (2006) data

5. CONCLUSIONS AND FUTURE WORK

We have introduced a novel class of expert finding methods founded on the following twofold principle. First, it states that expert finding is a process of walking (consulting) in an *expertise graph* of candidate experts and topical documents. Second, it advocates that the relevance appearing from the documents should be propagated not once, but multiple times, further through various connections in such a graph. We showed that one of the most effective among existing methods is a special case of our approach: one-step relevance propagation on our expertise graph. Notably, experiments conducted on the data crawled from web-sites of two large organizations validated the effectiveness of our methods. We empirically demonstrated that the use of multi-step relevance propagation by different probabilistic random walk based methods sharing the same above-mentioned principle leads to significant improvements over the baseline one-step propagation. We also found the benefit of utilizing direct organizational links among candidate experts.

There are several promising directions to extend the presented research. First, it is reasonable to continue with making the random walk more relevance-directed. For example, a surfer could move from persons to documents with a probability proportional not only to their association degree, but also to their relevance. Second, new entities (e.g. dates of mailing lists) can be introduced into expertise graphs to better model the relevance flow. Furthermore, not only documents, but paragraphs or sentences may serve as sources of flowing out relevance.

Future work in this area should also include a wider use of professional connections between employees or any other knowledge about information flows in the organization. In cases when an appropriate organizational structure is not available, one can try to infer it through hierarchical clustering of persons using their pre-computed static profiles or links inferred from their e-mail communications. It is also interesting to study how the propagation of document relevance through persons to other documents may affect the performance of the initial document retrieval run.

6. ACKNOWLEDGMENTS

We want to sincerely thank Hugo Zaragoza from Yahoo! Research, Barcelona for useful discussions and extensive comments on the first versions of this paper.

REFERENCES

[1] IBM Professional Marketplace matches consultants with clients. White paper. November 2006.

[2] Enterprise search from Microsoft: Empower people to find information and expertise. White paper. Microsoft, January 2007.

[3] M. S. Ackerman, V. Wulf, and V. Pipek. *Sharing Expertise: Beyond Knowledge Management*. MIT Press, Cambridge, MA, USA, 2002.

[4] E. Agichtein, C. Castillo, D. Donato, A. Gionis, and G. Mishne. Finding high-quality content in social media. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 183–194, 2008.

[5] K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, 2006.

[6] K. Balog, T. Bogers, L. Azzopardi, M. de Rijke, and A. van den Bosch. Broad expertise retrieval in sparse data environments. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 551–558, 2007.

[7] K. Balog and M. de Rijke. Finding experts and their eetails in e-mail corpora. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 1035–1036, 2006.

[8] I. Becerra-Fernandez. Facilitating the online search of experts at NASA using expert seeker people-finder. In *PAKM'00, Third International Conference on Practical Aspects of Knowledge Management*, 2000.

[9] M. Bilenko and R. W. White. Mining the search trails of surfing crowds: identifying relevant websites from user activity. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 51–60, 2008.

[10] C. S. Campbell, P. P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 528–531, 2003.

[11] Y. Cao, J. Liu, S. Bao, and H. Li. Research on expert search at enterprise track of trec 2005. In *Proceedings of 14th Text Retrieval Conference (TREC 2005)*, 2005.

[12] H. Chen, H. Shen, J. Xiong, S. Tan, and X. Cheng. Social Network Structure behind the Mailing Lists: ICT-IIIS at TREC 2006 Expert Finding Track. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

[13] K. Collins-Thompson and J. Callan. Query expansion using random walk models. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 704–711, 2005.

[14] N. Craswell, A. de Vries, and I. Soboroff. Overview of the trec-2005 enterprise track. In *Proceedings of TREC-2005*, Gaithersburg, USA, 2005.

[15] N. Craswell, D. Hawking, A.-M. Vercoustre, and P. Wilkins. Panoptic expert: Searching for experts not just

for documents. In *Ausweb Poster Proceedings*, Queensland, Australia, 2001.

[16] N. Craswell and M. Szummer. Random walks on the click graph. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 239–246, 2007.

[17] F. Crestani. Application of spreading activation techniques in informationretrieval. *Artif. Intell. Rev.*, 11(6):453–482, 1997.

[18] T. Davenport. Knowledge Management at Microsoft. White paper. 1997.

[19] T. Davenport. Ten principles of knowledge management and four case studies. *Knowledge and Process Management*, 4(3), 1998.

[20] H. Fang and C. Zhai. Probabilistic models for expert finding. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007*, pages 418–430, 2007.

[21] L. Fields. 3 great databases for finding experts. *The Expert Advisor*, (3), March 2007.

[22] S. Harabagiu, F. Lacatusu, and A. Hickl. Answering complex questions with random walk models. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 220–227, 2006.

[23] D. Hawking. Challenges in enterprise search. In *ADC '04: Proceedings of the 15th Australasian database conference*, pages 15–24, Darlinghurst, Australia, Australia, 2004.

[24] D. Hiemstra. *Using Language Models for Information Retrieval*. Phd thesis, University of Twente, 2001.

[25] D. Hiemstra, H. Rode, R. van Os, and J. Flokstra. Pftijah: text search in an xml database system. In *Proceedings of the 2nd International Workshop on Open Source Information Retrieval (OSIR)*, pages 12–17, August 2006.

[26] M. Idinopulos and L. Kempler. Do you know who your experts are? *The McKinsey Quarterly*, (4), 2003.

[27] G. Jeh and J. Widom. Scaling personalized web search. In *WWW '03: Proceedings of the 12th international conference on World Wide Web*, pages 271–279, 2003.

[28] P. Jurczyk and E. Agichtein. Discovering authorities in question answer communities by using link analysis. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 919–922, 2007.

[29] J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, 1999.

[30] O. Kurland and L. Lee. Respect my authority!: HITS without hyperlinks, utilizing cluster-based language models. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 83–90, 2006.

[31] J. Lafferty and C. Zhai. Document language models, query models, and risk minimization for information retrieval. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 111–119, 2001.

[32] R. Lempel and S. Moran. Salsa: the stochastic approach for link-structure analysis. *ACM Trans. Inf. Syst.*, 19(2):131–160, 2001.

[33] X. Liu, W. B. Croft, and M. Koll. Finding experts in community-based question-answering services. In *CIKM '05: Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 315–316, 2005.

[34] W. Lu, S. Robertson, A. Macfarlane, and H. Zhao. Window-based Enterprise Expert Search. In *Proceeddings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

[35] C. Macdonald and I. Ounis. Voting for candidates: adapting data fusion techniques for an expert search task. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 387–396, 2006.

[36] M. T. Maybury. Expert finding systems. Technical Report MTR06B000040, MITRE Corporation, 2006.

[37] M. A. Najork, H. Zaragoza, and M. J. Taylor. Hits on the web: how does it compare? In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 471–478, 2007.

[38] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Stable algorithms for link analysis. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 258–266, 2001.

[39] L. Page, S. Brin, R. Motwani, and T. Winograd. The pagerank citation ranking: Bringing order to the web. Technical report, Stanford University, 1998.

[40] D. Petkova and W. B. Croft. Proximity-based document representation for named entity retrieval. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 731–740, 2007.

[41] M. Richardson and P. Domingos. The intelligent surfer: Probabilistic combination of link and content information in pagerank. In *NIPS '01: Advances in Neural Information Processing Systems*, 2001.

[42] P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *ECIR*, pages 309–320, 2008.

[43] P. Serdyukov, D. Hiemstra, M. Fokkinga, and P. M. G. Apers. Generative modeling of persons and documents for expert search. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 827–828, 2007.

[44] P. Serdyukov, H. Rode, and D. Hiemsta. Exploiting sequential dependencies for expert finding. In *SIGIR '08: Proceedings of the 31th annual international ACM SIGIR conference on Research and development in information retrieval*, 2008.

[45] P. Serdyukov, H. Rode, and D. Hiemsta. Modeling expert finding as an absorbing random walk. In *SIGIR '08: Proceedings of the 31th annual international ACM SIGIR conference on Research and development in information retrieval*, 2008.

[46] A. Shakery and C. Zhai. A probabilistic relevance propagation model for hypertext retrieval. In *CIKM '06: Proceedings of the 15th ACM international conference on Information and knowledge management*, pages 550–558, 2006.

[47] X. Song, B. L. Tseng, C.-Y. Lin, and M.-T. Sun. Personalized recommendation driven by information flow. In *SIGIR '06: Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 509–516, 2006.

[48] K. Toutanova, C. D. Manning, and A. Y. Ng. Learning random walk models for inducing word dependency distributions. In *ICML '04: Proceedings of the twenty-first international conference on Machine learning*, page 103, 2004.

[49] T. Tsikrika, P. Serdyukov, H. Rode, T. Westerveld, R. Aly, D. Hiemstra, and A. de Vries. Structured Document Retrieval, Multimedia Retrieval, and Entity Ranking using PF/Tijah. In *INEX 2007*, 2007.

[50] H. Zaragoza, H. Rode, P. Mika, J. Atserias, M. Ciaramita, and G. Attardi. Ranking very many typed entities on wikipedia. In *CIKM '07*, Lisbon, Portugal, 2007.

[51] J. Zhang, M. S. Ackerman, and L. Adamic. Expertise networks in online communities: structure and algorithms. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 221–230, 2007.