# SEARCH FOR EXPERTISE

## GOING BEYOND DIRECT EVIDENCE

by

Pavel Serdyukov

PhD dissertation committee:

*Chairman and secretary*

| | |
|---|---|
| Prof. dr. ir. A. J. Mouthaan | (Universiteit Twente) |

*Promotor*

| | |
|---|---|
| Prof. dr. P. M. G. Apers | (Universiteit Twente) |

*Assistant promotor*

| | |
|---|---|
| Dr. ir. D. Hiemstra | (Universiteit Twente) |

*Members*

| | |
|---|---|
| Prof. dr. D. Hawking | (Australian National University) |
| Prof. dr. T. W. C. Huibers | (Universiteit Twente) |
| Dr. I. Ounis | (University of Glasgow) |
| Prof. dr. M. de Rijke | (Universiteit van Amsterdam) |
| Prof. dr. R. J. Wieringa | (Universiteit Twente) |


CTIT Ph.D. thesis Series No. 09-144
Centre for Telematics and Information Technology
P.O. Box 217 - 7500 AE Enschede - The Netherlands


SIKS Dissertation Series No. 2009-22
The research reported in this thesis has been carried out under the auspices of SIKS, the Dutch Research School for Information and Knowledge Systems.

Printed and bound by Ipskamp Drukkers B.V.

ISBN 978-90-365-2845-0
ISSN 1381-3617; (CTIT Ph.D. thesis Series No. 09-144)
http://dx.doi.org/10.3990/1.9789036528450

# SEARCH FOR EXPERTISE
# GOING BEYOND DIRECT EVIDENCE

**DISSERTATION**

to obtain
the degree of doctor at the University of Twente,
on the authority of the rector magnificus,
prof.dr. H. Brinksma,
on account of the decision of the graduation committee,
to be publicly defended
on Wednesday the 24th of June at 13:15

by

**Pavel Serdyukov**

born on the 25th of May 1980
in Volgograd, Russia

This dissertation is approved by:

Prof. dr. Peter M. G. Apers (promotor)

Dr. ir. Djoerd Hiemstra (assistant-promotor)

# Contents

# Acknowledgements

It would not have been possible to complete this work without the support of all the kind people surrounding me during the last four years. Thanks are due to all of them and some deserve a special mention.

First of all, I would like to express my deep gratitude to my promoter, Peter Apers, who always had confidence in me and my capability to do research. His advice, as well as his unfailing optimism and patience, made it possible for me to go through the uneasy first year of my life as a PhD student and retain my enthusiasm till the very end.

I am greatly indebted to my daily supervisor, Djoerd Hiemstra, who gave me a helping hand when I needed it so much. Since then I have always been inspired by his, indeed, everyday willingness to discuss my work and his invaluable input into the papers that we co-authored and this thesis.

I am very thankful to Maarten Fokkinga who also supervised me during a short, but one the most decisive periods of my career. Our meetings did an important job in forming my research vision.

There are many other people who contributed to my work at Twente. Special thanks go to Henning Rode, whose team spirit helped us be so prolific in publications. It was a great pleasure to work together with Robin Aly, Arthur van Bunningen, Sander Evers, Harold van Heerde and my office mate Rongmei Li. I am happy to admit that my life would not be so enjoyable without the generous support of Ida den Hamer-Mulder, Suse Engbers and Sandra Westhoff. I wish to extend my heartfelt thanks to all current and former members of the Database group for the unique friendly and cozy working atmosphere. That is why, I feel grateful to Gerhard Weikum who not only helped me to do my first steps in academic research, but also advised me to join this group, and to Ling Feng who offered me the job and also supervised me in the first year.

Besides my work in the Netherlands, I enjoyed a few months working at

Yahoo! Research and living in Barcelona. I want to thank Ricardo Baeza-Yates, Vanessa Murdock and Roelof van Zwol for this dreamlike opportunity. I am especially thankful to Vanessa for her unstoppable passion toward our project that eventually led us to a great result and her daily support inside and outside the lab. I was also very happy to share this time with Aris, Mia, Hugo, Borkur, Katja, Adam, Antonio, Michele and other great lab mates.

I am honored that David Hawking, Theo Huibers, Iadh Ounis, Maarten De Rijke and Roel Wieringa agreed to join my dissertation committee. Many thanks to David and Maarten for their comments and suggestions for the final version of my thesis.

My dearest friends always helped me to make my life more worth living in one way or another: Natalie, Sergey, Olga, Yana, Pavel. My sincere and warm thanks to my friends from Deurningerstraat, Makandra and other places for saving me from boredom. I would really like to hug many more close friends from Moscow, Volgograd, Munich and other cities and villages.

Last, but not least, I am inexpressibly thankful to my parents and grandparents for their love and sacrifice. And Alisa, my wife, for turning my life into a sequence of wonderful dreams coming true.

Pavel Serdyukov
Enschede, 15 May 2009

# 1

# Introduction

## 1.1 The task of automated expert finding

In large enterprises people often search not only for relevant documents, but also for their colleagues that know something on the topic of their information need (Hertzum and Pejtersen, 2000). Sometimes the required knowledge is just not freely accessible in digital format. It might not be considered important enough to be digitized and stored or it can be hardly expressible in written language. In these cases asking other people becomes the only way to find an answer (Craswell et al., 2001). Those people who are able to satisfy certain information needs, give correct answers to specific questions, explain them and even guide the user further to other sources of relevant information are usually called *experts*. 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, recruiters for talented employees, even consultants for other consultants to redirect inquiries and decrease the risk of losing clients (Idinopulos and Kempler, 2003).

The need in finding a well-informed person may be critical for any kind of project. However, any attempt to identify experts by manual browsing through organizational documents or social networks may fail in very large enterprises, especially when they are geographically distributed. A standard text search engine may be of great help, but still is not able to fully 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 (Maybury, 2006). It allows to either save time and money on hiring a consultant when a company's own human resources are sufficient, or 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. For ensuring traceability, the system usually returns not only the ranking of people, but also a list of evidences that indicate each person's expertise (e.g. summaries of relevant documents related to the person) (Hawking, 2004).

Expert finding (also known as expertise search, expert recommendation, expertise location or expertise identification) inherits a lot from document retrieval and information filtering tasks, and is thereby traditionally regarded as a subject of research on Information Retrieval. It is also often considered that we should restrict the scope of our search to the experts who are all employees of the same organization. Despite that this limitation is not an obvious requirement for this task, expert finding is a part of the functionality of a typical Enterprise Search system, which usually operates within the scope of a single company. It is also important to distinguish between expert search and the search for someone whom users know or vaguely remember, like a celebrity or a classmate. Popular examples of people search engines are `Spock.com`, `pipl.com`, `people.yahoo.com`. Typically, these search systems ask for the name of a person, although some keywords describing the person's interests or expertise may be used for disambiguation of persons with similar names. Note that an expert finding system aims to find *any* person with the certain knowledge, even though the restriction of the search to a specific subset of people is possible. Disambiguation of personal names also adds up to the complexity of this task, but still is not usually regarded as a primary concern.

**The need for indirect expertise evidence** Finding an expert is a challenging task, because *expertise* is a loosely defined concept which is hard to formalize. It is common to refer to expertise as to "tacit knowledge" (Baumard, 2001), the type of knowledge that people carry in their minds and which is, therefore, difficult to access. It is opposed to "explicit knowledge", which is already captured, described, documented and stored. An expert finding system aims to assess and access "tacit knowledge" in organizations by finding a way to it through artifacts of "explicit knowledge". It analyzes organizational documents in order to find some evidence about the expertise of the people they mention. Its final goal is to help people discover and transfer knowledge that otherwise would stay unused, and hence stimulate their "socialization". According to this mission, a meeting planning component is often viewed as a functional requirement for an expert finding system (Serdyukov et al., 2008a).

It is however unclear what amount of personal knowledge should be con-

sidered enough to name somebody "an expert". It depends not only on the specificity of the user query, but also on characteristics of the respective expertise area: on its age, depth, and complexity. This means that the time spent to become a world-class expert in *Java Serialization* is probably only enough to gain a beginner level in *Atomic physics*. Despite that fact, however, expert finding systems usually do not infer the actual level of expertise or any quantitative estimate that may be easily explained or mapped to a qualitative scale. They just provide some estimate that may be used to rank people by their expertness on a topic. These estimates are often not even comparable across topics. As a result, a serious, but hardly avoidable drawback of the existing systems is that they recommend people in any case, even when there is no one in the organization who merely deserved to be named an expert on the topic. However, this issue is common for most tasks of ranked retrieval (Hawking and Robertson, 2003).

The relevance of organizational documents explicitly related to the person is usually used as *direct evidence* for personal expertise. This may include documents authored by the person, e.g. his/her publications, emails, forum messages, resumes, home pages, and even personal query logs (Maybury, 2006). However, other documents that just mention the person are also regarded as primary sources of direct expertise evidence. Since document relevance can be estimated only with high uncertainty, estimates of personal expertise are often not less uncertain. Even directly related content is not always a reliable evidence, since it may, for instance, contain discussions, showing the interest of involved people, but not their competence. We can also suppose that people might become authors of documents not only because of their direct contribution to the content, but due to some other kind of relation to their co-authors (e.g. if they are project managers or participated in related discussions at some point) (Bennett and Taylor, 2003). In other words, the direct relation to sources of topical information often implies personal experience from participation in topical activities, but not necessarily deep expertise on the topic.

The high uncertainty of direct evidence suggests that a possible way to improve our guess about someone's expertise is to increase the amount of evidence by also taking *indirect evidence* into account. This includes organizational documents that are implicitly related to candidate experts, their colleagues and documents found outside of the organization. This thesis proposes several ways to deal with direct evidence, but, most importantly, it studies the utility of indirect evidence that can be found within the organization and outside.

**From expertise identification to expertise selection** The list of issues relevant to expert finding research is not reduced only to the inference of personal expertise. A practically usable expert finder should help not only to *identify knowledgeable people*, but also to *select the most appropriate experts* among them for a face-to-face contact (Ackerman et al., 2002). Since expert finding is a tool for improving organizational communication, it must be able to predict various features of a planned communication to help it be successful.

In the first place, it should assume that the communication should be physically doable. So, the availability and interruptibility of experts that may depend on their location and/or workload should be considered. Sometimes, an intelligent meeting planner, taking into account agenda records of several employees including the user and predictions for their future location, is required.

Second, it must estimate communication skills of persons along with their expertise. The knowledge exchange is often hardly reachable due to cultural or language differences, or due to lack of communication and presentation skills of an expert. The ability to present own work is always consonant with a talent to give and explain answers and may be inferred, for example, from the frequency of public talks.

An expert finder should also try to predict whether the communication is likely to be desired by both parts. Various human factors like expert's mood or mental stress may be considered. Preferences of experts and users on communication with certain people (e.g. based on their positions/ranks or reputation in a company) should also be integrated. Most of the above-mentioned issues are the topic of the dedicated research on pervasive and ubiquitous computing, assuming that personal context can be inferred from measurements made by sensors of various types (Fogarty et al., 2005).

However, we consider that our work is orthogonal to the line of research described above. We envision a system with a decision function taking all kinds of evidence into account to provide the final ranking of candidate experts. The goal of this thesis is to improve the quality of the informed guess about personal expertise, assuming that other features are observed with sufficient confidence. The influence of each feature may be set by an automatic machine learner, the system administrator or inferred from explicit preferences of a specific user.

## 1.2 Research objectives

The research presented in this thesis attempts to find new ways to improve performance of expert finding systems. We seek for better understanding of how to extract the direct expertise evidence for a person from the organizational documents where he/she is mentioned, and how to utilize indirect evidence from the organizational documents implicitly related to the person and also those documents that can be found outside of the enterprise.

We describe various techniques of automatic expertise inference from organizational documentation, intra-organizational social networks and the World Wide Web. First, we propose a novel way of expertise evidence extraction from documents where the person is mentioned. Second, we show how to utilize features of organizational network consisting of documents and employees to gather additional evidence. Third, we explain how to combine local organizational evidence with the evidence acquired from the Web. Finally, to validate proposed methods we demonstrate how to utilize similar techniques for the following related tasks: entity search in Wikipedia and location ranking for placing Flickr images on a map. To achieve the results demonstrated in this thesis, we pursued the following research objectives and answered a number of associated research questions.

**RO1: Going beyond independence of terms and experts** State-of-the-art expert finding methods infer personal expertise using sources of direct documentary evidence. Some of them, often the most effective ones, aggregate probabilities of relevance of organizational documents mentioning the person to get an estimate of personal expertise. They assume that the more often the person is detected in the documents containing many words describing the topic, the more likely we may rely on this person as an expert on this topic. However, these methods also consider that persons as well as terms occur in the document independently and do not influence the appearance of each other. Although, the assumption about independence among terms is a de facto standard in probabilistic approaches to IR (Crestani et al., 1998), the independence of terms from persons does not seem obvious. Topical words and persons mentioned in the text are entities of quite different nature and often appear in the text for different purposes. In this respect, we sought to answer the following research questions.

Does the assumption about dependence of terms and persons in a document lead to better performance of expert finding methods measuring the degree of their co-occurrence? How to model this dependence and estimate its strength? How to use the assumption of dependence to infer expertise?

**RO2: Going beyond the scope of directly related organizational documents** Most expert finding methods aggregate direct expertise evidence arising from the organizational documents mentioning candidate experts and do not notice indirect sources of expertise evidence. Particularly, they ignore the evidence that can be found by following implicit links between documents and persons. In other words, they do not propagate relevance probabilities further than to persons explicitly mentioned in the documents, even though persons and documents relevant to a query can be represented as a directed graph with paths of different length. In order to compensate for this inconsistency in approaches, we attempted to find the answers for the following research questions.

What sources of expertise evidence in the organization, besides those documents that mention the person, can be used for estimation of the personal expertise? Should we stop after the first step of relevance probability propagation from retrieved documents to directly related candidate experts? How to model multi-step relevance propagation in a graph of documents and persons?

**RO3: Going beyond the scope of the organization** While the intranet of an organization still should be regarded as a direct source of expertise evidence for its employees, the amount and quality of supporting organizational documentation is often not sufficient. At the same time, leading people search engines, such as `Zoominfo.com` or `wink.com` claim that none of their information is anything that one could not find on the Web (Arrington, 2007). Neglecting expertise evidence which can be easily found within striking distance is not practical. Consequently, our research implied answering the following questions:

What information sources outside of the organization are useful for finding experts? What measures can be used to get high-quality estimates of expertise from these sources? Is there any benefit in combining direct organizational and indirect web-based expertise evidences?

**RO4: Going beyond the scope of the expert finding task** Expert finding is an example of a task estimating the relevance (expertise) of an entity (person) when it has no explicit textual description or when such a description is incomplete. Since, there are tasks besides expert finding that encounter similar problems, it was promising to expand our research focus and to apply similar techniques and principles to other applications of the same type. Particularly, we tried to find the answers to the following questions by studying two tasks: entity ranking in Wikipedia and placing images on a map using

their user-generated descriptions.

Do other applications benefit from the principles used to develop expert finding algorithms? To what degree should solutions be adapted for related tasks?

## 1.3 Thesis outline

The structure of this thesis follows the above-described research objectives. It starts from breaking the widely popular assumption about independence of persons and terms in documents. At its next step, it alleviates two restrictions: one stating that the evidence should be searched only in directly related documents and another one limiting the analysis only to documents hosted within the enterprise employing candidate experts. It finally arrives to the point which demonstrates how similar principles and techniques can be adopted in related tasks. The thesis is organized into the following chapters.

The next chapter describes state-of-the-art research on expert finding and related tasks. Chapter 3 presents our expert finding methods using the assumption of dependence between terms and persons in a document. Origins of this work can be found in (Serdyukov et al., 2007b; Serdyukov and Hiemstra, 2008a; Serdyukov et al., 2008c). Chapter 4 explains how to utilize indirect expertise evidence in the enterprise. The material used in this chapter is taken from (Serdyukov et al., 2007c, 2008b,d). Chapter 5 describes several ways to go outside of the enterprise and search for expertise evidence on the Web. They were initially proposed in (Serdyukov and Hiemstra, 2008b; Serdyukov et al., 2009a). Chapter 6 demonstrates how to apply similar and other task-specific techniques to applications resembling expert finding: entity ranking in Wikipedia and ranking locations for placing Flickr images on a map. The first part of this chapter is published as (Tsikrika et al., 2007), the second part is submitted as a patent and also published as (Serdyukov et al., 2009b). Chapter 7 concludes the thesis with an overview of its contributions and recommends directions for future research.

# 2

# State of the Art

In this chapter, we give an overview of a number of effective and sophisticated expert finding methods known from academic publications, and, to show a complete picture, we also describe industrial solutions promoted in media and independent market studies. Later on, we draw and illustrate an insightful analogy between approaches popular in expert finding research and a series of related information retrieval technologies. Finally, we describe evaluation standards and test collections traditionally used by researchers on the topic of this thesis.

## 2.1 Expert Finding in research

Expert finding systems compelled close attention of the IR community only recently, but a large amount of work has been already done. In this section we describe the most cited approaches that we classify into four categories: profile-based, document-based, window-based and graph-based methods.

### 2.1.1 Profile-based expert finding

Early pioneering approaches to expert finding could be classified as *profile-based* (Craswell et al., 2001; Liu et al., 2005b). This technology was the first step in full automation of expert finding in organizations and generally aimed to avoid manual maintenance of personal profiles (resumes or home pages). In these approaches, all documents related to a candidate expert are merged into a single personal profile prior to the actual retrieval process. The proof of the relation between a person and a document can be an authorship or just the occurrence of personal identifiers (e.g. full names or email addresses) in the text of documents located in the organizational intranet: e.g. we may

consider external publications, descriptions of personal projects, sent emails or answers in message boards. Resulting personal profiles are ranked like documents with respect to user queries using standard text similarity measures and corresponding best candidate experts are suggested to the user.

However, a number of advanced profile-based approaches, using latest progress in text retrieval research, have been suggested. Streeter and Lochbaum (1988) proposed to solve the task of finding the organization with the highest expertise by, first, building profiles using all organizational documentation and, second, applying latent semantic indexing techniques to profile-term vector space. Balog et al. (2006) (Model 1) followed language modeling approach to IR and ranked candidate experts by the probability of generating a query by the language model of a candidate's profile. In their model, each term generated by the profile language model is produced by one of documents used to create the profile with the probability that the candidate expert is actually related to the document. Later, they suggested using *topical profiles* and ranked candidate experts by the richness of their profiles in topics expressed in a query (Balog and de Rijke, 2007b). Petkova and Croft (2006) grouped documents by type/format and weighted the contribution of documents from each group to a candidate's profile.

#### Document and Profile-based query expansion

Regarding expert finding as a task of profile retrieval motivated the application of advanced document retrieval techniques for further improvement. Among them, query expansion techniques using the top retrieved expert profiles are the most popular. First, Macdonald and Ounis (2007a) applied Divergence From Randomness (DFR) based weighting of terms from the top profiles to select a few expansion terms. Serdyukov et al. (2007a) suggested massive query expansion approach through representing the query as a mixture of document- and profile-specific relevance language models. Later, Macdonald and Ounis (2007b) measured topical cohesiveness of an expert profile (averaged similarity of individual documents to the entire profile) to select coherent profiles for expansion and hence avoid topic drift. Alternatively, they used only those documents included in expert profiles that were already highly relevant to the topic. However, the latter approach looks similar to the classic query expansion from a document collection (Lavrenko and Croft, 2001), since most relevant documents in most relevant profiles will supposedly be among the top retrieved documents from the collection anyway. Query expansion from organizational documents is actually popular in expert finding research and appears in works of Petkova and Croft (2006) and Balog et al. (2008b)

### 2.1.2 Document-based expert finding

Early *profile-based* approaches demonstrated that it is reasonable to consider that the more often a person is mentioned in documents rich in query terms, the higher chance that the person has some knowledge on the query topic. However, as it became clear later, it is better to analyze the co-occurrence of personal identifiers and topical words on the lower, and hence less ambiguous levels: within the scope of documents or text windows. Our confidence in that the piece of text containing many query terms is relevant should be inversely proportional to its size (i.e. proportional to the density of topical words). And the chance that a candidate mentioned in the text actually relates to its relevant part also increases if we consider smaller text fragments, or at least not all related documents as a single fragment. Consequently, the follow-up *document-based* approaches proposed to analyze the content of each document separately and let their individual relevance probabilities add up to the probability of expertness of related persons.

#### Language model based expert finding

One of the most cited document-based expert finding methods was simultaneously proposed by Balog et al. (2006) and Cao et al. (2005). It is based on the probabilistic language modeling (LM) principle of IR and considers that the expertise of candidate person $e$ with respect to the query $Q$ is proportional to the probability $P(Q, e)$:

$$P(Q, e) = \sum_{D \in Top} P(Q|D)P(e|D)P(D) \tag{2.1}$$

where $P(Q|D)$ is the probability of the document $D$ to generate the query $Q$, which is assumed proportional to the unknown probability that the document $D$ is relevant. $P(e|D)$ is the probability of association between the candidate $e$ and the document $D$. $Top$ is the set of documents retrieved and the prior probability $P(D)$ is distributed uniformly over the $Top$. $Top$ can be unbounded or limited to the predefined number of top ranked documents, selected by rank or relevance probability.

The probability of the query to be generated by the document language model, considering independence assumption about term generation, is expressed as:

$$P(Q|D) = \prod_{q \in Q} P(q|D), \tag{2.2}$$

where the product is taken over all individual occurrences of query terms. Term generation probabilities are estimated as:

$$P(q|D) = (1 - \lambda_G)\frac{c(q, D)}{|D|} + \lambda_G \frac{\sum_{D' \in C} c(q, D')}{\sum_{D' \in C} |D'|} \tag{2.3}$$

where $c(q, D)$ is the term count of $q$ in the document $D$, $|D|$ is the document length and $\lambda_G$ is a Jelinek-Mercer smoothing parameter - the probability of a term to be generated from the global language model calculated over the entire collection $C$ (empirically set to 0.8 in our experiments). This specific kind of smoothing outperformed Bayesian smoothing in their experiments with profile- and document-based models (Balog et al., 2006).

Document-candidate association probabilities are calculated empirically using the following equation:

$$P(e|D) = \frac{a(e, D)}{\sum_{e'} a(e', D)}, \tag{2.4}$$

where $a(e, D)$ is the non-normalized association score between the candidate $e$ and the document $D$ proportional to their strength of relation (in most cases, to the importance of a field where the candidate occurred).

Balog's document-based method (often referred by authors as *Method 2*) not only outperforms profile-based methods according to their evaluation (where their own profile-based method is named *Method 1*), but is also based on theoretically-sound language model based information retrieval framework. These circumstances motivated us to use this approach as a baseline in the majority of experiments described in this thesis.

#### Data fusion based expert finding

Although, this thesis pays particular attention to the previous model as a baseline in our empirical comparisons, another class of similar methods also deserves extensive mention. Macdonald and Ounis (2006) proposed a number of data fusion based methods slightly deviating from the principle of summing relevances of documents related to a candidate. For example, one method summed not relevance scores, but *reciprocal ranks* of documents (RR), another one, called Votes, just used *the number of documents with the person mention in the top*. Methods using *minimum, maximum and average of relevance scores* were also evaluated. BM25 weighting model (Robertson et al., 2000) was used to measure relevance scores of documents. Later Macdonald et al. (2008a) suggested to cluster documents related to a candidate, measure relevance of these clusters and sum reciprocal ranks of clusters for each candidate. They also utilized query-independent evidence of document relevance in the organizational collection: URL path length (inversely proportional to relevance) and the number of inlinks (directly proportional to relevance).

### 2.1.3 Window-based expert finding

Some recent works attempted to avoid propagation of relevance of those document 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 (150-200 words) surrounding the person's mention was considered (Lu et al., 2006). Balog and de Rijke (2008) later expanded this model to consider windows of various sizes at the same time, weighted by their importance. In another approach, the partial relevance of each query term instance found in a document contributed to the probability of a candidate's expertness, but proportionally to its word distance from the nearest mention of the candidate in this document. Different distance functions have been applied and some of them lead to window-based approaches (Petkova and Croft, 2007).

### 2.1.4 Graph-based expert finding

It is important to mention another line of research that proposed finding experts by measuring their centrality in organizational or public social networks. These approaches often ignore the relevance of content related to candidate experts and utilize documents only as the context establishing relations between candidates based on the fact of their co-occurrence. Sometimes they are even designed as query independent measures of prior belief that a person is authoritative within some knowledge community and therefore able to answer questions on topics popular in the community. It seems that while for very specialized communities this assumption seems plausible, there is no guarantee that central users from multidisciplinary knowledge networks are "know-it-alls"

First, Campbell et al. (2003) compared the HITS algorithm (Kleinberg, 1999) against a simple document-based approach, similar to the *Votes* method (see Section 2.1.2) on email corpora from two different organizations. The directed social graph was created using e-mail headers and *from/to* fields - so, contained only persons as nodes and e-mails as edges. Using HITS (only *authority* scores) for candidate ranking resulted in better precision, but lower recall than for the simple method. Zhang et al. (2007) analyzed a large highly specialized (in Java programming) help-seeking community in order to identify users with high expertise. The social graph was built from post/reply user interactions with edges directed from questions to answers to reward answering activity. Three measures were compared: answers/questions ratio and graph centralities: HITS and PageRank. The former measure outperformed centralities what meant that answering questions of those users who answer a lot themselves is not an activity indicating high expertise. Another

study demonstrated that rankings produced by both HITS and PageRank are inferior to the ranking by a standard document-based method (Chen et al., 2006). This result is especially relevant and significant to our research, since authors also experimented with TREC 2006 data (W3C corpus, only mailing lists) used in this thesis as well.

Finding experts in topic-focused communities or for random topics expressed in user queries is a more novel and complex task than the long known problem of finding authoritative people in large social networks. For instance, the authority (citation index) of scientists in co-authorship networks is traditionally defined by centrality measures: closeness, betweenness, Markov centrality (PageRank) etc. (Liu et al., 2005a). These measures do a good job for tasks of identifying globally important social actors, so not necessarily active in the scope of a certain topic. The illustrative example is the task of finding influential bloggers (see Section 2.3.3). However, as was already mentioned, these approaches are not known to be successful in query-dependent expert finding scenarios for which it is hard to detect a well-developed and homogeneous social community on the topic of each possible query.

## 2.2 Real-world Expert Finding

### 2.2.1 Commercial expert finding systems

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 (Davenport, 1997, 1998; Becerra-Fernandez, 2000). They basically represented repositories of employee profiles with simple search functionality. These profiles contained a summary of personal knowledge, skills, affiliations, education, and interests, as well as contact information. Surprisingly expert finding is not considered an integral part of enterprise search systems nowadays. This situation seems to be the consequence of the current immaturity of the enterprise search market and should improve with the expected growth of competition in the future (Owens, 2008).

However, expert finding services can be found within the enterprise search frameworks of major vendors. Autonomy (`www.autonomy.com`), the undisputed market leader, provides a classic expert finding service with a feature to use a document as a query. FAST (`www.fastsearch.com`), the runner-up, does not provide its own solution, but supports AskMe (`www.askme.com`), a company that develops an expert finder on the top of the FAST platform. Their expertise search engine is not fully automatic: AskMe expects users to upload personal documents to the server on their own for profiling purposes.

However, AskMe considers the workload aspect: it enables experts to specify the number of questions that they are willing to answer per day. FAST is acquired by Microsoft in April 2008, and Microsoft recently started to advertise their expert finding solution built on the basis of their Office and Outlook products. Microsoft Knowledge Network is a profile based expert finder using personal emails as the primary expertise evidence (Microsoft, 2007). It recommends those experts who are found in proximity to the user in the organizational social network. Endeca (`www.endeca.com`), the third enterprise search market leader, does not offer a "plug and play" expert search engine, but with their powerful entity search technologies, an expert finder can be rapidly developed on the client side. Its Guided Navigation technology also allows to refine a query by specifying different aspects extracted from the initially returned result: experts' position in a company or their areas of expertise. The famous case of using Endeca's expert finder in IBM (where it was called "Professional Marketplace") is described by Maybury (2006).

Among specialized tools for expert finding, three are several well-known and worth mentioning: Tacit Illumio (`www.illumio.com`), Triviumsofts SEE-K (`www.triviumsoft.com`) and Recommind Mindserver (`www.recommind.com`). Illumio, in contrast to the traditional centralized approach (Craswell et al., 2005a), accounts on the distributed arrangement of the data in the enterprise. Its client monitors personal desktop, extracts expertise evidence from its content and serves as a filter for incoming requests for expertise that are intelligently disseminated by the central server to user desktops. Mindserver provides advanced faceted search and query refinement capabilities: it groups experts by a project or location and shows keywords representing aspects of their expertise. SEE-K can be also distinguished for its extraordinary approach to result visualization: each expert's skills are represented as a tree with most characteristic skills placed closer to the root and minor skills depicted as leaves.

Almost all above-described solutions provide highly intelligent expertise identification functionality (although in terms of effectiveness they are not known to be ahead of research community (see Section 2.1)) and many of them offer powerful ways to represent and manually navigate search results. However, while dedicated expert finders and expert search systems with expertise location functionality are of a great help to improve organizational communication and knowledge flow, they are still too far from providing a complete and tolerable solution. According to recent surveys, only 55 percent of professional services employees and a mere 27 percent of public sector employees are able to locate expertise using their current enterprise search systems (Recommind, 2009). Moreover, there are still no applications that would assist users at each step of expertise sharing and acquisition, as it

is envisioned in early research on expert finding (McDonald and Ackerman, 1998; Johansson et al., 2000).

### 2.2.2 Free on-line expert finding

Some large-scale free on-line people search (`www.spock.com`) and expert finding (`www.zoominfo.com`) systems are already quite well-known in consultancy business (Fields, 2007). The ZoomInfo's PowerSearch offers a search over 34 million business professionals and 2 million companies across virtually every industry. Analogous specialized search engines exist for journalists (`www.profnet.com`) and lawyers (`www.expertwitness.com`). Some on-line resume databases (`www.monster.com`) and social networking systems (`www.linkedin.com`) are often used as expert search engines for recruiting purposes (King, 2006; Kolek and Saunders, 2008).

Advanced academic search engines already allow search for people who are influential in the certain research area. Google Scholar shows key authors for the topic in addition to the ranking of relevant publications. The Community of Science (`cos.com`) contains profiles of more than 480,000 experts from over 1,600 institutions worldwide. Besides searching, it can be used for browsing a hierarchically organized expertise taxonomy.

## 2.3 Related tasks

Many tasks where we basically try to search or analyze people and their artifacts have a lot in common with expert finding. Thus, it comes as no surprise that some techniques used in novel application domains look like inspired by expert finding approaches, while some other methods deserve to be regarded as their predecessors. We give a detailed review of allied problems and their state-of-the-art solutions in this section.

### 2.3.1 Finding similar experts

Balog and de Rijke (2007a) identify the task of *finding similar experts in an organization* and propose a simple ranking solution based on measuring overall similarity of candidate experts to those specified in the *example list*. Similarity is measured by Jaccard coefficient measuring the magnitude of overlap between sets of documents related to compared candidates. In their follow-up work they use various query-independent measures of personal reputation, popularity and social activity to recommend only top-notch experts (Hofmann et al., 2008).

There is a line of research on *link prediction in social networks* that by implication strongly relates to the above-described task and people recommendation tasks in general (Liben-Nowell and Kleinberg, 2003). Its methods for measuring similarity between two graph nodes are not limited to *common neighbors* based approach employed by Balog and de Rijke (2007a) for finding similar experts. Alternatively, methods calculating *hitting time* for two nodes (expected number of steps required for a random surfer to reach one node starting from another) (Jeh and Widom, 2002) or clustering nodes (Cadez et al., 2000) allow to also take high-order co-occurrences into account.

### 2.3.2 Finding experts at question/answering portals

A clever way to find experts is to find a place where these experts not only dwell and willingly share their expertise, but also regularly get evaluated by other experts or regular users. Community-driven question/answering portals are the places where people ask questions, give answers and vote for the best of them. The most popular example is Yahoo! Answers [1], the largest community of its kind nowadays with a market share approaching 100%. The typical research problem of finding the best answer for a question resembles expert finding, if we consider features of answerers only, not looking at features of answers.

The history of a user activity at the portal, namely the number of questions and answers are often used to predict the quality of fresh answers given by that user. According to recent studies, the most predictive content-independent feature of answer quality is the ratio of previously given and promoted (selected as best by askers or collectively by user votes) answers of the answerer (Agichtein et al., 2008). Furthermore, focusing on a particular category correlated with obtaining best ratings for answers in categories where questions centered on factual or technical content (e.g. *Programming*) (Adamic et al., 2008). Dom and Paranjpe (2008) suggested a number of Bayesian smoothing techniques using overall population statistics to get a better estimate of the above-mentioned probability that a randomly selected answer from the user history is also the best one for the corresponding question. The HITS algorithm on the user-answer graph was also utilized recently and authority (or hub) score of a user was proposed as a better predictor of new answers' quality (Jurczyk and Agichtein, 2007).

Surprisingly, the language model of user expertise, mined from the history of questions/answers was never used to predict the quality of new answers. Moreover, while these systems explicitly (Zhang et al., 2007) or implicitly

[1] `answers.yahoo.com`

(Adamic et al., 2008) search for experts in the community, their analysis is always query/question independent and actually akin to finding the most helpful, demanded and active users, but not necessarily experts in very specific topics. From the other hand, being successful in these communities does not necessarily mean to readily provide top-quality expertise. According to Adamic et al. (2008) only 1% of questions posted at Yahoo!Answers requires expertise level above average.

### 2.3.3 Finding influential/relevant blogs

The challenge of finding trend setters in Blogosphere is one the most intriguing in web-based social analysis (Agarwal et al., 2008). Despite that a significant number of blogs are maintained by communities, and not by individuals, it is obvious that they represent a collection of documents (posts) containing firsthand evidence about expertise of their author(s). Since blogs are web-sites, it comes as no surprise that their influence is often determined by classic authority measures, Indegree, HITS or PageRank, measuring how often blog posts are cited by reputable sources. It is also suggested to consider not only popularity, but also novelty of posted stories, since many influential posts start a trend only after being re-posted by already famous bloggers (Song et al., 2007).

The search for blogs relevant to a query (Ounis et al., 2008) not only looks similar to expert finding, but even borrows its ready-made solutions. Elsas et al. (2008) evaluate both profile-based (all posts are merged into a "virtual" document) and document-based (each post relevance is measured separately) expert finding approaches. In the latter case, the contribution of each post relevance to the overall blog relevance is made proportional to the similarity between the blog and the post. Seo and Croft (2008) evaluates these methods and also proposes own hybrid approach based on clustering posts within a blog and summing relevance of these clusters. Actually, authors of both papers claim that their solutions are adopted not from research on expert finding, but on distributed information retrieval, which is reviewed in the next section.

### 2.3.4 Resource selection

Resource (database, server, or collection) selection is a well-known IR problem with first solutions published in mid 90's (Callan et al., 1995; Voorhees et al., 1995; Gravano, 1998). It appeared first as a task of web search service selection by a metasearch engine (Selberg and Etzioni, 1995; Dreilinger and Howe, 1997) and then grew into an independent sub-topic of research on

Distributed IR dealing with thousands of autonomous collections, e.g. nodes of a Peer-to-Peer web search engine (Chernov et al., 2005). Since it is impossible to forward a query to all databases, they are usually pre-ranked by their potential to return relevant document in response to a query, based on aggregated statistics of their documents. There are two principal approaches assuming that collections are either *cooperative*, i.e. voluntarily sharing their aggregates, or *uncooperative*, so permitting only their sampling with queries (Craswell, 2000).

Resource selection in a cooperative environment comes to ranking collections as single documents and hence has much in common with profile-based expert finding (see Section 2.1.1). While methods just merging collection documents to get aggregated statistics are among earliest known (Zobel, 1997; Xu and Croft, 1999), the most popular approach uses a task-specific *tfidf* approach. Document frequency in the collection is used instead of the sum of term frequencies, and inverted document frequency is approximated by inverted collection frequency (Callan et al., 1995). However, the most effective method for uncooperative environment copies (samples) a part of the collection and then sums the relevance of these documents w.r.t. to a query to rank collections (Si and Callan, 2003). It bears a great resemblance with document-based expert finding approaches, if we do not consider the fact that only a sample of documents is used (see Section 2.1.2).

### 2.3.5 Information filtering/routing

In some situations users approach the expert finding system not with a short query, but with a thorough description of their expertise need. A particular case of such a scenario is an automatic search for reviewers, when a conference management system assigns knowledgeable researchers to review articles using a submitted paper or a conference description as a text pattern describing the meaning of the appropriate expertise. Traditionally, expertise of reviewers (candidate experts) is described by their profiles mined from the documents they authored. The task then comes to matching a paper to these profiles using either vector-space (Dumais and Nielsen, 1992; Hettich and Pazzani, 2006) or probabilistic approaches (Karimzadehgan et al., 2008).

A similar task is intelligent message addressing, i.e. finding potential recipients of a chat/email message. This technology becomes indispensable when users constantly receive a lot of impersonal emails or newsletters from their employer, although not being able to unsubscribe from them because of the company's rules. A message addressing (at the sender's side) or filtering (at the recipients's side) mechanism learns a user model from the user's personal data (e.g. sent emails) and uses it to provide binary classification

of messages. This approach is similar to spam filtering with the classifier of legitimate emails trained on the user's personal data. Recently, it was demonstrated that traditional expert finding methods, although not being the best possible solution, are able to successfully cope with this task (Carvalho and Cohen, 2008).

### 2.3.6 Faceted search

Faceted search is an information aggregation technology for structuring and visualizing search results. It aims to facilitate the interaction between a user and the search system by helping the navigation through results and consequent query refinement (Hearst, 2006; Knabe and Tunkelang, 2007). The most popular related technology focuses on presenting retrieved documents in groups (facets) each corresponding to a single document feature and further grouping documents within each facet by feature values. These features might be extracted from a document's metadata (e.g. date of issue, owner, purpose) or inferred from its content (e.g. topics, sentiments or real-world entities mentioned in it). If a faceted search system is able to group documents along such facets as "authors", "personalities" or "employees", then it is reasonable to think of it as of an expert finding system.

However, there was little research done on ranking facet-value pairs, so basically most faceted search interfaces output them in alphabetical order. Recently, it was also proposed to order facet-value pairs either by the number of documents where each of them appear, or by the correlation (e.g. measured by the pointwise mutual information score) between the probability of association of a facet-value pair with the document and the probability of its relevance (topicality) (Koren et al., 2008). Both approaches, if applied to the "people" facet, closely resemble document-based expert finding methods (see Section 2.1.2): the Votes (Macdonald and Ounis, 2006) and the language model based methods respectively (Balog et al., 2006).

### 2.3.7 Entity ranking

Expert finding can also be regarded as a specialized entity ranking task with restriction of search to only entities of such types as "people" or "employees". In general, many more kinds of entities are usually mentioned in documents and hence can be searched by matching their context to a query. This context can be often specified explicitly, e.g. as an article in an encyclopedia. Searching with graph-based methods for typed entities (*images, dates, phone numbers* etc.) on the Web was explored recently in several publications (Cheng et al., 2007; Zaragoza et al., 2007). An entity ranking track started in 2007

at the INEX workshop [1]. Using the Wikipedia collection, where entities are described by Wiki-articles and highly interlinked, INEX evaluates the search for entities by limiting the result set for each query to a specific entity type (category). We propose our own solution to this task in Chapter 6 based on expert finding techniques we proposed in Chapter 4.

## 2.4 Evaluation standards

The expert search task is a part of the Enterprise track of the Text REtrieval Conference (TREC) since its first run in 2005 (Craswell et al., 2005a). It is also the only enterprise search task being run each year since then until 2008. The TREC community created experimental data sets consisting of organizational document collections, lists of candidate experts and sets of search topics, each with a list of actual experts. The evaluation measures were borrowed from the text retrieval tasks and applied to the submitted ranked lists of candidate experts as otherwise for documents. We also review a number of alternative collections with interesting properties.

### 2.4.1 TREC 2005-2006: W3C corpus

The collection used in Enterprise Track of Text REtrieval Conference (TREC) in 2005 and 2006 represents the internal documentation of the World Wide Web Consortium (W3C) and was crawled from the public W3C (*.w3.org) sites in June 2004. As shown in Table 2.1, the data consists of 331,037 documents from several sub-collections: web pages, source code, mailing lists etc. Not the entire data is useful - for instance, the *dev* part is rarely used despite its size. While there are not so many near-duplicates in the *lists* part, only about 60,000 e-mails are single messages and the rest of them belongs to about 21,000 multi-message threads (Wu and Oard, 2005). In contrast, *www* part contains a lot of "almost near-duplicates", e.g. revisions of the same report document describing W3C standards and guidelines.

The W3C data is supplemented with a list of 1092 candidate experts represented by their full names and email addresses. Two quite different sets of queries were used by participants. In 2005, 50 queries were created using names of working groups in W3C as titles and members of these groups were considered experts on the query topic. Judgments were therefore binary, 1 for experts (members) and 0 for non-experts (non-members). In 2006 the TREC community collectively and manually judged each candidate for each

[1] http://inex.is.informatik.uni-duisburg.de/2007/xmlSearch.html

| Part | Description | # docs | size(GB) |
|---|---|---|---|
| lists | public e-mails | 198,394 | 1.8 |
| dev | source code | 62,509 | 2.6 |
| www | web pages | 46,975 | 1.0 |
| esw | wiki | 19,605 | 0.18 |
| other | miscellaneous | 3,538 | 0.05 |

**Table 2.1**: Summary of W3C collection

of 49 developed queries using the provided list of *supporting* documents for each candidate. *Supporting* meant that such a document is on the query topic to some extent and mentions the candidate. The judgment scale was not binary and participants could mark candidates not only as experts and non-experts, but also as "unknown" when they were not sure which category a candidate belongs to. While queries from 2006 allow to reproduce a classic expert search scenario, queries from 2005 actually simulate the search for sub-groups within an organization (a search for any person in the group working on the query topic problem). However, since judgments are made without human expert opinion about knowledgeability of candidates, it is rather unclear if they make realistic evaluations possible.

### 2.4.2 TREC 2007-2008: CSIRO corpus

The collection used in Enterprise Track of Text REtrieval Conference in 2007 and 2008 represents a crawl of publicly available pages hosted at the official web sites (about 100 `*.csiro.au` hosts) of Australia's national science agency (CSIRO), done in March 2007 (Bailey et al., 2007b). The collection, often referred as CERC (CSIRO Enterprise search collection), contains 370,715 documents with a total size of 4.2GB. There is no official division into sub-collections, but according to Jiang et al. (2007) about 89% of documents are HTML pages, 4% are pdf/word/excel documents and the rest is a mix of multimedia, script and log files. At least 95% of pages have one or more outgoing links as reported by Bailey et al. (2007a).

TREC 2007/2008 participants were provided not with a list of candidates, but with only a structural template of email addresses used by CSIRO employees: *firstname.lastname@csiro.au* (e.g. *John.Doe@csiro.au*). Thus, in order to build the list of CSIRO employees through extracting their e-mail addresses from the corpus, most participants had to get around spam protection, check if similarly looking addresses belong to the same employee and filter non-personal addresses (e.g. *education.act@csiro.au*). While such an ap-

proach makes the expert finding task more complex, it is doubtful whether it becomes more realistic. Usually, all employees are registered with a staff department and hence it should be possible to automatically inquire for the list of current employees and avoid recommending those who have left the company.

The topic set used in 2007 was created with the help of CSIRO's Science Communicators. Their everyday responsibilities include interacting with industry groups, government agencies, media and the general public. Sometimes, they actually act as expert finders on demand, since often questions they answer are requests for employees with specific knowledge. Organizers asked about 10 science communicators to develop topics in areas of their expertise. That resulted in 50 queries, each supplemented with a few "key contacts" - the most authoritative and knowledgeable employees on the query topic. On average, the number of key contacts per topic was 3 (from 1 to 11) and 152 in total. The primary requirement was that topics should be broad and important enough to deserve a dedicated overview page at the CSIRO web-site. While it was unknown whether the collection actually contains any evidence of expertise for the proclaimed experts, the realism of experimental setting certainly increased comparing to previous years when experts were elected by non-experts (participants). In 2008 topic descriptions were created again with the help of science communicators, but judgments where made by participants in the same way as in 2006.

### 2.4.3 Other collections

The UvT Expert collection is the most popular among alternative datasets and developed using public data about employees of Tilburg University (UvT), the Netherlands. The total collection size in XML format is 380MB and contains information (in English and Dutch) about 1168 experts. This often includes a page with contact information, research and course descriptions and publications record (full text of 1,880 publications is available). In addition, each expert details his/her background by selecting expertise areas from a list of topics. Balog et al. (2007) suggested to use 981 of these topics which have both English and Dutch translations.

There are a few less acknowledged collections used once for expert finding. Hogan and Harth (2007) describe an expert finding test collection made of DBLP and CiteSeer databases containing abstracts of computer science publications. Authors crawled them, integrated and converted into RDF format what results in the corpus of 18GB size including 715,690 abstracts. Demartini and Niederee (2008) proposed the task of finding experts using only the data from personal desktops. The data was gathered from desktops

of 14 users (researchers) in November 2006. The collection included 48,000 items of 8GB size, mainly e-mails, publications, address books and calendar appointments, as well as *desktop activity logs* (Chernov, 2008). All participants developed queries, related to their activities, and performed search only for people mention in documents from their own desktops. Demartini (2007) examines the task of finding experts in Wikipedia and suggests two ways of using it. First, for finding world-known personalities described by Wiki-pages under categories *People* or *Living_People*. Second, for finding experts among ordinary users contributing to Wiki Community, considering the text and semantic markup of their contributions.

### 2.4.4 Performance measures

In accordance to the tradition established by TREC community, expert finding methods are evaluated in exactly the same way as document retrieval systems. It is reasonable, since the quality of rankings can be estimated independently of what we rank if quality measures for individual items are alike. As long as expert judgments for candidate experts are binary as relevance judgments for documents, the same evaluation strategy can be applied.

The following performance measures are standard for TREC official evaluations and also used to evaluate the methods proposed in this thesis. Usually the macro-average of these measures over all queries in a test set is used to compare expert finding systems. Note that instead of *documents* we talk about *candidate experts* or just *candidates* and in place of *relevant documents* we refer to *experts*.

- **Precision at $K^{th}$ rank**: probability to find an expert by picking a random candidate from those with ranks lower or equal to $K$. In other words, it is the share of experts among top $K$ ranked candidates.
- **Average Precision**: probability to find an expert by first taking an expert randomly and then picking a random candidate among those with ranks lower or equal to the rank of the initially selected expert. In other words, it is the average of precisions calculated at ranks that the system assigned to experts,
- **Reciprocal Rank**: the inverted rank of the highest ranked expert. In other words, it is the precision calculated at the rank of the highest ranked expert.

In most of our comparisons we rely on Average Precision as on the primary performance indicator, since it measures the overall capability of a

system to distinguish between non-experts and experts, even when they appear deep down the ranking. However, it is clear that the critical demand for high precision at low ranks distinguishes users of expert finding systems even in comparison to users of web search engines. The cost of a false recommendation in expert search is much higher than in web search: a conversation with an ignorant person or even reading documents supporting the incorrect system's expertise judgment takes much longer time than taking a glance at a single irrelevant web page. By similar reasons, measures purely based on recall are used on quite rare occasions in expert finding research.

# 3

# Beyond independence of terms and experts

One of the most popular assumptions leveraged by many expert finding methods states that the expertise of a person should correlate with the co-occurrence of personal identifiers (such references as full names, e-mail or home-page addresses) and topical terms in organizational documents (Westerveld, 2006). According to this belief, the more often a person is mentioned in documents containing many topical terms, the higher the chance that this person actually has some knowledge on the topic. However, expert finding methods using the above assumption also consider that persons as well as terms occur in documents independently and do not influence the appearance of each other. Although, the independence of terms in documents and queries is accepted as a standard in probabilistic information retrieval models (Crestani et al., 1998), mainly due to performance advantages of such simplifications, the independence of terms from persons given the document is not so obviously grounded and needs re-thinking.

In this chapter, we answer research questions posed in the beginning of this thesis and related to **Research Objective 1** (see Section 1.2). We propose two models that break the assumption of independence between terms and candidate experts. The first model claims that the occurrence of terms in the document may be explained by the presence of candidate experts. We propose a method regarding people as generators of the relevant document's content. Our generative modeling combines the features of both so-called profile- and document-based approaches: it ranks candidates using their language models built from the retrieved documents, but also takes the frequency of candidate's mentions in the top ranked documents as supporting evidence of his/her expertise on the search topic. Our second model does not strictly assume that people generate the content of documents they are

mentioned in, but tries to capture the strength of association between the document's relevance and persons by looking at how their personal identifiers are positioned in the document relative to positions of query terms.

## 3.1 Person-centric expert finding

The key approaches to expert finding discussed in Chapter 2.1 state that the level of personal expertise can be determined by the aggregation of document scores related to a person. However, their intuition is generally based on measuring the co-occurrence degree of query terms and personal identifiers within the context of topical documents. It indeed seems reasonable to think that good experts should be mentioned in documents containing many query terms. This assumption is also supported by the fact that expanding the query usually leads to the improvement of not only document retrieval, but also expert finding (Macdonald and Ounis, 2007a).

In probabilistic terms, effective expert finding often comes to the estimation of the joint probability $P(e, q_1, ..., q_k)$ of observing the candidate expert $e$ together with query terms $q_1...q_k$ in a sample generated by language models of documents $D$ from the set of top ranked documents $Top$. For instance, the document-based model by Balog et. al. (see Section 2.1.2) defines this joint probability as:

$$P(e, q_1, ..., q_k) = \sum_{D \in Top} P(e|D)(\prod_{i=1}^{k} P(q_i|D))P(D) \tag{3.1}$$

As we may notice, this model assumes the independent generation of all query terms and the candidate by a document. The assumption of independent generation of terms by document language models is widely accepted. For example, the above mentioned model looks similar to the popular query expansion method by Lavrenko and Croft (2001), also assuming term independence, if only one regards the candidate expert $e$ as a candidate term for expansion. However, persons mentioned in the document are often responsible for its content, either explicitly as authors, or implicitly as recipients. We may also think of candidate experts as strong indicators for a document topic, even when they are just mentioned in the text. For example, the reference to an information source often contains a person's name what implies that the person is the source of terms (or, at least, ideas) mentioned in the document around his/her name.

We demonstrate two graphical models that we compare in this work on Figure 3.1. So, while according to a typical document-based method, a doc-

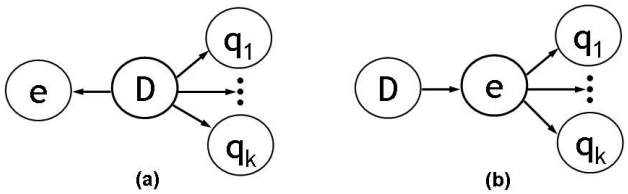



**Figure 3.1**: Dependence networks for two methods of estimating $P(e, q_1, ..., q_k)$

ument has its own unique document language model that produces terms for that document (see Figure 3.1a), in our method, (see Figure 3.1b), a document does not have the language model, but requests candidate experts to generate its terms using their personal language models. Note that we still consider that the global (collection) language model is also partly responsible for generating terms in a document.

This section is further organized as follows. In the next section, we show how to utilize the assumption that persons mentioned in a document influence the generation of terms it consists of. In Section 3.1.2, we explain how personal language models can be mined from retrieved documents and used further to predict the quality of personal expertise. Experimental results supporting our assumptions are presented in Section 3.1.3. Now, we define our person-centric model formally.

### 3.1.1 Making persons responsible

The person-centric method, which is the main contribution of this chapter, can be viewed as a hybrid method combining the features of both document- and profile-based methods (see Section 2.1). It builds its prediction by analyzing the top retrieved documents and summarizing the expertise evidence found. However, the estimation of a personal language model (see Section 3.1.2) becomes the crucial step in this prediction.

Our approach is based on the assumption of dependency between the query terms and a candidate. We suppose that candidates are actually responsible for the generation of terms within retrieved documents. According to the model presented in Figure 3.1b, we calculate the required joint probability as follows:

$$P(e, q_1, .., q_k) = \sum_{D \in Top} P(q_1, .., q_k|e)P(e|D)P(D) = P(q_1, .., q_k|e) \sum_{D \in Top} P(e|D)P(D) \tag{3.2}$$

where $P(q_1, ..., q_k|e)$ is the probability of generating the query from the personal language model of the candidate $e$. It reflects the amount of relevant

knowledge of the candidate. The sum in the right part of this formula can be considered as a prior probability $P(e)$ that one can name the candidate $e$ an expert on the topic even without looking at the term distribution in his/her personal language model:

$$P(e) = \sum_{D \in Top} P(e|D)P(D), \tag{3.3}$$

which basically measures the influence/activity of the candidate in the topic area. It is proportional to the frequency of appearance of the candidate in the topical documents. We take a ranked document prior to be inversely dependent on the document rank: $P(D) = 1/rank(D)$ in order to distinguish the importance of a document in covering the aspects of the query topic. In our experiments we also show the performance with uniformly distributed $P(e) = 1/m$, where $m$ is a number of candidate experts in the system and hence such a prior does not affect the ranking. We could also consider not ranks, but actual scores as in Equation 3.1. However, since we do not use this prior solely for ranking, but multiply it with another probability, it was necessary to avoid strong variation of its estimate across queries and make it not score-, but only ranking-dependent.

We also consider that query terms occur independently given a candidate expert, what results in:

$$P(q_1, ..., q_k|e) = \prod_{i=1}^{k} P(q_i|e) \tag{3.4}$$

Now we present our algorithm of mining for personal language models from the top retrieved documents.

### 3.1.2 Mining for personal language models

As we see, the personal query term generation probabilities $P(q_i|e)$ are the only estimates we miss so far. Of course, we can calculate them in the way similar to the one which profile-based methods use: merge those retrieved documents that relate to the person $e$ into a single profile document and calculate corresponding maximum likelihood estimates of term generation probabilities (see Section 2.1.1). However, it would be justifiable if there was only one person per document. Since we have already postulated that all candidates may be responsible for generating query terms in the documents they are mentioned in, such approach would give us only a very rough approximation of personal language models in most cases. Another quick solution may be to measure the "share" of document term frequency that

should contribute to personal language models using the probability $P(e|D)$ as proposed by Balog et al. (2006). However, in this case, we do not account for the fact that some persons may appear in a document accidentally and hence their responsibility for the document's content should not be inferred without the knowledge about their occurrence in other topical documents.

Guided by these considerations, we represent a document from the set $Top$ of retrieved documents as a mixture of personal language models and the global language model. In formal terms, we define the likelihood of top ranked documents as:

$$\prod_{D \in Top} \prod_{w \in D} ((1 - \lambda_G)(\sum_{i=1}^{m} P(e_i|D)P(w|e_i)) + \lambda_G P(w|G))^{c(w,D)} \tag{3.5}$$

Here $e_1, ..., e_m$ are the persons occurring in the documents from the $Top$, $c(w, D)$ is the count of term $w$ in document $D$, $(1 - \lambda_G)$ is the probability that a term will be generated from one of the personal models and not from the global language model. $\lambda_G$ controls the ability of the algorithm to build personal models which are discriminative only for the terms which are topic-specific. Those terms which have high probability in the collection in total will get low generation probabilities over all persons. We could also think of "hidden" persons, or leave the assumption that there still exists the document language model, also responsible for terms generation. However, we relied on a somewhat generic model expressing the principle that we suppose allows to better distinguish among candidate experts by simply breaking the independence assumption.

Our approach to candidate experts modeling is partially inspired by the similar hypothesis used in pseudo-relevance feedback method for document retrieval by Zhai and Lafferty (2001). It claims that the model of relevance behind a user query can be mined from the top retrieved documents, if only we consider them as mixtures of relevance (unknown) and global (known) language models (see Figure 3.2 (left side)). The significant difference is that we define the relevance model as a mixture of personal models of candidate experts mentioned in documents on the topic (see Figure 3.2 (right side)). These people, we suppose, actually hold and share the knowledge (relevance) which can potentially satisfy the user information need.

Actually, any personal language model acquired from top ranked documents is query-specific and hence only one of many the person would use in different contexts. If it was necessary to analyze the entire collection in the same way, we could estimate a much more detailed personal term distribution. However, we assumed that candidates should be judged by their ability

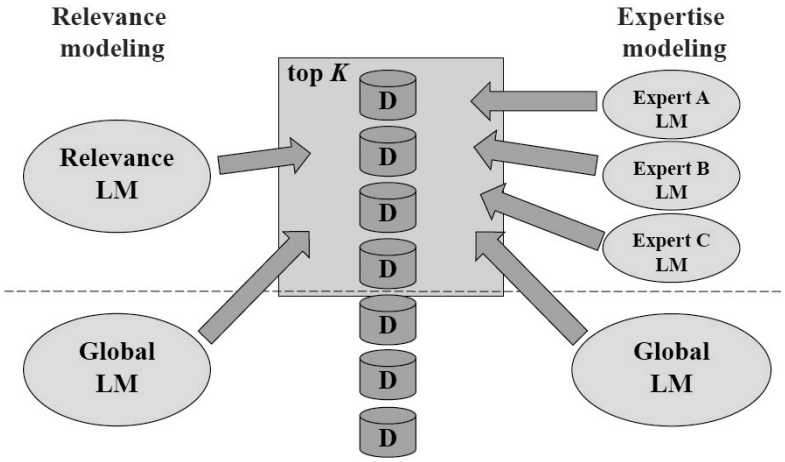



**Figure 3.2**: Relevance vs. Expertise modeling

to "say something" on topics related to the query. In other words, we are interested only in the language model the person uses for generating documents that cover the query topic to some extent, so it is reasonable to get it dynamically: at the query execution time from retrieved documents. Moreover, it is even safer to bound the analysis to the scope of top documents, since otherwise the ambiguity of personal language models may increase dramatically not being topic-specific.

### Using fixed personal contribution probabilities

Considering that all parameters, including P($e_i|D$) are given, we are able to calculate the maximum likelihood estimates of term generation probabilities from personal language models $P(w|e_i)$. In order to do that, we apply the EM algorithm (Dempster et al., 1977), traditionally used to estimate unknown parameters. We propose the following formulas updating likelihood of the document set *Top* (see Equation 3.5) to be used recursively for its maximization:

E-step:

$$P(e|w,D) = \frac{(1-\lambda_G)P(e|D)P(w|e)}{(1-\lambda_G)(\sum_{i=1}^{m} P(e_i|D)P(w|e_i)) + \lambda_G P(w|G)} \tag{3.6}$$

M-step:

$$P(w|e) = \frac{\sum_{D \in Top} c(w, D) P(e|w, D)}{\sum_w \sum_{D \in Top} c(w, D) P(e|w, D)} \tag{3.7}$$

#### Measuring personal contribution

So far we relied on the assumption that the probability $P(e|D)$ is fully determined by the type of the association between $e$ and $D$, and the number and the types of associations of other candidates with the document. This practically means that if we have a document with the probability distribution $P(e|D)$, then for another document with the same number of persons mentioned in the same way, the probability $P(e|D)$ will be distributed likewise. In other words, neither the content of the document, nor any preliminary information about candidates will influence that distribution. However, in our method we extract not only personal language models, but also probability distributions $P(e|w, D)$, which show who is the most probable generator of the term $w$ in the document $D$. It allows us to estimate the probability of contribution for each person $e$ also based on the document's content and on our current knowledge about the language model of $e$. For that purpose, we no more fix probabilities $P(e|D)$ and calculate them at every M-step of EM algorithm presented in Section 3.1.2 as follows:

$$P(e|D) = \frac{1 + \sum_{w \in D} c(w, D) P(e|w, D)}{m + \sum_{i=1}^{m} \sum_{w \in D} c(w, D) P(e_i|w, D)}, \tag{3.8}$$

where $m$ is the number of candidate experts extracted from the retrieved documents in total, used here for the purposes of Laplace smoothing.

### 3.1.3 Experiments

#### Experimental setup

For the evaluation we utilized the W3C corpus - the data from the expert search task in the Enterprise track of the TREC used in 2005 and 2006 - and its largest (1.8 GB) 'lists' part containing discussions within the W3C consortium. We focus our experiments with only this part of the collection for several reasons specified below.

At first, this part has a standardized format (emails of average length 450 words) what means that its properties and hence our conclusions should not change significantly across different enterprises, at least for the data of specific kind. Besides that, the nature of TREC 2005 topics implies that for

each topic there is a working group with exactly the same name. W3C working groups are organized to finally produce standards/solutions for the topic and the result of their work is described in reports almost always containing members of these groups (who are experts according to given judgments) as main authors and contributors. This means that using these report documents (the corpus often contains many revised versions of the same report) makes such simulation rather unrealistic. Topics created for TREC 2006 are seemingly inspired by the same approach, although they do not directly use names of working groups. Moreover, it is possible to find persons in the text of e-mails (e-mail headers) by just using unique email addresses that are much less ambiguous than personal names. Since, we do not apply sophisticated personal name disambiguation techniques, it was important to avoid uncertainty in determining person-document relations wherever possible. Finally, since email addresses occur in specific email fields in most cases, we are able to differentiate the types of person-document relations and hence fairly compare two personal language model mining approaches: one using prior knowledge about field importance and another one dynamically estimating probability of personal contribution $P(e|D)$ (see Section 3.1.2).

TREC also provides a list of 1092 candidate experts with supplemented full names and email addresses. Experiments were conducted by considering only these candidates as persons in our person-centric model. We also tested inclusion of other person entities by taking any unique email found in the collection as a new person id. This caused only a small degradation of performance, probably due to the rapid increase in the number of competing and sometimes non-existing candidates with each new document retrieved, so we do not report these results here. We provide results separately for two sets of TREC queries with relevance judgments: used in 2005 (50 queries) and in 2006 (49 queries). The data is parsed with the Snowball stemmer and indexed using Java and the Lucene open-search engine.

#### Results discussion

First of all, we accomplish recognition of candidate experts by looking for their email addresses in *from*, *to*, *cc* and *body* email fields. We additionally search for candidates in a *body* field using their full names. Different types of associations are weighted using the following values: $a(e, D^{from}) = 1.5$, $a(e, D^{to}) = 1.0$, $a(e, D^{cc}) = 2.5$ and $a(e, D^{body}) = 1.0$, what is the setting similar to the one used in recent studies of the 'lists' subcollection (Balog and de Rijke, 2006). If some person appears in several fields, only its highest association score is considered. Note that we did not omit candidates from the list of experts that do not occur in the 'lists' part of the W3C corpus.

The standard language model based IR approach, as defined in Equations 2.2 and 2.3, was used for the retrieval of documents.

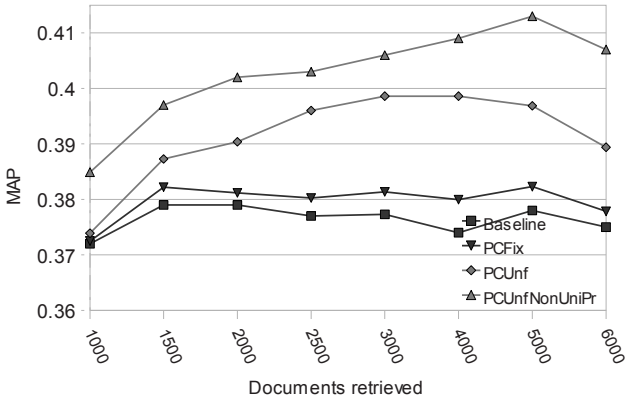
In the first place, we wanted to test the quality of the mined personal language models and decide on whether they are sufficient for efficient expert finding (see Section 3.1.2). We start from presenting the performance of our methods considering that person's priors $P(e)$ are uniformly distributed and then using non-uniform priors, as defined by Equation 3.3, with the best of them. So, the following methods are evaluated:

- **Baseline**: the baseline document-based method (see Equation 3.1),
- **PCFix**: the person-centric method using fixed person-document association scores and uniform personal priors (see Equations 3.6, 3.7),
- **PCUnf**: the person-centric method using unfixed dynamically calculated association scores and uniform personal priors (see Equations 3.6, 3.7, 3.8),
- **PCUnfNonUniPriors**: the person-centric method using unfixed dynamically calculated association scores and non-uniform personal priors (see Equations 3.6, 3.7 3.8 and 3.3).

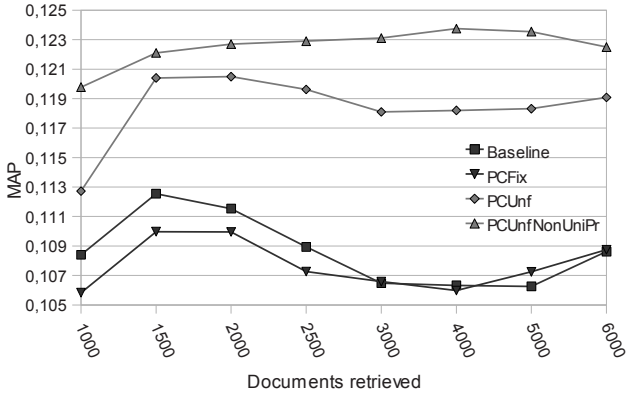
We have only two parameters in all models including the baseline model: $\lambda_G$, used in Equation 3.6 and the number of retrieved documents. Different values for $\lambda_G$ between 0.1 and 0.9 showed negligible differences in performance, but 0.8 was slightly better than others. The second parameter was much more influential. It is always rather unclear how many top documents describe each query topic to the sufficient extent. So, a good algorithm should be robust to the size of a query result set. We vary its size from 1000 to 6000 of top ranked documents. We analyze the performance using the classic IR evaluation measures: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and Mean Precision at top 5 ranked candidates (P@5) (see Section 2.4.4). We show MAP, P@5 and MRR values for both sets of queries in Figures 3.3, 3.4 and 3.5 respectively.

We see that **PCFix** method performs slightly better than **Baseline** on average. For the MAP and MRR measures the positive difference is not obvious: **PCFix** is better only in half of the cases. However, its advantage is clearly visible for P@5 (see Figure 3.4) for both query sets.

Moreover, **PCUnf** method shows better performance than both **Baseline** and **PCFix** methods on all measures/queries, especially for MAP (but not so notably for MRR). It demonstrates that query-specific and purely




(a)

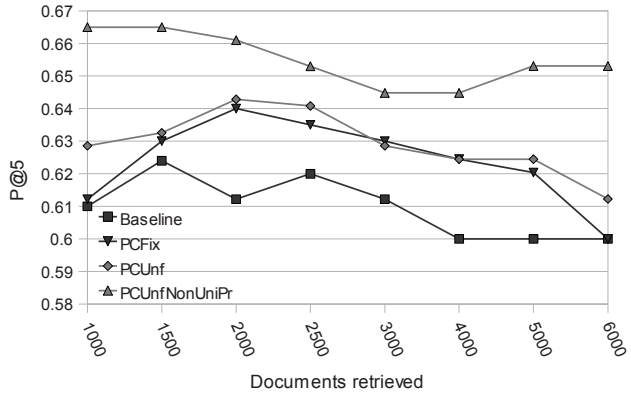



(b)

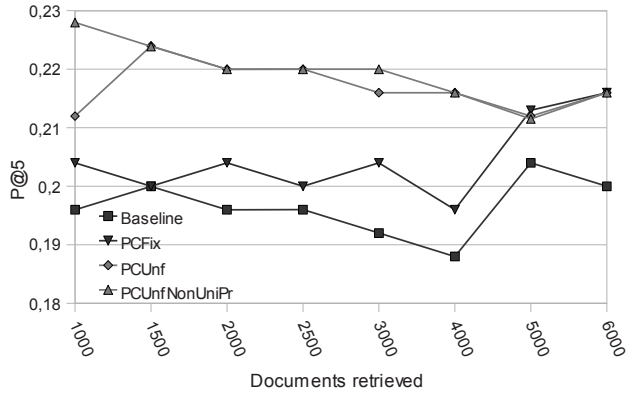
**Figure 3.3**: MAP over different numbers of documents retrieved, for the queries from 2006 (a) and for the queries from 2005 (b)

content-based estimation of personal contribution to the document is beneficial in personal language modeling.

Moreover, using non-uniform priors $P(e)$, as in Equation 3.3, with **PCUnf** method (**PCUnfNonUniPriors** method) improves performance even further for all MAP and P@5 measures at almost all numbers of retrieved documents. The frequency of participation in discussions on the topic is of course a significant evidence of personal expertise. However, from a statistical point of view, this prior penalizes the score of those candidates whose models are



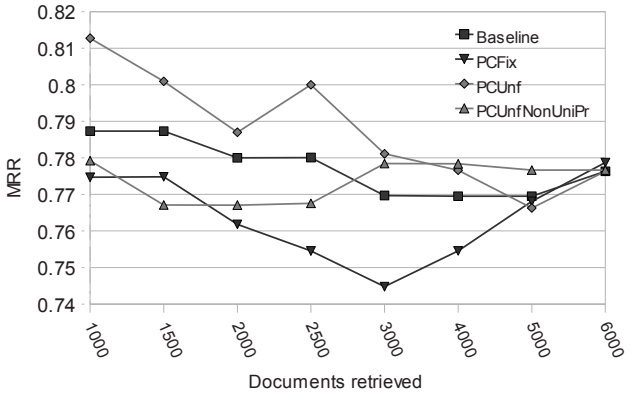

(a)



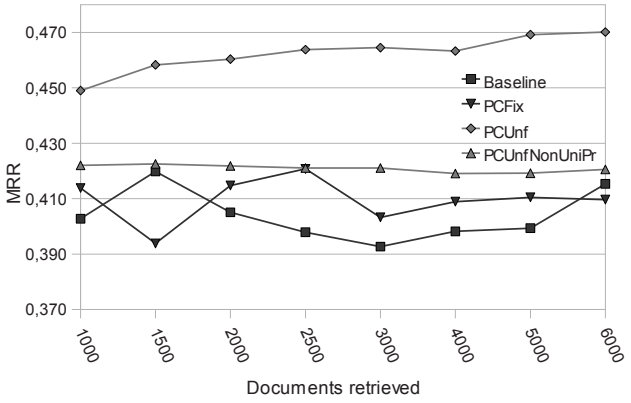

(b)

**Figure 3.4**: Precision at 5 over different numbers of documents retrieved, for the queries from 2006 (a), and for the queries from 2005 (b)

built using insufficient amount of training data, i.e. related documents. Both effects in total prevent incidental persons from getting high scores. However, using non-uniform priors spoils the performance of **PCUnf** in case of MRR measure. So, if the user information need can be effectively satisfied with only one expert (and he/she is always available for requests), then **PCUnf** is more preferable (it is not worse than our baseline for 2006 queries, but notably better on 2005 queries). In order to find the additional evidence for the observations we made, we measured the statistical significance of the im-




(a)




(b)

**Figure 3.5**: MRR over different numbers of documents retrieved, for the queries from 2006 (a) and for the queries from 2005 (b)

provements we had over both sets of queries. We considered improvements with $p < 0.05$ for the paired t-test to be statistically significant. We found that our improvements according to MAP measure are statistically significant for **PCUnfNonUniPriors** method for all numbers of documents retrieved. MAP improvements of **PCUnf** method are not significant if we retrieve less than 2000 documents. Improvements according to P@5 for both methods become significant starting from 2000 documents as well. Improvements for MRR are significant only with 2005 queries for **PCUnf** method.

To sum things up, the experiments indicate that our person-centric model is built on supposedly more realistic and more beneficial assumptions than the baseline document-based model.

## 3.2 Using sequential dependencies

As it is already observed in this chapter, despite that the assumption of independence is very popular in various IR tasks, it does not always lead to the best performance in each and every case. Including such features of the task that better characterize co-occurrence of terms often helps to improve. In this section we propose to take not only the fact of co-occurrence into account, but also the important property of this co-occurrence: the sequential order of a candidate's identifier and the query terms mentioned in a document. While we do not try to assign any specific semantics to types of the order, we hope that treating them differently may be beneficial even for simple expert finding approaches.

Two ways of considering positions of terms in documents are especially popular in document retrieval, namely, in query expansion. While one approach measures the degree of proximity of a term to the query terms in the scope of a document (Gao et al., 2002), the other also takes the sequential order of terms into account (Metzler and Croft, 2007). Once, it was shown for expert finding that the overall pairwise distance between a candidate's mention and the query terms in a document expresses the degree of association between the document and the candidate (Petkova and Croft, 2007). At the same time, the importance of the order in which personal identifiers and query terms occur in documents was never studied to the best of our knowledge.

The section is structured in the following way. The next section explains the details of our method taking orders into account. Then it is followed by a section describing empirical evaluations of the proposed method and resulting conclusions.

### 3.2.1 Weighting orders differently

We propose to use the sequential orders of query terms and candidate experts in documents for estimating the amount of document relevance probability that should be propagated to the candidate. While in the previous section we applied sophisticated analysis to determine the strength of association between a document and the persons mentioned, in this section we propose a lightweight method mainly helpful in cases when there is no specific information about importance of document parts where persons are mentioned. Note that in contrast to our previous method which is based on dependence of terms on persons, here we relax this assumption and assume that terms are generated independently by a document. At the same time, we suppose that candidate experts relate to the document's content up to the degree explained by their position in respect to the document's most relevant part.

Since according to the above-mentioned assumption the probability of generating a person from a document is query-dependent, we change the definition for the classic document-based expert finding method, described by Equation 2.1 (see Section 2.1.2) in the following way:

$$P(Q,e) = \sum_{D \in Top} P(Q|D)P(e|Q,D)P(D) \tag{3.9}$$

$$P(e|Q,D) = \frac{a(e,Q,D)}{\sum_{e'} a(e',Q,D)}, \tag{3.10}$$

where $P(e|Q,D)$ is the probability of association between the candidate and the document and $a(e,Q,D)$ is the non-normalized association score between the candidate and the document given the query. Both *the probability and the association score are query-dependent*. They depend on where the candidate's personal identifier is mentioned in the text with respect to the positions of the query terms. We recognize the following types of sequences to weight them differently:

- $a(e,Q,D) = w^{before}$: The candidate $e$ is mentioned before any query term is mentioned $(e, q_1, ..., q_k)$,
- $a(e,Q,D) = w^{after}$: The candidate $e$ is mentioned after all query terms are mentioned $(q_1, ..., q_k, e)$,
- $a(e,Q,D) = w^{between}$ The candidate $e$ is mentioned in between of the query terms $(q_1, ..., e, ..., q_k)$.

Although we do not try to define any semantics for these orders, they, in fact, may have various meanings, depending on specifics of a collection.

For instance, authors of documents are usually mentioned before any topical words, people which are used to describe the topic probably occur somewhere in between of query terms and those who made lesser contributions are mentioned in the acknowledgments after all topical words. The proposed approach allows to distinguish these roles even when documents are not structured and persons mentioned in them are not semantically annotated. The experiments described in the next section simulate this widespread case.

### 3.2.2 Experiments

The CSIRO collection's data is largely unstructured (see Section 2.4.2), in contrast to the data from W3C crawl, which mostly consists of e-mails of a predefined format. Since we expect the benefit of our method to appear for collections like CSIRO, we conduct our experiments with this corpus. 50 queries with judgments made by CSIRO employees and 3500 candidates found in the collection were used for the evaluation. At the collection preparation stage, we extracted associations between candidate experts and documents by searching for the candidates email addresses and full names in the text of documents. It was enough to retrieve 50 documents containing at least one candidate's mention for the best performance of the baseline method.

Two expert finding methods are evaluated: the baseline method based on the assumption of independence between query terms and the candidate's mention (see previous section, Equation 3.1) and our method using the assumption of their sequential dependence described in the previous section (see Equations 3.9, 3.10). For the baseline method the association score between the document and any candidate mentioned is always equal to 1.0. Since our method has only 3 parameters, we calculated their optimal setting with a simple hill climbing search method using cross-validation with a 80/20 split. The best performance of our method was always reached with roughly the following values for association scores: $w^{before}(e, Q, D) = 10.0$, $w^{after}(e, Q, D) = 0.1$, $w^{between}(e, Q, D) = 1.0$. In the case of several mentions of the same candidate in a document (what rarely happened), the maximum weight was used. This result suggests that the most important persons in the most relevant documents are often mentioned before topical words (for instance, it often happens in resumes that are supposedly very important sources of expertise evidence in CSIRO). Of course, this may vary between organizations and should actually depend on what kinds of documents prevail in the organization. However, even when the structure of documents is hard to define, but it is still possible to group them by type, these weights should better be tuned for each such type individually.

| | MAP | MRR | P@5 |
|---|---|---|---|
| Independence | 0.361 | 0.508 | 0.220 |
| Seq. Dependence | 0.384 | 0.543 | 0.232 |

**Table 3.1**: The performance of expert finding methods using two assumptions: independence and sequential dependence between candidates and query terms

Both methods are compared in terms of common Information Retrieval performance measures officially used in TREC evaluations: Mean Average Precision (MAP), precision at top 5 ranked candidate experts (P@5) and Mean Reciprocal Rank (MRR). Table 3.1 demonstrates the advantage of our method based on the assumption of sequential dependence over the baseline method assuming sequential independence. Improvements for MAP and MRR are statistically significant for the paired t-test with $p < 0.05$ and are not significant at this level for P@5 measure.

## 3.3 Summary

We presented two novel methods for expert finding in organizations. The first one is based on modeling retrieved documents as mixtures of personal language models. In other words, it assumes that terms in documents are generated by those persons who are mentioned in them. It finally ranks candidate experts by combining the following evidence of their expertise: the probability of generation of the query by the personal language model and the prior probability of being an expert expressed in terms of a candidate's activity in the discussions on the topic. We proposed two ways of personal models extraction from top ranked documents. In one case, we considered that person-document relation probabilities are fixed and fully depend on the field of a document where the person appeared. In another case, we obtained these probabilities dynamically by predicting the contribution of persons to a document based on their intermediately estimated language models. When our method used this second way of modeling, it outperformed one of the best state-of-the-art approaches which we used as a baseline.

Our second method takes the sequential dependence of a person and query terms occurring in a document into account. We claim that it is useful to differentiate the orders of their occurrence in a document to estimate the strength of the relation of a candidate expert to the document's content. We supposed that these orders bear semantics that is quantifiable, though not necessarily precisely defined. We formalized our assumptions and successfully justified them with experiments.

# 4

# Beyond the scope of directly related documents

In most systems the prediction of personal expertise is made through the analysis of textual content of documents the person is directly related to (Maybury, 2006; Craswell et al., 2005a; Bailey et al., 2007b). The majority of the proposed approaches shares the principle claiming that the relevance of the local textual context of a person adds up to the evidence of his/her expertness. Furthermore, methods estimating relevance of the textual content related to a person on the lower and hence less ambiguous level (e.g. paragraph or sentence level) usually more effective (Petkova and Croft, 2007; Balog and de Rijke, 2008) (see Section 2.1.2). Most of these approaches 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 the immediate neighborhood of a candidate. Being in fact based on the principle of relevance flow in the direction from documents to persons, most methods assume that the propagation of relevance should stop after the very first step.

Alternative approaches find experts by measuring network centrality of persons in specialized professional communities or just in a sub-set of documents with minimum acceptable relevance to a query (see Section 2.1.4). However, these methods use documents just to set up relations among candidate experts and ignore the fact that such relations may exist only in the context of certain topic. So, while being effective for query-independent tasks like finding the most authoritative experts in question/answering portals and forums (Agichtein et al., 2008), they are inferior in performance to the state-of-the-art query-dependent expert finding methods (Chen et al., 2006).

In this chapter, we answer research questions posed in the beginning of this thesis and related to **Research Objective 2** (see Section 1.2). We ad-

vance previous work by combining features of both above-mentioned classes of expert finding methods. We claim that even when local context deserves to be the primary source of evidence, there is no obvious reason that global context located not in the scope of the documents mentioning the candidate should be ignored. 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 (Crestani, 1997), we allow the probability of document relevance to flow further through reciprocal connections between persons and documents. Initial ideas and first versions of the algorithms described in this chapter were demonstrated in (Serdyukov et al., 2007c; Rode, 2008) and further developed in (Serdyukov et al., 2007c, 2008b). To sum up, our contributions are as follows.

- 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 chapter is organized as follows. In the next section 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. The related research on link-based analysis is described in detail in Section 4.2. Our experiments with two real-world test collections supporting our assumptions are discussed in Section 4.3. Finally, Section 4.4 summarizes our insights and outlines ideas for follow-up research.

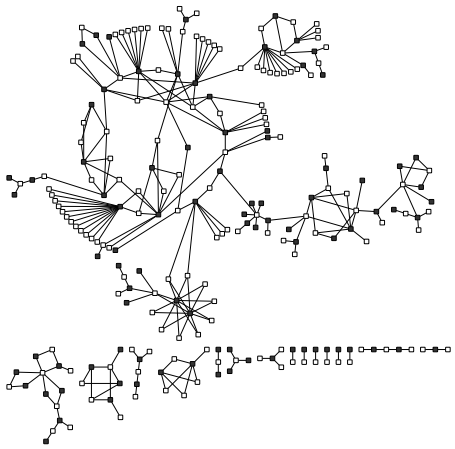


**Figure 4.1**: A fragment of the real expertise graph with links between documents (white nodes) and candidate experts (black nodes) for the query "sustainable ecosystems"

## 4.1 Expertise estimation by relevance propagation

### 4.1.1 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. We will further on call them *expertise graphs*. 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.

The simplest form of expertise graphs is always bipartite, since all edges point from documents to experts only and back. Figure 4.1 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. Inter-document links are represented by directed edges following the link if not specified otherwise. Person-to-person edges are usually reciprocal in organizations since professional contacts are often tight and face-to-face within one enterprise, especially for employees working in the same or related sub-units. It is not generally the rule for more informal social networks. In Twitter (`www.twitter.com`), for example, *followers* are not necessarily *followed* by those whom they follow.

By including additional edges, the graph gains a higher density and enables more intensive relevance propagation, however by losing its strict bipartite property. Moreover, it allows to discover those experts that are not well represented in organizational documentation by some reason (e.g. do not share it with expert finding system due to privacy concerns). These experts, if found, should relieve the load of requests for those people often mentioned in documents.

*Including Further Nodes.* 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. Moreover, organizational units may serve either as mediators in search for employees, or as actual objects of search. Persons outside the company and other companies might reveal interesting connections as well, if they are mentioned in the documents together with candidate experts. We may also incorporate relations and entities extracted from other external global professional networks (e.g. `LinkedIn.com`).

*Controlling Topical Specificity.* The size and density of an expertise graph depends on the number of retrieved documents. In a simplest case such a graph may be query-independent and include the entire collection. However, in this thesis we assume that it is important for such a graph to have topical focus. Since, it is dangerous to reward documents and candidates not initially supported with sufficient evidence of relevance/expertise from their textual content/context, we intentionally ignore highly irrelevant documents when building expertise graphs.

### 4.1.2 Baseline: one-step relevance propagation

As it was mentioned in the beginning of this chapter, both aggregated relevance and centrality based expert finding methods start from making too rough simplifications. Most importantly, document-based based methods do not include relations between experts into their models and do not notice

documents that relate to persons indirectly. One of the most theoretically sound representatives of these methods, proposed by Balog et al. (2006), follows the probabilistic language modeling principle of IR (Hiemstra, 2001) and defines the probability of expertness for the candidate expert $e$ with respect to the query $Q$ as described by Equations 2.1 and 2.4 (see Section 2.1.2).

If we carefully examine these equations, 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 2.1 as our baseline. Here and further on we also use the probability of selecting a document given a candidate:

$$P(D|e) = \frac{a(e, D)}{\sum_{D'} a(e, D')} \tag{4.1}$$

where $a(e, D)$ is the non-normalized association score between the candidate $e$ and the document $D$ proportional to their strength of relation. The probability of selection a candidate given a document is defined by Equation 2.4. Our way of distributing these scores over candidate experts in a document is described in Section 4.3.1).

### 4.1.3 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 described by Figure 4.1.3 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 (Ackerman et al., 2002).

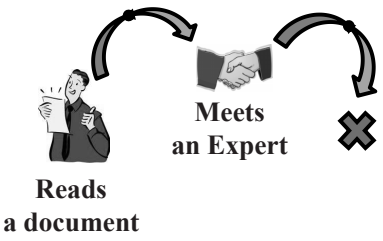



**Figure 4.2**: Expert finding as a one-step process

We may imagine that the search for expertise may consist of the following repeating stages of gradual knowledge acquisition (see Figure 4.1.3):

(1) At any time: (a) randomly reading a document, or just picking a random candidate,

(2) After reading a document: (a) consulting a person mentioned in this document, or (b) checking for other linked documents and reading one of them, or

(3) After consulting a person: (a) reading other documents mentioning this person, or (b) consulting another candidate expert which is recommended by this person.

Note that while modeling the 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 documents and people. First we present three models considering that expertise graphs are bipartite (graphs used for 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.

### 4.1.4 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

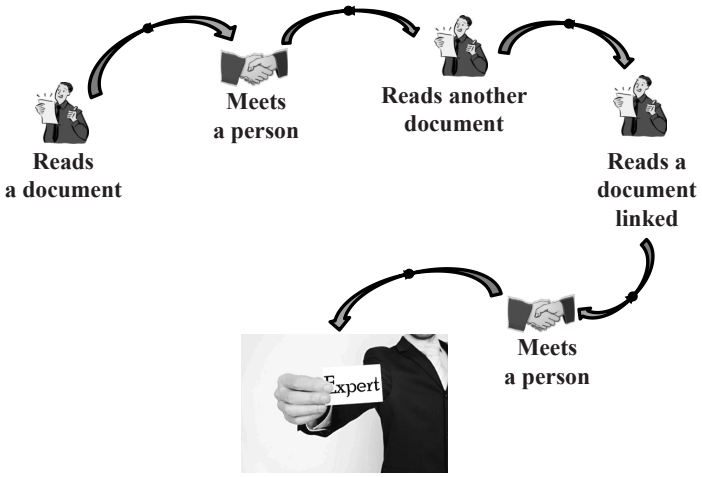



**Figure 4.3**: Expert finding as a multi-step process

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. Thus, after getting the list of ranked documents with the list of related candidate experts attached, the user:

- selects a document,
- makes K-steps of two kinds: (a) if a user is in the document node, then either one of related candidate experts is selected, or the reading of the document is continued, or (b) if a user is in the candidate node, then one of documents related to this candidate is selected.

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 a starting document $D$ as a starting point for the walk is proportional to its probability of relevance $P(Q|D)$, and

(2) the probability to stay at a document node at any step is also proportional to its probability of relevance $P(Q|D)$, while the probability to leave the node is proportional to the probability of its irrelevance $(1 - P(Q|D))$.

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:

$$P_0(D) = P(Q|D), P_0(e) = 0, \tag{4.2}$$

$$P_i(D) = P(Q|D)P_{i-1}(D) + \sum_{e \to D} P(D|e)P_{i-1}(e), \tag{4.3}$$

$$P_i(e) = \sum_{D \to e} (1 - P(Q|D))P(e|D)P_{i-1}(D) \tag{4.4}$$

The probabilities $P(e|D)$ and $P(D|e)$ are defined in Equations 2.4 and 4.1. Finally, we consider that the expertise of $e$ is proportional to $P_K(e)$.

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 $P_1(e) \dots P_K(e)$. 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 our method can be easily utilized in this way, we experiment only with unsmoothed probabilities.

### 4.1.5 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 of 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:

$$P_i(D) = \lambda P_J(D) + (1-\lambda) \sum_{e \to D} P(D|e) P_{i-1}(e), \tag{4.5}$$

$$P_i(e) = \lambda P_J(e) + (1-\lambda) \sum_{D \to e} P(e|D) P_{i-1}(D) \tag{4.6}$$

$$P_J(D) = P(Q|D), P_J(e) = \frac{n(e, Top)}{|Top|}, \tag{4.7}$$

where $\lambda$ is the probability that at any step the user decides to make a jump and not to follow outgoing links anymore, $n(e, Top)$ 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 the expertise of $e$ is proportional to the stationary probability $P_\infty(e)$. Although our method is computationally more intensive than the baseline one-step propagation, it converges quickly (after 200-300 iterations) for typical expertise graphs containing at most 2000 nodes according to our experiments.

### 4.1.6 Absorbing random walk

In our next approach we represent the search for an expert as *an absorbing random walk* in a document-candidate graph (Serdyukov et al., 2008b). We calculate the probability of finding a candidate if we consider that this candidate is the required expert. The candidate node which we want to evaluate

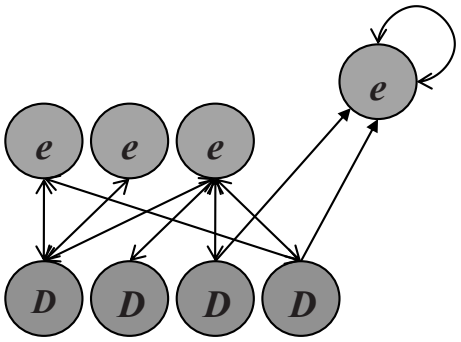



**Figure 4.4**: Absorbing random walk

is only self-transient, since we assume it to be the final destination of the walk. Formally speaking and as illustrated by Figure 4.1.6, we remove all outgoing edges from the measured candidate, add the self-transition edge to it and use the following equations iteratively:

$$P_0(D) = P(Q|D), P_0(e) = 0, \tag{4.8}$$

$$P_i(D) = \sum_{e \to D} P(D|e) P_{i-1}(e), \tag{4.9}$$

$$P_i(e) = \sum_{D \to e} P(e|D) P_{i-1}(D) + P_{i-1}(e) P^{self}(e|e) \tag{4.10}$$

Finally, we consider that the expertise of $e$ is proportional to the probability $P_\infty(e)$. Note that $P^{self}(e|e)$ equals 1.0, since we removed all edges from any node $e$ under study. Making the full run of iterations for each candidate is unnecessary. If we rewrite the above equations in a matrix form, we get:

$$\mathbf{p} = \mathbf{p}_0 \mathbf{A}^i, \tag{4.11}$$

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:

$$\mathbf{p} = \mathbf{p}_0\mathbf{B} \tag{4.12}$$

The absorbing random walk based method has several advantages over the previously presented methods. Considering that we defined the size of our expertise graphs, this method does not need to tune any other parameters. It is also a direct generalization of the one-step propagation method. This means that in contrast to the one-step approach using one-step probabilities, our multiple-step method in fact calculates the probability $P^{mult}(e|D)$ of finding candidate expert $e$ by making *minimum sufficient number of steps* starting from document $D$:

$$P(Q,e) = \sum_{D \in Top} P^{mult}(e|D)P(Q|D)P(D) \tag{4.13}$$

In other words, the equation 4.13 provides the opportunity to propagate relevance to a candidate not only from directly related documents, but also from any documents from which there is a path to the candidate. It should be mentioned that in strongly connected graphs with an absorbing node the probability of absorption in the infinity is effectively close to 1.0. 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. Alternatively, in the cases when graphs are dense, we could calculate the probability to reach the node in some $K$ number of steps, what would mean doing $K$ iterations using the above equations. It was actually clear from the preliminary experiments that the relevance of documents situated farther than 3 nodes from a measured candidate has almost no influence on its probability of absorption, since the probability of following such paths is too low.

### 4.1.7 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. This often helps to model their mutual relations and directions of influence better. The scenario of search 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 (Ackerman et al., 2002).

We experimented with adding these new transitions to our expertise graph and using them for Infinite Random Walk method. The iterations specified in Equations 4.5 and 4.6 are updated in the following way:

$$P_i(D) = \lambda P_J(D) + (1-\lambda)((1-\mu_D)\sum_{e \to D} P(D|e)P_{i-1}(e) + \\ + \mu_D \sum_{D' \to D} P(D|D')P_{i-1}(D')), \tag{4.14}$$

$$P_i(e) = \lambda P_J(e) + (1-\lambda)((1-\mu_e)\sum_{D \to e} P(e|D)P_{i-1}(D) + \\ + \mu_e \sum_{e' \to e} P(e|e')P_{i-1}(e')), \tag{4.15}$$

where $\mu_D$ is the probability of following document-document connections, $\mu_e$ is the probability of following candidate-candidate connections. The new transition probabilities are calculated as:

$$P(D|D') = 1/N_{D'}, P(e|e') = 1/N_{e'}, \tag{4.16}$$

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. Alternative approach is suggested by Balog et al. (2007) when the scores of candidates are linearly combined with scores of the best candidates from the same organizational units.

## 4.2 Related work on 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 (Page et al., 1998), HITS (Kleinberg, 1999) (its random walk based version is described by Ng et al. (2001)) and SALSA (Lempel and Moran, 2001) are probably

the most popular. Several attempts have been made in the last years to make these models query and content dependent. Note that HITS is content dependent only to the extent that it starts from retrieval of top relevant documents. However, it later ignores their relevance scores, considering only links among them. The Intelligent Surfer (Richardson and Domingos, 2001) walks to linked pages biased by their relevance to the query. Personalized Pagerank allows to put preferences on certain web-pages, so that the centrality of any document would depend on its proximity to preferred ones (Page et al., 1998; Jeh and Widom, 2003). Both ideas were further combined in a unified framework which considered also the bi-directional walk over hyperlinks (Shakery and Zhai, 2006). Random walks on graphs containing queries and clicked links (or entire search trails) were recently utilized for web search result expansion (Craswell and Szummer, 2007; Bilenko and White, 2008). Searching with graph-based methods for typed entity classes on the Web was explored recently in several publications (Cheng et al., 2007; Zaragoza et al., 2007; Tsikrika et al., 2007).

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 (Lafferty and Zhai, 2001; Kurland and Lee, 2006). Finite random walk over terms through thesaural and syntactic relations is applied to query expansion (Toutanova et al., 2004; Collins-Thompson and Callan, 2005) and question answering (Harabagiu et al., 2006) tasks. Erkan and Radev use implicit links between similar sentences to compute their centrality for text summarization (Erkan and Radev, 2004). It was also used as a model of preference flow between users with shared interest in a recommendation system (Song et al., 2006).

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

## 4.3 Experiments

### 4.3.1 Experimental setup

We conduct our experiments with two data sets provided by the TREC community (see also Sections 2.4.1 and 2.4.2. Although both testbeds originate from the real-world organizations and allow to realistically simulate classic expert finding scenarios, they have some clear differences.

**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 (see details in Section 2.4.1). 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. The substantiation of our choice can be found in Section 3.1.3. The W3C data contains a 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. Instead of a list of candidate experts, only the structure of candidates' email addresses was provided: *firstname.lastname@csiro.au*. For the purpose of finding candidate experts, we extracted all email addresses from the collection with *csiro.au* domain and *firstname.lastname*-like first part using a regular expression. Additionally, we bypassed spam-protection by recognizing several anti-spam aliases, like *[at]* for @, and codes in Javascript for dynamic generation of e-mail addresses. We also made an automatic match of emails with the same first part, but with different subdomains to one candidate identifier. For example: Alan.Smith@cmis.csiro.au, Alan.Smith@ento.csiro.au → Alan.Smith@csiro.au. If the email address without subdomains did not exist in the collection for the specific person, it was made up. We also had a list of email addresses to be banned which were not personal, but organizational addresses (e.g. publishing.photos@csiro.au). Using this strategy we built our own candidates list by finding about 3500 candidates in the collection. 50 queries with judgments made by 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 (as in Section 3.1.3): $a(e, D^{from}) = 1.5$, $a(e, D^{to}) = 1.0$, $a(e, D^{cc}) = 2.5$ and $a(e, D^{body}) = 1.0$ respectively. If a person appeared in several fields, only the maximum of association scores is considered.

The results analysis is based on calculating popular IR performance measures also used in official TREC evaluations (see Section 2.4.4): Mean Aver-

age 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 multi-step relevance propagation approach. However, for the sake of a fair comparison, we also show the performance of the simplest of known methods, called Votes by Macdonald and Ounis (2006), 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 (Macdonald and Ounis, 2006),
- **Baseline**: the baseline one-step relevance propagation method (see Section 4.1.2),
- **FRW**: the multi-step relevance propagation with Finite Random Walk method (see Section 4.1.4),
- **IRW**: the multi-step relevance propagation with Infinite Random Walk method (see Section 4.1.5).
- **ARW**: the multi-step relevance propagation with the Absorbing Random Walk method (see Section 4.1.6).

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 Section 2.1.2) 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. During our preliminary experiments with one-step propagation, we observed that the optimal number of retrieved documents varied considerably over the data sets. We had to retrieve 1500 documents from the W3C collection and just 50 documents from the CSIRO collection for the maximum performance of the one-step propagation method (see Figures

4.5 and 4.6). We considered this performance as our baseline. 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.

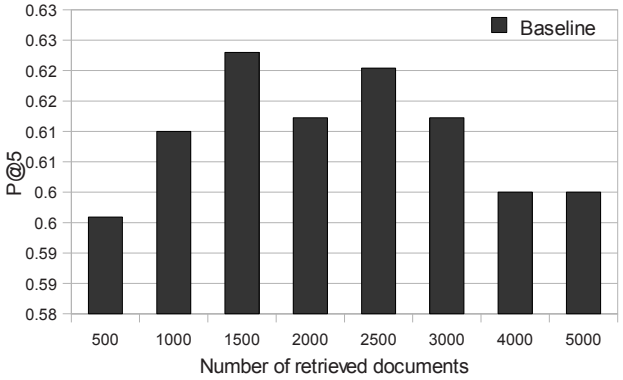



**Figure 4.5**: P@5 over different numbers of retrieved documents for the baseline method, W3C (2006) data

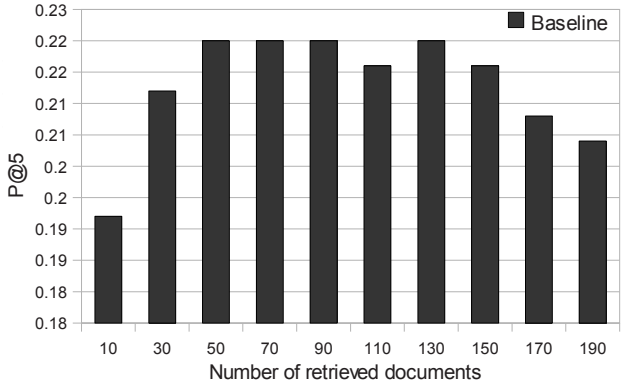



**Figure 4.6**: P@5 over different numbers of retrieved documents for the baseline method, CSIRO (2007) data

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.

### 4.3.2 Experiments with multi-step relevance propagation

Both FRW and IRW methods depend on one parameter. In case of FRW this is the number of relevance propagation steps $K$ to be done. Figures 4.7 and 4.8 compare performance of Baseline with the performance of FRW after from 1 to 43 propagation steps for both data sets.

We see that the maximum MAP is reached after making on average 13 steps for the W3C data and 33 steps for the CSIRO data. This is not a very long walk in our graph, since we have a probability to stay at document nodes and also because a typical expertise graph (see Figure 4.1) 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 increasing the number of propagation steps from one to more also leads to improvement of MAP for both datasets, making this noticeable already after a few steps.

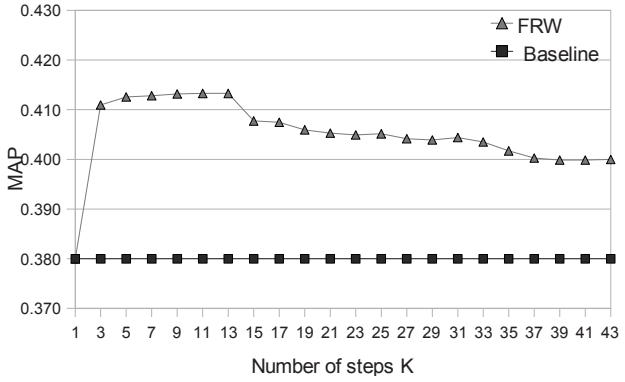



**Figure 4.7**: MAP for the Finite Random Walk method (FRW) with different numbers of propagations steps taken, W3C (2006) data

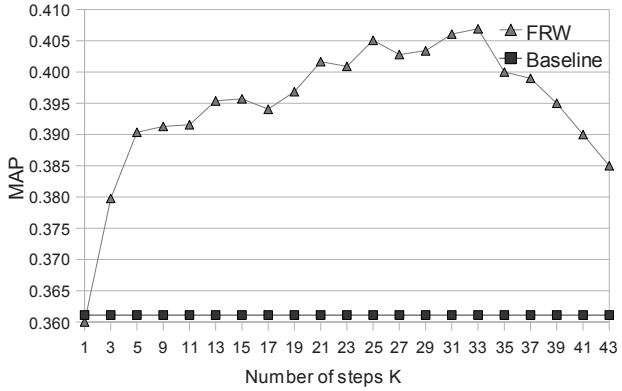



**Figure 4.8**: MAP for the Finite Random Walk method (FRW) with different numbers of propagations steps taken, CSIRO (2007) data

The only parameter to tune for IRW method is $\lambda$ - the probability of a walk restart. Empirically, in Figure 4.9 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 $\lambda$ to the value between 0 and 0.15 is also typical for the use of random walks in web retrieval (Najork et al., 2007).

To avoid the effects of over-training for IRW and 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 cannot be used to tune parameters for each other. Since ARW method is parameter-free given an expertise graph, 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 4.1. As hoped, we see that actually both Infinite and Finite Random Walk methods are equally effective and outperform Baseline method for all measures. ARW method also shows the improvement over the baseline for all measures, but still seems inferior to IRW and 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 MAP is significant for all three methods at the $p < 0.001$ level. For MRR it is significant at the $p < 0.01$ level for IRW method and at the $p < 0.05$ level for FRW method. For P@5 it is significant at the $p < 0.01$ level for IRW and ARW methods, and at the $p < 0.05$ level for FRW method. The improvement we got is also comparable with the advantage of Baseline over Votes method, which is one of the simplest methods known. Having in mind that Baseline is the one of the most effective methods known, we may conclude that the improvements of our methods demonstrate their importance for the research in expert finding.

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 4.8, 4.7 and 4.9) 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.

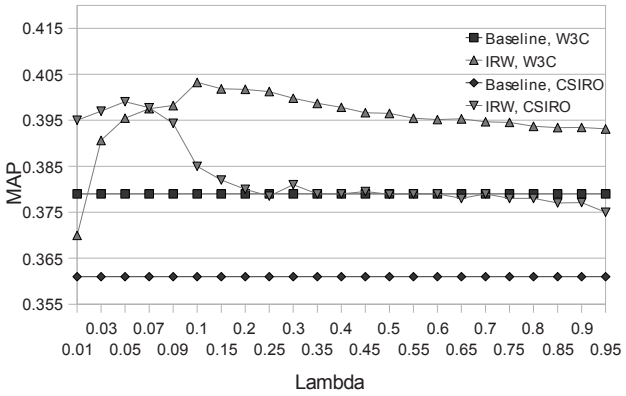



**Figure 4.9**: MAP for Infinite Random Walk method (IRW) over different values for jumping probability

| | 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 4.1**: Performance for all measures, both data sets and all tested methods

### 4.3.3 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. 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 4.10 (values for $\mu_e$ and $\mu_D$ probabilities are on X-axis) demonstrate only the benefit from adding organizational links. When we set the probability of relevance propagation between related candidate experts $\mu_e$ 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 $\mu_D$.

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

[1] `wikipedia.org/wiki/Organizational_structure`, visited in December 2008

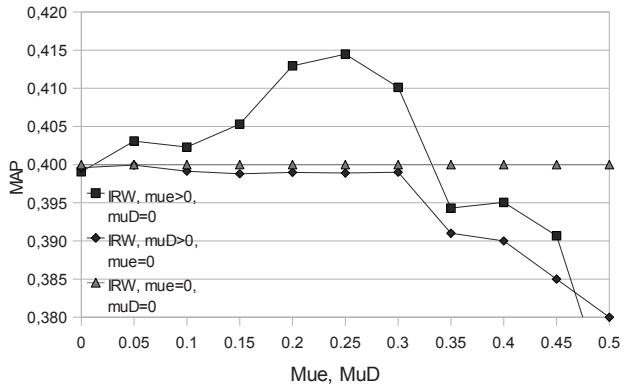



**Figure 4.10**: MAP for Infinite Random Walk method (IRW) with additional links, CSIRO (2007) data

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 interconnect 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 4.11 that using simulated organizational links increases the performance of Infinite Random Walk method for all values of $\mu_e$ 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.

## 4.4 Summary

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, used as a baseline in this chapter, 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

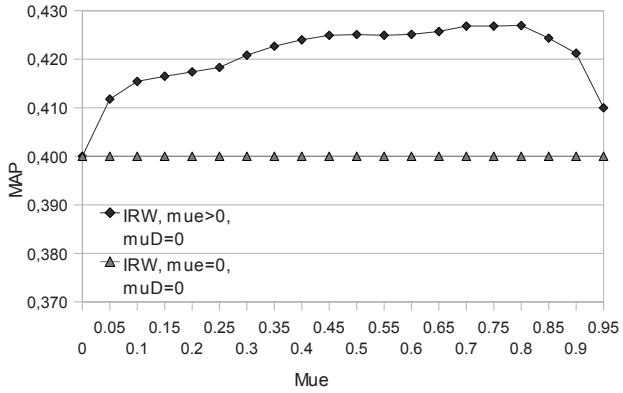



**Figure 4.11**: MAP for Infinite Random Walk method (IRW) with simulated organizational links, W3C (2006) data

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.

# 5

# Beyond the enterprise

In the previous chapter, we demonstrated that expertise evidence found not in the immediate vicinity of candidate experts is still capable to significantly influence their ranking. Despite that we expanded the person-specific document space to those documents that are indirectly related to candidates, we still assumed that we restrict our analysis to the documents stored in a single enterprise.

In this chapter, we answer research questions posed in the beginning of this thesis and related to **Research Objective 3** (see Section 1.2). We propose to avoid the above-mentioned limitation and explore the predicting potential of expertise evidence acquired from sources publicly available on the Web and not only originating from the organization under study. Using APIs of two major web search engines, we show how different types of expertise evidence, found in the organization and outside, can be extracted and combined together. Finally, we demonstrate how taking the web factor seriously significantly improves the performance of expert finding in the enterprise.

The remainder of this chapter is organized in two sections. Section 5.1 explains our strategy of expertise evidence acquisition from various web sources using generic and vertical search engines. It shows how to combine evidence coming from different verticals by means of rank aggregation. Section 5.2 proposes to treat each search result item (URL, title and snippet) differently by using query-dependent and query-independent measures of its quality.

## 5.1 Acquiring Expertise Evidence from the Web

Analysis of personal expertise should not be necessarily undertaken using only organizational data. Such a limitation may lead to erroneous recom-

mendations due to training data incompleteness and often contradicts with user expectations. Typical users rely on expert finding systems not only in their everyday need in helpful people. They look forward to expand their social networks and set up long-run relations with professionals respected also outside of the organization. In these and similar cases, users would like the system to suggest not only knowledgeable and intelligent people, but also authorities in the area of their specialization. This means popular persons who are socially active and often quoted or cited on a topic of interest not only by their co-workers.

Consequently, a user-friendly expert finder needs to leave an organizational "cage" in search for additional expertise evidence in public sources. The obvious solution for finding expertise evidence outside of the enterprise is to search for it on the Web, what means getting access to various kinds of data: web pages, blogs, newsgroups, electronic libraries etc. The expertise evidence found in these sources may have a different meaning, but anyway adds up to the overall trust in a person as an expert. There are basically two ways of acquiring it:

- **Crawling and RSS Monitoring**. Many web data mining systems rely on focused crawling and analyzing discovered RSS feeds (Gruhl et al., 2006; Ziegler and Skubacz, 2006). However, it is often not even necessary to develop your own web spider - topical monitoring can be implemented by means of such powerful aggregating tools as Yahoo! Pipes (`pipes.yahoo.com`) or Google Alerts (`google.com/alerts`).
- **Search Engine APIs.** It is still possible to avoid "downloading the Internet" and use open APIs of the famous web search engines - Google (`code.google.com/apis`), Yahoo (`developer.yahoo.com`) or Live Search (`dev.live.com`) (McCown and Nelson, 2007). Google has no limits on number of queries/day, Yahoo limits it to 5000 (at the time the present research was conducted, but now unlimited), Live Search to 25000. All engines provide the access not only to their generic web search services, but also to vertical search in *maps*, *images*, *news* etc. Unfortunately, it is not possible to automate data collection from services not accessible via APIs, even when it is easy to create wrappers for their web interfaces. Search engines usually have a right to ban IPs sending automated queries according to their Terms of Service.

### 5.1.1 Fast evidence acquisition with search engines APIs

One could imagine an expert finder that is equipped with a web crawler focusing on retrieval of employee-specific information from the Web. Such

a spider would provide us with plenty of information about how the organization is positioned in the world or regional markets, how influential and wide-spread its organizational knowledge is. However, in case when an expert finder should be made cheap but efficient, the enterprise may rely on powerful mediators between people and the Web: leading search engines and their public search APIs.

Since it is basically infeasible even for a wealthy organization to maintain an effective web search crawler, we focus on using APIs of two leading web search engines: Yahoo! and Google (Live Search API is still in unstable beta state). We extract expertise evidence for each person from their databases using the following strategy.

First, we build a query containing:

- the quoted full person name: e.g. *"tj higgins"*,
- the name of the organization: e.g. *csiro*,
- query terms without any quotes: e.g. *genetic modification*,
- the directive prohibiting the search at the organizational web site (in case of Web or News search):e.g. *-inurl:csiro.au*.

Adding the organization's name is important for the resolution of an employee's name: the ambiguity of personal names in web queries is a sore subject. It was shown that adding the personal context to the query containing a name or finding such context automatically significantly improves the retrieval performance (Shen et al., 2008). Of course, one could possibly improve by listing names of all organizations where the person was ever employed (using OR clause) or by adding such context as the person's profession or title. However, the latter may still decrease the recall, cause this information is rarely mentioned in informal texts. It is also possible to apply more sophisticated strategies for names representation (e.g. using the first name's diminutive forms and abbreviations), but we avoided using them for the sake of fast implementation and also as a quick solution for ambiguity resolution. In some cases, namely when using Web and News search services, we also added a clause restricting the search to URLs that do not contain the domain of the organization. It was done to separate organizational data from the rest of available information. In some cases, when an organization's domain is not unique, it is useful to just enlist all organizational domains, each in separate *-inurl* clause.

As the second step of acquiring the evidence of a certain type, we send the query to one of the web search services, described further in this section. The

returned *number of results* is considered as a measure of personal expertness. In other words, we ask a specific search engine: "Please, tell us how many times *this person* occurs in documents containing *these query terms* and not hosted at *the domain of her/his own organization*". The answer shows the degree of relation of a person to the documents on the topic what is a common indicator of personal expertness (see Section 2.1). Our technique is akin to the Votes method measuring a candidate's expertness by the number of organizational documents retrieved in response to a query and related to the candidate (Macdonald and Ounis, 2006) (see Section 2.1.2).

Due to limits of the Search Engine API technology we used, we need to restrict the number of persons for which we extracted global expertise evidence. In case of large organizations with thousands of candidate experts, it is unrealistic and unnecessary to issue thousands of search engine queries containing each person and the initial user query. So, pre-selection of candidates using enterprise data only makes a lot of sense. In our experiments, described later in Section 5.1.8, we proceed further with only 100 most promising candidate experts per query according to the evidence found in the enterprise. Processing one query takes less than a second. So, it usually took from 15 to 70 seconds to issue queries for all candidates, to wait for all responses of one search engine and to download all search result pages. While, it seems intolerable for document search, expert finding is the task in which people re-formulate their queries less often and appreciate the quality of recommendation much higher. Moreover, expected time spent in pointless conversations with unqualified people is usually longer than one minute, considering that meetings are not always short-time and spontaneous. Due to a significant increase in performance that we achieve with our techniques (see Section 5.1.8), we hope that other issues are of minor importance.

Apart from the ranking built on fully indexed organizational data, we built rankings using 6 different sources of expertise evidence from the Web: Global Web Search, Regional Web Search, Document-specific Web search, News Search (all via Yahoo! Web search API), Blogs Search and Books Search (via Google Blog and Book Search APIs). We describe each type of evidence and details of its acquisition further in this section.

### 5.1.2 Acquiring evidence from Enterprise

Despite the presence of vast amount of personal web data hosted outside of the corporate domain, the enterprise itself stays the main repository of structured and unstructured knowledge about its employees. Moreover, large part of enterprise documentation is often not publicly accessible and hence not indexed by any of web search engines. Even traditionally public Web 2.0

activities are often insistently popularized to be used fully internally within organizations for improving intra-organizational communication (GuideWire-Group, 2005). According to recent surveys (Levine, 2008), 24% of companies have already adopted Web 2.0 applications. Internal corporate blogging (Huh et al., 2007) and Project Wiki technologies (Buffa, 2006) are the most demanded among them. For instance, it is reported that Microsoft employees write more than 2800 blogs and about one third of them is only internally accessible (Efimova and Grudin, 2007).

Since it is usually possible to have fast access to the content of indexed documents in an Enterprise search system, we build Enterprise data based rankings using state-of-the-art expert finding algorithm proposed by Balog et al. (2007). We also use its performance as our baseline. Note that it measures candidate's expertness by calculating a weighted sum of scores of documents retrieved to a query and related to the candidate as defined in Equation 2.1. In contrast to that measure, we aggregate expertise evidence from the Web by simply counting all documents matched to a query and related to the person. Therefore, the only difference is that we consider all document scores equal and do not assume that the amount of a document score propagated to a mentioned candidate depends on the number of candidates in that document. We demonstrate the utility of relevance probabilities calculated for URLs, titles and snippets later in the next section. We also do not expect too many candidates from the same organization to co-occur in documents hosted elsewhere since there are many more chances for professional encounters in a bounded organizational space than in the immense World Wide Web.

### 5.1.3 Acquiring evidence from Web search

The importance of the World Wide Web for finding information about people is unquestionable. Especially, since everyone cares much about "online reputation" and wants to be found, since it is often crucial to be searchable in the Internet Era (Pang and Lee, 2008). The word "Google" is officially added to the Oxford English Dictionary as a verb. "Googling" a person is one of the most popular search activities with dedicated manuals and howtos (Sherman, 2005). 30% of all searches on Google or Yahoo! are for specific people or people related (Arrington, 2007). The increasingly used practice for employment prescreening is to "Google" applicants (Jones et al., 2007). A 2006 survey conducted by `CareerBuilder.com` found that one in four employers use Internet searches to learn more about their potential employees and actually more than half of managers have chosen not to hire an applicant after studying their online activity.

There is however a huge controversy on what search engine is better: Google or Yahoo! Almost everyone has his own opinion on this topic. From one point of view, Google has much larger share of searches in U.S. (59% in February 2008[1]), but Yahoo! is still a bit ahead of Google according to The American Customer Satisfaction Index[2]. After all, we preferred Yahoo! Web Search API by two reasons. Yahoo's search APIs are more developer-friendly and have less usage limitations, namely they are limited to 1000 results per query as opposed to Google API which returns not more than 32 results.

In order to analyze different scopes of a person's mentioning on the web, we built expertise rankings based on several kinds of web searches: without any restrictions (except those mentioned in Section 5.1.1) and with restrictions on domains location and on the type of documents:

- **Global Web Search**. The search without restriction of the scope.
- **Regional Web Search**. The search only at web-sites hosted in Australia (by using Yahoo's *country* search option). The purpose was to study whether we may benefit by expanding the search scope gradually, first searching for the expertise evidence in a company's region.
- **Document-specific Web Search**. The search only in PDF documents (by using Yahoo's *format* search option). The purpose was to study whether it is beneficial to differentiate document types. The PDF format was selected as a de-facto standard for official on-line documents (white papers, articles, technical reports) that we regarded as one of the main sources of expertise evidence.

### 5.1.4 Acquiring evidence from News Search

Good experts should be a bit familiar to everybody. However, to be searchable and broadly represented on the Web does not always mean to be famous and authoritative. What really matters is to be familiar to a large group of people interested in a search topic. It is well-known that news reflect internet buzzes, especially in blogosphere, serving as a filter for events and topics interesting for a broad audience (and vice versa is also true) (Lloyd et al., 2006). It basically means that being on the top of the news often means to be distinguished for your professional achievements: for making a discovery, starting a trend, receiving an award.

[1] www.comscore.com
[2] www.theacsi.org

Yahoo! (`news.yahoo.com`), Google (`news.google.com`) and Live Search offer APIs for their News Search services. However, their significant limitation making them useless for expertise evidence acquisition is that they allow to search only in news from the past month. Since employees are not celebrities and hence are not mentioned in news daily, it is almost impossible to extract sufficient expertise evidence from these services. Google also has News Archive Search (`news.google.com/archivesearch`), but has no API for accessing it.

To realistically simulate the usage of News Search, we took our usual query (see Section 5.1.1), added *inurl:news* clause to it and sent it to Yahoo! Web Search service. In this way we restricted our search to domains and sub-domains hosting only news or to pages most probably containing news.

### 5.1.5 Acquiring evidence from Blog Search

As it was already mentioned in Section 5.1.2, blogs are very rich sources of knowledge about personal expertise. The larger part of corporate professional blogs is public and indexed by major blog search engines. Leading recruiting agencies predict the rapid increase of interest in candidates passionate about writing their blogs (Golta, 2008). Actually, the retrieval task of finding relevant blogs resembles the task of finding experts among bloggers in the Blogosphere. Recently, Balog et al. (2008a) successfully experimented with expert finding methods for *blog distillation* task on TREC 2007 Blog track data.

Two major blog search engines are fiercely competing with each other leaving others far behind: Technorati and Google Blog Search. According to the spreading Internet hype and recent random probings Google has significantly better coverage for blogs (Thelwall and Hasler, 2007). Its Blog Search API is much more developer-friendly than Technorati's, which is often reported to be very unreliable (and it was even impossible to get an Application ID at `technorati.com/developers` at the time of writing). Despite that Google Blog Search API also has its own inconvenient limitations (it can only return up to 8 links in a result set), we use it for building Blog Search based ranking (see Section 5.1.1).

### 5.1.6 Acquiring evidence from Academic Search

Academic publications are a great source of expertise evidence, especially for R&D companies. Not all of them can be found at corporate web-sites, since their free public distribution may be forbidden by copyright terms. There

are two major multidisciplinary Academic Search engines: Google Scholar [3] and Live Search Academic[4]. The others like *Scopus* or *Web of Science* index significantly less publications on many subjects, do not consider unofficial publications and are sometimes restricted to specific types of articles (e.g. to journals). Several studies have shown that it is effective to calculate bibliometric measures for estimating reputation of scientists using citations found in Google Scholar (Bar-Ilan, 2008). It also becomes more popular among researchers to specify in their resumes the number of citations in Google Scholar for their publications. Google Scholar can actually be regarded as a world-wide expert finding system, since it always shows 5 key authors for the topic at the bottom of the result page.

Unfortunately, there is no possibility to access any academic search engine via an API. However, Google provides an API for a very similar search service: Book Search[5]. While its publication coverage is not as large as Google Scholar's, there is a high overlap in the data they both index, since Google Scholar always returns items indexed by Book Search for non-fiction subjects. Using Books Search also naturally allows to search for expertise evidence in not strictly academic sources. So, we build an Academic Search based ranking by sending queries (see Section 5.1.1) to Google Book Search service.

### 5.1.7 Combining Expertise Evidences Through Rank Aggregation

The problem of rank aggregation is well known in research on metasearch (Liu et al., 2007). Since our task may be viewed as *people metasearch*, we adopt solutions from that area. We also decided to use only ranks and ignore the actual number of results acquired for each candidate expert and a query from each search service. It was done for the sake of comparability and to avoid the need for normalization of values.

In our preliminary experiments with different rank aggregation methods we found that the simplest approach is also the best performing. To get the final score we just sum the negatives of ranks which the person $e$ is assigned to if ranked together with the other candidates using expertise evidence found in sources from the set $K$:

[3] `scholar.google.com`
[4] `academic.live.com`
[5] `books.google.com`

$$\sum_{i=1}^{K} -Rank_i(e) \tag{5.1}$$

This approach is often referred as Borda count (Aslam and Montague, 2001). We also tried to learn weights of sources with the Ranking SVM algorithm, using its $SVM^{map}$ version which directly optimizes Mean Average Precision[6] (Yue et al., 2007). However, its performance was surprisingly nearly the same as Borda count's.

### 5.1.8 Experiments

We experiment with the collection used by the Enterprise TREC community in 2007 (see Section 2.4.2). It represents a crawl from Australia's national science agency's (CSIRO) web site and includes about 370 000 web documents (4 GB) of various types. We built our own candidates list by finding about 3500 candidates in the collection (see Section 4.3.1 for details). 50 queries with judgments created by CSIRO Science Communicators (a group of expert finders on demand) were used for the evaluation.

The results analysis is based on calculating popular IR performance measures also used in official TREC evaluations: Mean Average Precision (MAP), Mean Reciprocal Rank (MRR) and precision at top 5 ranked candidate experts (P@5) (see Section 2.4.4).

In our experiments discussed below we compare our methods with a baseline ranking and also study the effectiveness of combinations of rankings. The performance of the following rankings and their combinations is discussed further:

- **Enterprise**: Baseline enterprise search based ranking (see Section 5.1.2),
- **YahooWeb**: Yahoo! Global Web search based ranking (see Section 5.1.3),
- **YahooWebAU**: Yahoo! Regional Web search based ranking (see Section 5.1.3),
- **YahooWebPDF**: Yahoo! Document-specific Web search based ranking (see Section 5.1.3),
- **YahooNews**: Yahoo! News search based ranking (see Section 5.1.4),
- **GoogleBlogs**: Google Blog search based ranking (see Section 5.1.5),

[6] `projects.yisongyue.com/svmmap/`

| | Enterprise | YahooWeb | YahooWebAU | YahooWebPDF | YahooNews | GoogleBlogs |
|---|---|---|---|---|---|---|
| YahooWeb | 0.287 | | | | | |
| YahooWebAU | 0.254 | 0.502 | | | | |
| YahooWebPDF | 0.259 | 0.513 | 0.359 | | | |
| YahooNews | 0.189 | 0.438 | 0.400 | 0.395 | | |
| GoogleBlogs | 0.069 | 0.424 | 0.412 | 0.422 | 0.494 | |
| GoogleBooks | 0.111 | 0.419 | 0.411 | 0.412 | 0.453 | 0.202 |

**Table 5.1**: The normalized Kendall tau distance between all pairs of rankings

| | MAP | MRR | P@5 |
|---|---|---|---|
| Enterprise | 0.361 | 0.508 | 0.220 |
| YahooWeb | **0.423** | 0.547 | **0.248** |
| YahooWebAU | 0.372 | 0.462 | 0.220 |
| YahooWebPDF | 0.358 | 0.503 | 0.200 |
| YahooNews | 0.404 | 0.554 | 0.216 |
| GoogleBlogs | 0.406 | **0.582** | 0.200 |
| GoogleBooks | 0.373 | 0.517 | 0.200 |

**Table 5.2**: The performance of rankings

- **GoogleBooks**: Google Book search based ranking (see Section 5.1.6).

Before starting analyzing the quality of each ranking, we compare them using normalized Kendall tau rank distance measure (Fagin et al., 2003). Low tau scores indicate similarity of rankings and since all candidates are taken from **Enterprise** ranking, each pair of rankings contains the same set of candidates. As we see in Table 5.1, the **Enterprise** ranking appears to be very similar to **GoogleBlogs** and **GoogleBooks** rankings. While the similarity of the latter is also supported by its similar performance with the **Enterprise** (see Table 5.2), the **GoogleBlogs** obviously improves the **Enterprise** not being considerably different. It probably happens because it is different mostly at more important lower ranks. It is also interesting that all four rankings acquired using the same Yahoo Web Search API differ very substantially. This result approves that at least the decision to segregate different information units within one source was reasonable. On the contrary, rankings acquired from Google and even from its different search services disagree at a much lower level. We may suppose that it is explained by the fact that both sources provide only a limited amount of evidence. The Google Blog Search API returns at maximum 8 results, so all candidate experts mentioned more than 8 times in blogs are regarded equal. Google Book search basically allows us to distinguish only between noted specialists and does not provide us with all sorts of academic expertise evidence.

The performance of each ranking is presented in Table 5.2. We see that

| **YahooWeb +** | MAP | MRR | P@5 |
|---|---|---|---|
| Enterprise | **0.460** | **0.604** | 0.240 |
| YahooWebAU | 0.390 | 0.483 | 0.224 |
| YahooWebPDF | 0.402 | 0.525 | 0.208 |
| YahooNews | 0.406 | 0.543 | 0.232 |
| GoogleBlogs | 0.427 | 0.562 | 0.223 |
| GoogleBooks | 0.452 | 0.567 | **0.244** |

**Table 5.3**: The performance of combinations of the **YahooWeb** ranking with the other rankings

| | MAP | MRR | P@5 |
|---|---|---|---|
| YahooWebPDF + GoogleBooks | 0.440 | 0.567 | 0.232 |
| YahooNews + GoogleBlogs | 0.420 | 0.571 | 0.216 |

**Table 5.4**: The performance of additional combinations inferring better Academic and Social Media evidences

restricting the scope of web search to the regional web or to specific file format does not lead to better results. Both **YahooWebAU** and **YahooWebPDF** rankings are inferior to **YahooWeb** ranking and to **Enterprise**. However, all other rankings built on web evidence are better than **Enterprise** in terms of MAP and MRR measures. It is hard to decide which of them is the best: **YahooWeb** is much better in MAP and P@5, but if user needs to detect the most knowledgeable person fast, using evidence from news and blogs seems a better idea according to the performance of the MRR measure. **GoogleBlogs** ranking outperforms the baseline only slightly, so its use without combining it with other evidences is questionable.

We also experimented with combinations of rankings (see Section 5.1.7). Following the principle that we should give a priority to the best rankings, we combined the most effective **YahooWeb** ranking with each other ranking (see Table 5.3). We surprisingly found that the combinations of that ranking with **Enterprise** and **GoogleBooks** rankings, which are not the best alone, are the best performing. Probably, since according to the normalized Kendall tau distance (see Table 5.1) these rankings are more similar to **YahooWeb** ranking, their combination produces a more consistent result. We also combined **Enterprise** ranking with each other ranking, but found that its combination with **YahooWeb** ranking is still the best.

In order to study the future potential of web evidence combinations, we decided to simulate the inference of web evidences which we can not currently acquire through APIs. First, we combined **YahooWebPDF** and **GoogleBooks** rankings to infer a better academic search based evidence. Consider-

| **YahooWeb + Enterprise +** | MAP | MRR | P@5 |
|---|---|---|---|
| YahooWebAU | 0.463 | **0.606** | 0.240 |
| YahooWebPDF | 0.446 | 0.589 | 0.240 |
| YahooNews | **0.468** | 0.600 | **0.252** |
| GoogleBlogs | 0.452 | 0.591 | 0.244 |
| GoogleBooks | 0.449 | 0.597 | 0.232 |

**Table 5.5**: The performance of combinations of the **YahooWeb** and the **Enterprise** rankings with the other rankings

ing that a lot of official and unofficial publications are publicly accessible in PDF format, we hoped to simulate the output of Google Scholar-like search service. As we see in Table 5.4, the performance of that combined ranking approved our expectations: it is better than each of these rankings used alone. Second, we tested the combination of **YahooNews** and **GoogleBlogs** rankings considering that it would represent an output from some future Social Media search service as it is envisioned by many (Firestone et al., 2007). The advantage of this combination is visible, but less obvious. It is certainly better than **YahooNews** ranking, but outperforms **GoogleBlogs** ranking only according to the MAP measure.

As we see in Table 5.5, further combination showed that when we combine **Enterprise** ranking, **YahooWeb** ranking and **YahooNews** ranking, we get improvements for the MAP and the P@5 measures. In total using that combination we had 29% improvement of MAP, 20% of MRR, and 14% of P@5. Combinations of 4 and more rankings only degraded the performance. To test statistical significance of the obtained improvement, we calculated a paired t-test for each measure. Results indicated that the improvement is significant at the $p < 0.01$ level with respect to the baseline.

## 5.2 Measuring the quality of a web search result

In the previous section we demonstrated a simple, but effective measure using web-based evidence for personal expertise. We counted the number of information units in a web source that contain all query terms and a candidate mention. Since we consider every link returned by a search service as a partial evidence of personal expertness, the next step would be to differentiate the strength of these evidences by taking various properties of these links into account. After all, the majority of expert finding approaches measures the quality of a person-specific result set returned by the search engine by summing document relevance probabilities. Generally speaking, in order to

estimate the overall $Quality$ of the person-specific search $Result_e$, one usually aggregates calculated quality measures over result items from the $Result_e$:

$$Quality(Result_e) = \sum_{Item \in Result_e} Quality(Item) \tag{5.2}$$

Besides that relevance is the main document quality measure used by popular document-based methods 2.1.2, query-independent measures are also recently studied. (Macdonald et al., 2008a) reported findings for the enterprise data only (e.g. all inlinks are only from pages of the Enterprise (CSIRO) domain). They used their expert finding method (described in Section 2.1.2) as a baseline. Using Inlinks and URL length improved MAP by a few percents. Similar document quality measures for document retrieval task can be found in some groups' reports on TREC Enterprise Track 2007 (Zhu et al., 2007; Duan et al., 2007; Wu et al., 2007). Measuring the quality of web result set to predict users' satisfaction with a search engine was proposed by White et al. (2008).

We decided to approach this problem from a different point of view. First, we measure the quality of the global web evidence, since it has shown itself to be so valuable in the previous section. Second, we again rely on rank aggregation to combine evidence originated from the enterprise and the web. Alternative approaches are used by Balog and de Rijke (2009) and He et al. (2009). They linearly combined retrieval scores for personal profiles as pseudo-documents built on the enterprise and web data. In one case, only web result snippets were used, in another case, entire documents were downloaded.

A result set returned by a typical web search engine consists of a list of result items described by their URLs, titles and summaries (snippets). Certainly, downloading web pages using URLs of web result items for the deeper analysis of web result quality may lead to the better perfomance, but in our experiments we restrict ourselves to quality measures calcualted just from the search result pages or using such page statistics that can be quickly acquired from a search engine without downloading the full content of a page. All measures that we considered in this paper could be classified into two types: query-dependent and query-independent.

### 5.2.1 Query independent quality measures

In our experiments we focused on four kinds of query independent quality measures of a result item (web page).

### URL length

Previous studies indicated that URL length is inversely proportional to the usefulness of the page it refers to (Kraaij et al., 2002; Craswell et al., 2005b). We apply a simple quality measure based on this assumption: $Quality(Item) = 1/Length(Item_{URL})$. The URL length is expressed in levels: the number of backslashes in the URL after its domain part. It should be mentioned that expressing the URL length in symbols performed much worse in our preliminary experiments.

### Inlinks for domain

Another quality estimate we used is an approximation of the result item's authority. It was recently proposed to measure the strength of expertise evidence extracted from a web page by the number of its inlinks (Macdonald et al., 2008a). There are web services providing similar statistics: Yahoo! Search API (Site Explorer) returns the number of inlinks for a provided URL, sites like `Prchecker.info` even show the estimate of Google PageRank. Academic search engines like Google Scholar usually return the number of citations per publication in their result set.

However, it was hard to calculate sophisticated web graph centrality measures without downloading the content of all web pages returned in results sets of all test queries. Moreover, we were interested in a lightweight solution. Since most pages are not often linked by pages outside of their domain, we used a simple inlink authority measure for the domain of the result item, considering that in many other authority measures (e.g. Pagerank) this value anyway propagates to all pages hosted at the result item's domain: $Quality(Item) = Inlinks(Domain(Item))$. The authority estimate was acquired using the ***link:*** clause plus the domain name to query Google Web Search API that returned the number of pages citing the given domain.

### Domain size

We also supposed that the importance of the domain which hosts the returned result page should also be expressed by its size: $Quality(Item) = Size(Domain(Item))$ The main intuition was that large domains usually become so only due to the time and money spent on their maintenance what in turn demonstrates their respectability. The size estimate was acquired using ***site:*** clause plus the domain name to query Google Web Search API that returned the number of pages indexed by Google at the given domain.

#### Freshness

We supposed that a page's last date of modification shows how much trust we should put in expertise evidence found in it. Supposedly, the freshness of expertise evidence implicitly indicates the freshness of candidate's expert knowledge. In our preliminary experiments it appeared that considering only those results that where at least once modified (or created) after 2006 was better than just treating all of them equally useful:

$$Quality(Item) = \begin{cases} 1, Year(Item) \geq 2006 \\ 0, Year(Item) < 2006 \end{cases}$$

### 5.2.2 Query dependent quality measures

The state-of-the-art methods, including the one we use to get Enterprise based ranking (Balog et al., 2006), often rank candidates by the sum of relevance probabilities of pages that contain their mentions (see Section 2.1.2). In our case, it is possible to issue a query without a person's name within and get only topic based ranks of documents. But since most engines return only first thousand of matched pages, that strategy may fail for non-selective short ambiguous queries producing significantly larger result.

Since it is also very time- and broadband-consuming to download all pages in the result list in order to measure their relevance, we use a very simple measure of the *Item*'s (URL, Title or Summary) relevance which we sum over the result list:

$$Quality(Item) = \frac{N(q, q \in Item \wedge q \in Q)}{|Q|} \tag{5.3}$$

what is the number of query terms $q$ appearing in the result $Item$ divided by the number of terms in the query $Q$. Since it is hard to tokenize URLs, we just search for a query term as for a substring in this case.

### 5.2.3 Experiments

We again used the CERC collection as in experiments described in previous section. The results analysis is again based on calculating popular IR performance measures also used in official TREC evaluations. We analyzed the performance of the **Enterprise** ranking combined with one of the following rankings:

- **YahooWeb**: based on the number of web result items returned,

- **YahooWebURLLenInLevels**: based on the sum of URL Length based quality estimates for web result items,
- **YahooWebInlinksForDomain**: based on the sum of inlinks of domains of web result items,
- **YahooWebSizeForDomain**: based on the sum of sizes of domains of web result items,
- **YahooWebAfter2006**: based on the number of web result items modified or created after 2006,
- **YahooWebRelevURL**: based on the sum of URL relevance probabilities for web result items,
- **YahooWebRelevTitle**: based on the sum of title relevance probabilities for web result items,
- **YahooWebRelevSummary**: based on the sum of summary relevance probabilities for web result items,

We wanted to be sure that the difference in performance of the result size based method and the other methods using summed quality estimates of result items do not occur due to random assignment of ranks for items with equal values. So, we assigned these ranks more fairly than in the Section 5.1, when we just relied on a sorting algorithm for rank assignment. First, we considered that all candidates with zero expertise estimates are always assigned with the lowest negative rank possible in the system (-100 in our experiments, since we always start by taking top-100 candidates from the Enterprise based ranking). Second, we assigned equal ranks to the candidates with equal estimates. The comparison of results presented in Table 5.6 to the previous results, presented in Table 5.3, shows that the removal of this randomness certainly leads to the improvement.

Our initial intention was to improve baseline **Enterprise** ranking and **Enterprise+YahooWeb** rankings combination that we regard in this section as our new actual baseline. Only **YahooWebURLLenInLevels** ranking showed significantly degraded performance, the others were equally or better performing. Three rankings appeared to have slightly better performance in combination with **Enterprise** ranking: **YahooWebAfter2006**, **YahooWebRelevTitle**, **YahooWebRelevURL**. In the latter case, the result was also significantly better than for **Enterprise+YahooWeb** ranking for MAP and MRR measures at the level $p < 0.05$ (paired t-test). We also tried to further combine different rankings from the above list. However, we did not succeed to beat **Enterprise+YahooWebRelevURL**'s ranking performance with any of these combinations.

| **Ranking** | MAP | MRR | P@5 |
|---|---|---|---|
| **Enterprise** | 0.362 | 0.508 | 0.220 |
| **Enterprise +** | | | |
| **YahooWeb** | 0.485 | 0.627 | 0.256 |
| **YahooWebURLLenInLevels** | 0.386 | 0.532 | 0.216 |
| **YahooWebInlinksForDomain** | 0.477 | 0.632 | 0.252 |
| **YahooWebSizeForDomain** | 0.477 | 0.604 | 0.248 |
| **YahooWebAfter2006** | 0.491 | 0.620 | 0.256 |
| **YahooWebRelevURL** | 0.501 | 0.650 | 0.26 |
| **YahooWebRelevTitle** | 0.488 | 0.634 | 0.26 |
| **YahooWebRelevSummary** | 0.485 | 0.627 | 0.252 |

**Table 5.6**: The performance of quality-aware combinations of the web and enterprise based rankings

| **Ranking** | MAP | MRR | P@5 |
|---|---|---|---|
| **Enterprise +** | | | |
| **YahooWeb** | 0.371 | 0.740 | 0.469 |
| **YahooWebAfter2006** | 0.370 | 0.743 | 0.458 |
| **YahooWebRelevURL** | 0.373 | 0.765 | 0.487 |
| **YahooWebRelevTitle** | 0.371 | 0.754 | 0.480 |

**Table 5.7**: The performance of TREC 2008 queries

We finally submitted combinations of **Enterprise** ranking with **YahooWeb**, **YahooWebAfter2006**, **YahooWebRelevTitle**, and **YahooWebRelevURL** rankings as runs to TREC 2008 (see Table 5.7). The only difference with experiments with TREC 2007 queries is that we used our own infinite random walk based expert finding method Serdyukov et al. (2008d) to build the **Enterprise** ranking. In this case all methods were equally effective according to MAP measure, but according to MRR and P@5 measures, considering relevance of URLs was indeed beneficial.

## 5.3 Discussion

As it was demonstrated by our experiments, we are able to gain significant improvements over the baseline expert finding approach which analyzes only the data originating from the organization. We found again, as in Chapter 4, that the quality of inference of personal expertise depends on the amount of expertise evidence. When we search for the indirect evidence also outside of an organization on the World Wide Web, we increase our potential to guess

about the competence of its employees. It was also clear from experiments that combining different sources of evidence through simple rank aggregation allows to improve even more. We suppose that this improvement is first of all caused by diminishing the dominance of persons that appear in organizational documentation accidentally or by bureaucratic reasons (e.g. web-masters or secretaries). Such persons are often not related to the very meaning of most documents where they appear and are frequent only in the documents from one single source. Additional studies indicated not only the benefit from using diverse sources of expertise evidence, but also from accurate measuring the quality of evidence acquired from these sources.

In this thesis we focused our studies on the predefined subset of search services selected by their popularity and supposed richness in expertise evidence. However, there are more potentially useful sources that can be found among social networks, expert databases, vertical search engines and Web 2.0 applications built on the content generated by users. Some of them are already popular among professionals and therefore usable for expert finding. Other ones are just on the rise of their authority.

**Social Networks.** Social networks are essential sources of knowledge about personal skills and experience. They allow to extract expertise evidence not solely from a user profile, but also from its context: directly "befriended" user profiles or profiles connected implicitly through sharing the same attributes (e.g. places of work or visited events). However, while such huge networks as `LinkedIn.com` (33 million members as reported in January 2009) and `Facebook.com` (130 million members as reported in January 2009) are very popular for recruiting specialists (King, 2006; Kolek and Saunders, 2008), it is still hard to compare expertise of employees from the same organization using the data acquired from these sources. In many cases on-line profiles are not publicly accessible and still far not all employees care about being well-known.

**Expert databases.** Those who are not willing to create their own professional profile manually can be supplied with one for free. Such repositories of experts as `Zoominfo.com` and many others (Arrington, 2007) automatically summarize all information about people found on the Web to make them searchable. Many of them provide APIs for programmatic access to their databases[7].

**Vertical Search Engines.** Specialized topic-oriented search engines should be helpful for finding experts in specific industries: `SearchFinance.com` - for finding economists, `Medstory.com` - for doctors, Yahoo! Tech[8] and

[7] `www.programmableweb.com/apis/`

[8] `tech.yahoo.com`

Google Code Search[9] - for software engineers etc.

**User generated content.** There are other ways to share expertise besides blogging. Communities like `Slideshare.com` allow knowledge exchange with the minimum effort by just uploading personal presentation slides. However, professional advice at Yahoo![10] or LinkedIn[11] Answers or authoring Wikipedia articles (Demartini, 2007) are more significant activities. They indicate personal proficiency not only by relevance of the generated content, but also by feedback from involved users assessing the quality of advice (Adamic et al., 2008).

## 5.4 Summary

In this paper we proposed a way to gather additional expertise evidence apart from that available in organizations employing candidate experts. We used various kinds of web search services to acquire an additional proof of expertness for each person which was initially pre-selected by an expert finding algorithm using only organizational data. We basically developed two approaches to acquisition of expertise evidence that can be found on the web.

In our first approach, we used APIs of two major search engines, Yahoo! and Google, and built six rankings of candidate experts per query using different vertical search services. We empirically demonstrated that rankings from certain web sources of expertise evidence and their combinations are significantly better than the initial enterprise search based ranking. The presented study demonstrated that the predicting potential of the expertise evidence that can be found outside of the organization is invaluable. We discovered that combining the ranking built solely on the Enterprise data with the web based ranking may produce significant increases in performance. In our second approach, we tried to further improve the performance by using various quality measures to distinguish among web result items. While, indeed, it was beneficial to use some of these measures, it stayed unclear whether they are decisively important.

[9] `codesearch.google.com`
[10] `answers.yahoo.com`
[11] `linkedin.com/answers`

# 6

# Beyond expert finding

Expert finding is an example of the task that can be abstractly formulated as ranking entities that do not have their own descriptions, but are in certain and uncertain relations with some pieces of text in the collection. So, it comes natural, that after we had proposed improvements for expert finding, we applied similar techniques to other entity ranking tasks.

In this chapter, we answer research questions posed in the beginning of this thesis and related to **Research Objective 4** (see Section 1.2). We present first, yet effective, solutions for two novel tasks of entity ranking in specific domains. The first task is ranking entities in user-generated knowledge repositories, which is the most similar task to expert finding in our opinion. The application example that we use in this thesis is Wikipedia. The second task is location prediction for images using only their short descriptions (tags). As we show further this problem can be reduced to ranking locations (entities) in respect to the tagset (query) of an image that we need to place on a map. We focus our studies on images uploaded to photo-sharing services and particularly, Flickr. We can also imagine a very similar functionality within an expert finding application, which helps users to find a region with high expertise on the topic, i.e. with high density of experts or companies working in relevant areas.

## 6.1 Entity Ranking in Wikipedia

Entity ranking in Wikipedia is a task similar in nature to web search (where Wikipedia pages dominate in results), but has some distinguishing features. Queries here ask not for documents, but for a ranked list of *unique* entities, e.g. for movies, flags, or diseases, described by short labels and possibly but not necessarily by URLs of pages with detailed information about them.

| **Topic** | **#74** |
|---|---|
| Title | circus mammals |
| Description | I want a list of mammals which have ever been tamed to perform in circuses. |
| Narrative | Each answer should contain an article about a mammal which can be a part of any circus show. |
| Category | Mammals |
| Relevant | Asian Elephant, Brown Bear, Lion |
| Irrelevant | Gorilla |

**Table 6.1**: An example of INEX Entity Ranking topic

Users of entity ranking systems search for entities in the first place, rather than for any text which is "about" them. This in turn means that the relevance of pages describing entities is less of a concern for users than relevance of entities, although it may be helpful to find these entities. Most importantly, it leads to the need to estimate also the relevance of items that do not have any description (Zaragoza et al., 2007). If users still expect the system to return pages, like in Wikipedia-based entity search, they anyway want pages not just "about" the relevant entities, but pages whose primary purpose is to serve as a unique and complete description of an entity.

Entity ranking stresses on the fact that users experience *typed* information needs. So users search not for all kinds of relevant entities, as in default web search settings, but for their specific types. The type of an entity is defined in the context of Wikipedia by *categories* assigned to the entity's article. An entity can thus have several types. Furthermore, Wikipedia categories are hierarchically organized (although there are cycles (Zesch and Gurevych, 2007)). We can hence assume that an entity does not only belong the categories assigned to it, but also to ancestor categories. However, Wikipedia's category hierarchy does not form a strict tree, and thus moving too far away from the original categories can lead to unexpected type assignments.

To evaluate retrieval systems handling typed information needs for entities, the Initiative for Evaluation of XML Retrieval (INEX) started the XML Entity Ranking track (INEX-XER), with the aim to create a test collection for entity retrieval in Wikipedia (Vri, 2008). While the collection is the same that INEX uses for other tasks, e.g. ad-hoc XML Retrieval, test topics and relevance judgments are specifically designed for experiments with entity ranking (see example in Table 6.1).

Our approach to entity ranking can be summarized by the following processing steps:

(1) initial retrieval of articles,
(2) building of an entity graph,
(3) relevance propagation within the graph,
(4) filtering articles by the requested type.

The notion *entity graph* stands here for a query-dependent link graph, consisting of all articles (or entities) returned by the initial retrieval as vertices and the link-structure among them forming the edges. Links to other articles not returned in the initial ranking are not considered in the entity graph. *Entity graphs* can later be used for the propagation of relevance to neighboring nodes and serve the same purpose as *expertise graphs* from Chapter 4.

### 6.1.1 Entity retrieval by description ranking

The most simple and obvious method of entity retrieval could be the ranking of their textual descriptions with some classic document retrieval method. In case of expert finding, it is hard to expect from each and every candidate expert in the enterprise to maintain personal profile. At the same time, it is often not clear to what extent an expert finding system should trust such self-descriptions. Contrariwise, Wikipedia articles are collaboratively maintained by users. Their immense enthusiasm makes it possible to have an article for almost any more or less known entity. Discussions and change histories are public what decreases the risk of favoritism and spamming.

In our experiments we rank Wikipedia articles representing entities using a language-model based retrieval method:

$$P(Q|e) = \prod_{t \in Q} P(t|e), \tag{6.1}$$

$$P(t|e) = (1 - \lambda_C)\frac{tf(t,e)}{|e|} + \lambda_C \frac{\sum_{e'} tf(t,e')}{\sum_{e'} |e'|} \tag{6.2}$$

where $tf(q,e)$ is a term frequency of $q$ in the entity description $e$, $|e|$ is the description length and $\lambda_C$ is a Jelinek-Mercer smoothing parameter - the probability of a term to be generated from the global language model. In all our experiments it is set to 0.8, what is standard in retrieval tasks.

### 6.1.2 Entity retrieval by relevance propagation

Ranking entities using solely the content of their descriptions as the evidence of their relevance is a reasonable, but not the only step to be done towards

maximizing performance. Thousands of entities in Wikipedia have too short or empty descriptions, especially those that appear in novel evolving domains and just became known [1]. Wu et al. (2008) report that among the 1.8 million pages they crawled in July 2007, many are short articles and almost 800,000 (44.2%) are marked as stub pages, indicating that much-needed information is missing. For many popular queries implying an information need for widely known entities it may be not the issue, but unobvious queries with narrow focus may often fail to find a barely relevant entity. Moreover, even well-known entities are often described by associations with other entities and in terms of other entities. This means that query terms have lesser chance to appear in the content of a relevant description, since some concepts mentioned in its text are not explained and their details can be found in their own descriptions. It is known that internal links in Wikipedia normally link to another "relevant" Wikipedia page and this is imposed by official guidelines for contributors: "*Only make links that are relevant to the context...*". These considerations motivated us to continue adopting methods that we used for relevance propagation in Chapter 4.

#### Finite random walk

In order to model the relevance propagation between entities (articles), we model the process in which the user, after seeing initial list of retrieved entities:

- selects one document and reads its description,
- follows links connecting entities and reads descriptions of related entities.

Since we consider this random walk as finite, we assume that at some step a user finds the relevant entity and stops the search process. So, we iteratively calculate the probability that a random surfer will end up with a certain entity after $K$ steps of a walk started at one of the initially ranked entity. In order to emphasize the importance of entities to be in proximity to the most relevant ones according to the initial ranking, we consider that both (1) the probability to start the walk from a certain entity and (2) the probability to stay at the entity node are equal to the probability of relevance of its description. This finite random walk is similar to the one we adopted for expert finding task in Section 4.1.4.

$$P_0(e) = P(Q|e) \tag{6.3}$$

[1] `wikipedia.org/wiki/Wikipedia:WikiProject_Missing_encyclopedic_articles`

$$P_i(e) = P(Q|e)P_{i-1}(e) + \sum_{e' \to e}(1 - P(Q|e'))P(e|e')P_{i-1}(e'), \tag{6.4}$$

The probabilities $P(e|e')$ are uniformly distributed among links outgoing from the same entity. Finally, we rank entities by their $P_K(e)$.

**Linear Combination of Step Probabilities** It is also possible to estimate entity relevance using several finite walks of different lengths at once. In the following modification of the above-described method, we rank entities considering a weighted sum of probabilities to appear in the entity node at different steps:

$$P(e) = \mu_0 P_0(e) + (1 - \mu_0) \sum_{i=1}^{K} \mu_i P_i(e), \tag{6.5}$$

where $\mu_i$ is the prior probability that a random surfer stops the walk at $i$ step. Of course, the faster an article is "found" during the described random walk, the higher chance that it is relevant. However, for the sake of simplicity, we set $\mu_0$ to 0.5, distribute $\mu_1 \dots \mu_K$ uniformly and rely exclusively on varying number of steps $K$ in our experiments.

### Infinite random walk

In our second approach, we assume that the walk in search for relevant entities consists of countless number of steps. The stationary probability of ending up in a certain entity is considered to be proportional to its relevance. However, since the stationary distribution of a described discrete Markov process does not depend on the initial distribution over entities, so the relevance flow becomes unfocused. The probability to appear in a certain entity node becomes dependent only on its centrality, but not on its closeness to the sources of relevance. In order to solve this issue we introduce regular jumps to entity nodes from any node of the entity graph after which the walk restarts and the user follows inter-entity links again. We consider that the probability of jumping to the specific entity equals to the probability of relevance of its description. This makes a random walker visit entities which are situated closer to the initially highly ranked ones more often during normal walk steps. The following formula is used for iterations until convergence:

$$P_i(e) = \lambda_J P(Q|e) + (1 - \lambda_J) \sum_{e \to e'} P(e|e')P_{i-1}(e') \tag{6.6}$$

$\lambda_J$ is the probability that at any step the user decides to make a jump and not to follow outgoing links anymore. The described discrete Markov

process is stochastic and irreducible, since each entity is reachable due to introduced jumps, and hence has a stationary distribution. Consequently, we rank entities by their stationary probabilities $P_\infty(e)$. If to compare infinite and finite random walks, here we allow all entities with even a long path to some entity to influence its relevance estimate.

### 6.1.3 Experiments

The collection used for experiments with entity ranking is the Wikipedia XML Corpus based on an XML-ified version of the English Wikipedia of early 2006. Despite that actual Wikipedia was four times larger in size by the end of 2008[2], INEX collection still represents a significantly large experimental sample of 659,338 Wikipedia articles organized into 113,483 categories.

We trained and analyzed our models using those 28 queries from the Ad-Hoc XML Retrieval task of INEX 2006. Answers that are not the dedicated descriptions of relevant entities were removed. All our algorithms start from retrieval of articles from the collection using the baseline language modeling based approach to IR for scoring documents. Further we extract entities mentioned in these articles and build entity graphs. For the initial article retrieval as well as for the graph generation the PF/Tijah retrieval system was employed (Hiemstra et al., 2006). We tuned our parameters by maximization of the MAP measure and for 100 initially retrieved articles.

Experiments with the following methods are discussed further:

- **Baseline**: the baseline method ranking entities by the relevance probabilities of their Wikipedia-articles (see Equations 6.1, 6.2),
- **K-Step RW**: the K-step Random Walk method using multi-step relevance propagation with K steps (see Equations 6.3, 6.4),
- **K-Step RWLin**: the K-step Random Walk method using linear combination of entity relevance probabilities at different steps up to K (see Equation 6.5),
- **IRW**: the Infinite Random Walk method ranking entities by probabilities to reach them in infinity during non-stop walk (see Equation 6.6).

For the Entity Retrieval task we had a query and the list of entity categories as input. However, according to the track guidelines and our own intuition, relevant entities could be found out of the scope of given categories. Preliminary experiments have shown that using parent categories of any level spoiled the performance of the baseline method. However, it was

[2] `wikipedia.org/wiki/Wikipedia:Modelling_Wikipedia's_growth`

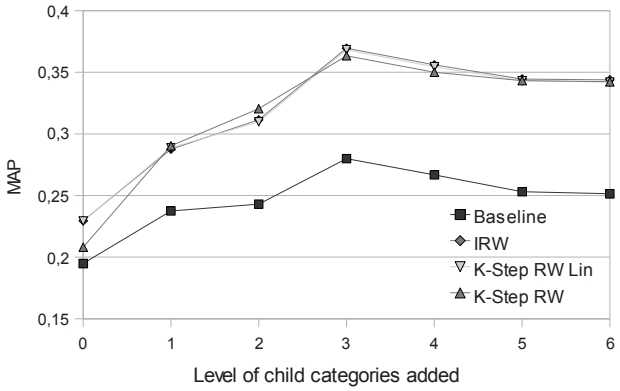



**Figure 6.1**: MAP performance of all methods for different levels of child categories added

important to include child categories up to the 3rd level (see Figure 6.1). This probably means that queries were created with an assumption that given categories should be greatest common super-types for the relevant entities. We used entities of all categories for the relevance propagation and filtered out entities using list of allowed categories only at the stage of result list output.

In all methods except the Baseline we had to tune one specific parameter. For the **K-step RW** and **K-step RWLin** methods we experimented with the number of walking steps. As we see in Figure 6.2 both methods reach their maximum performance after making 3 steps. **K-step RW Lin** method seems to be more robust to the parameter setup. It probably happens because it smooths the probability to appear in the certain entity after K steps with probabilities of visiting it earlier. The rapid decrease of performance for even steps for **K-step RW** method can be explained in the following way. A lot of relevant entities are only mentioned in the top ranked entity descriptions and do not have their own descriptions in this top, due to their low relevance probability or due to their absence in the collection. The relevance probability of these "outsider" entities entirely depends on the relevance of related entities, which are not relevant entities themselves (for example, do not match the requested entity type), but tell a lot about the ranked entity. So, all 'outsider' entities have direct (backward) links only to the entities with descriptions in the top and since we always start walking only from the latter entities, the probability to appear in 'outsider' entities at every even step is close to zero.

We also experimented with the probability to restart the walk from ini-

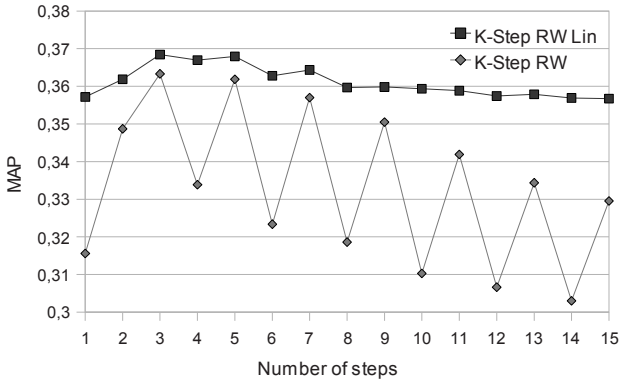



**Figure 6.2**: MAP performance for two methods and different numbers of steps

| **Method** | **MAP** |
|---|---|
| Baseline | 0.291 |
| K-step RW | 0.281 |
| K-step RWLin | 0.306 |
| IRW | 0.301 |

**Table 6.2**: MAP of 4 methods on the test data

tially ranked entities for the IRW method. According to results shown in Figure 6.3, values between 0.3 and 0.5 seem to be optimal. This actually means that making only 2-3 steps (before the next restart) is the best strategy what is also the case for the finite random walk methods.

To sum the things up, our experiments with the training data showed that all our three methods outperform the **Baseline** method. However, the **K-Step RW** method produced a bit worse results than the other two. In spite of this observation, it was necessary to examine the effectiveness of relevance propagation on the test data, fixing the best parameters tuned on the training 28 queries. We conducted the final experiment with 46 test queries (Vri, 2008).

Herewith, we present the performance of 4 sets of resulting entity lists submitted to the Entity track of INEX 2007. Table 6.2 shows that the performance instability of **K-step RW** method observed on the training data probably resulted in its low performance on the test data. However, both **K-step RWLin** and **IRW** methods equally outperformed **Baseline**. Their performance was also quite similar on the training data. That lets to conclude that the influence of entities situated further than in 3 links (number

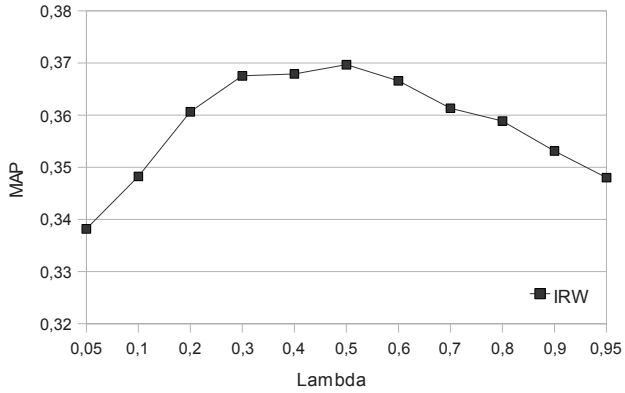



**Figure 6.3**: MAP performance of IRW method for different values of jumping probability

of steps for maximum performance of the **K-step RWLin**) from the current entity is marginal, when the probability of not following links at each step is as high as 0.5. In other words, it leads to empirical proof that the jumping probability is enough to control the size of a neighborhood that affects the relevance of an entity.

The main conclusion we arrived at is that the relevance propagation is an appropriate and beneficial mechanism for entity ranking even in such a semantically rich and well-structured corpus as Wikipedia with a complete textual description for almost every registered entity. While there is no classification of INEX queries into easy ones, asking for obviously popular entities, and hard ones, used for highly focused search, we expect the latter type of queries to take more advantage of relevance propagation. We also realize that Wikipedia is a user-generated encyclopedia successfully approaching the goal to describe every entity ever known to exist. This means that it solves the problem of entity search, which mainly consists in the lack of textual description for entities, by means of collaborative effort from users. In collections like we used for expert finding evaluation the benefit from multi-step relevance propagation is higher (see Section 4.3, Table 4.1), obviously due to the fact that reliable sources of expertise (relevance) evidence are not well-defined and need to be found automatically (e.g. personal pages/resumes of the candidates are not explicitly given). Nevertheless, the performance of runs produced by the **K-step RWLin** and the **IRW** methods was the best among 18 runs submitted by 8 groups participating in entity ranking track of INEX 2007 (Vri, 2008).

## 6.2 Placing Flickr images on a map

Nowadays, the number of context-aware Web 2.0 applications is constantly increasing. Due to the massive production of affordable GPS-enabled cameras and mobile phones (Naaman et al., 2004; Raper et al., 2007), location metadata such as *latitude* and *longitude* is automatically associated with the content generated by users. Users have the opportunity to spatially organise and browse their personal media (videos, photos or blog posts), and photo sharing services are leading the growing enthusiasm for personal location-awareness (Torniai et al., 2007). Geo-referenced photos can be organised in a browsable taxonomy of major locations or pin-pointed on a map to identify very small regions. Some of the most popular examples are Flickr Places[3] and Google Panoramio.[4]

While in theory every photo can be anchored to the location it was taken, in practice many photos are location agnostic. Furthermore, the majority of Flickr users does not own location-aware cameras. Thus a large proportion of photos uploaded to Flickr contains no location information, even when the photo merits localizing. When uploading photos on Flickr users can still geo-tag their photos by dragging the photos to a particular point on the World map. This process is time-consuming and results in less accurate geo-tagging of photos, compared to automatically geo-tagged photos from GPS-enabled cameras. When manually geo-tagging photos, Flickr initially suggests the location of the last uploaded photo or simply displays the World map. At the same time, users are likely to expect applications to be more pro-active by making some initial guess about the place where the photo was taken, since most users spend considerable effort to organize their "memory" geographically by describing photos with *tags* related to locations where they were taken (Ames and Naaman, 2007).

The objective of research presented in this section is to provide a more accurate starting point for geo-tagging photos, uploaded on Flickr, using the textual annotations provided by the user. According to recent literature (Ames and Naaman, 2007; Sigurbjornsson and van Zwol, 2008) users spend a considerable effort to organise their "memory" geographically by describing photos with *tags* related to locations where they were taken. The location specific tags (such as *Torre Agbar* which is only located in Barcelona), and location related tags (such as *elephants* which are related to locations such as zoos, Africa and Asia) provide essential cues as to where a picture was taken. For photos that are location agnostic (such as *dog*), location information may

[3]http://www.flickr.com/places/ visited January 2009
[4]http://www.panoramio.com visited January 2009

or may not be provided, but it is not relevant to the context of the photo.

The literature related to geo-tagging of photos and its use is extensive. In particular the reverse problem of discovering important landmarks and events, given a geographic co-ordinate has been studied extensively (Ahern et al., 2007; Rattenbury et al., 2007; Naaman et al., 2004). However, we believe that we are the first to investigate the problem of placing images on a map using the textual annotations provided by the user. While we focus on Flickr as our primary application, our approach can be applied to a wide range of service providers dealing with geo-referenced resources.

In this section we investigate generic methods for placing photos uploaded in Flickr on the World map. We construct an $n \times m$ grid based on the longitude/latitude co-ordinates of the globe, where each grid cell represents a location. Using a set of images whose locations are known, we place each image in its corresponding grid cell. As a baseline we employ the collective knowledge of Flickr users by estimating a language model from the terms people use to describe images taken at a particular location. We extend this model in several ways, using neighbouring cells under the assumption that "good" locations come from "good" neighbourhoods, and leveraging spatial ambiguity. Finally, we investigate how to incorporate external resources into the model, by boosting the importance of known location tags identified by their presence in GeoNames. We train, develop and evaluate our system using a snapshot of nearly 400,000 geo-tagged photos from Flickr with textual annotations. We predict the single most likely location of a photo in terms of accuracy at different levels of spatial granularity.

### 6.2.1 Spatial mining of user-generated content

The task of image classification exists for decades and all approaches can be roughly classified into two types: text-based and content-based (Datta et al., 2008). The text-based methods received considerably less attention, mainly due to common disbelief that images can be massively annotated in a manual, but cheap way. Nowadays, with the coming of new photo sharing services where users are highly motivated to generate descriptions for images to help their browsing and retrieval, various image text-based classification tasks become more realistic and doable, being already socially demanded. However, one work on content-based image classification is worth mentioning as the only research on image geo-mapping existing to the best of our knowledge. Hays and Efros (2008) proposed to use visual features of Flickr images to predict their geographic location with a nearest neighbor classification method. They report geo-locating 16% of test images within 200 km. Their data is limited to a sub-set of Flickr images tagged with at least

one name of a country, continent, densely populated city or popular tourist site and not tagged with specific non-geographic tags such as "birthday" or "concert". By contrast, our approach is potentially knowledge-free, highly scalable and not limited to photos that are known to contain locations in the textual annotations.

However, the task of finding a geographical focus of a web page was first proposed by Ding et al. (2000). Their approach was two-fold: finding locations of web pages with hyper-links to the analysed page and detection and disambiguation of toponyms in its content. Follow-on work by Amitay et al. (2004) and Zong et al. (2005) relied on propagating the confidence weights of found toponyms up to the root of the gazetteer taxonomy to find the most probable common ascendants (e.g. finding a *country* for several *cities* mentioned). Generally speaking, the research on finding geographical focus of text is almost entirely based on finding and resolving toponyms. For our task of placing images on a map, an obvious solution would be to simply resolve the toponyms in the Flickr tags. This would allow us to identify locations that are mentioned in the tags, but does not allow us to infer locations from tags that are not found in gazetteers or other resources. We briefly describe the state-of-the-art in toponym resolution. A complete and detailed description of toponym resolution heuristics is made by Leidner (2007), but most approaches are derived from the following ideas.

**Using location priors**. Even without any context it is possible to make an intelligent guess about the most probable referent for a toponym by just considering the prior probabilities of places. For example, places with larger populations, or more frequent mentions in text are more likely candidates.

**Search for disambiguators**. It is assumed that each place has a list of disambiguators, such as its neighbours in gazetteer hierarchy, which resolve it if found in the proximity of the place name mention. For example, the country of *France* and the state of *Texas* are both disambiguators for the city of *Paris*.

**Spatial minimality**. If several toponyms are mentioned in the text without disambiguators, then those places are selected as referents that minimise the *minimum bounding rectangle* containing them or the sum of their pairwise distances from each other. For example, if *Moscow* and *Helsinki* appear together in the text, then *Moscow, Russia* is selected instead of *Moscow, Idaho, US* since it is thousands of kilometers closer to Helsinki, and Helsinki is unambiguous.

Ranking locations by the probability of generating a tag set implies leveraging the above-mentioned approaches. Both spatial minimality and disambiguator-based methods are implicit in the language modeling approach because places are more likely to appear together with their disambiguators

in tags and tag sets from the same location. A significant number of users enlist disambiguators for place names in their tag sets by mentioning not only city, but also country tag, for example. However, as demonstrated on Figure 6.4, in addition to toponyms, other region-specific terms traditionally unnoticed by gazetteers can serve as disambiguators (e.g. *hermitage* for *St. Petersburg, Russia* and *pelican* for *St. Petersburg, Florida*). Using additional population and popularity based priors for locations is also not necessary due to the origin of tags representing locations: popular and highly populated locations get more image uploads and hence higher related tag counts.

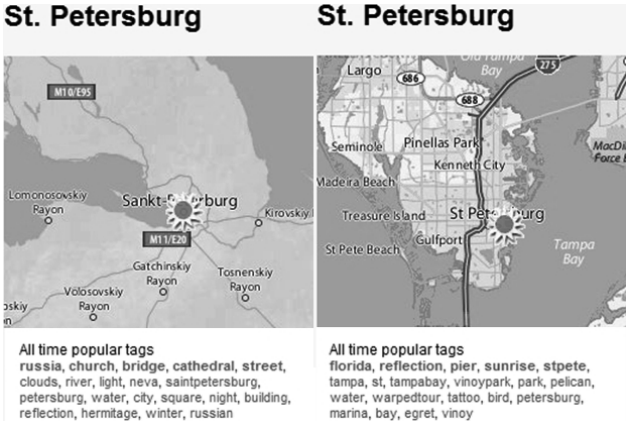



**Figure 6.4**: Tags for places with ambiguous names, generated with Flickr Places: http://www.flickr.com/places

A number of related approaches should also be mentioned to make the picture complete. In relation to the location identification of images, Ahern et al. (2007) propose a method for detection of Flickr tags exhibiting spatial patterns. They find dense areas using geodesic distances between images, and rank all tags in these areas with a $tf.idf$-based feature selection measure to select the most representative location-related tags. The focus of their research is on selecting tags, rather than localizing images. In later papers they propose a method for detection of tags that correspond to local events (Rattenbury et al., 2007). Naaman et al. (2003) look at the other side of the coin, recommending tags to the user, given a known location for an image. Working with blog data, Mei et al. (2006) present methods for finding latent semantic topics and their distribution over locations (states or countries) and Wang et al. (2007) propose a *Location-Aware Topic Model* based on Latent Dirichlet Allocation. The works are similar to ours in that they

attempt to discern whether a topic is location-related, however blog data has a considerably richer semantic representation than the tags associated with images on Flickr, which may be only two or three terms. Web queries are more similar to tags than blogs, in the sense that queries are two or three content terms representing much larger concepts. Backstrom et al. (2008) propose a method to measure the geo-specificity of a query, using the level of dispersion around the location of the query's highest frequency. With a similar goal in mind, Zhuang et al. (2008) calculate the inverse correlation of a query's click distribution over locations with their populations. Vadrevu et al. (2008) use the probability of co-occurrence of a query term with place names from each region to determine queries that might be related to a given region.

### 6.2.2 Representing locations on a map

For each photo in our collection we have the following sources of information: a *FlickrID*, a *geographic co-ordinate*, and a *set of tags*. The first task at hand is to map each geographic co-ordinate to a location on the map. For that purpose we overlay a grid over the World map, which allows us to define a location as a cell on this grid, described by a pair of universal geographical co-ordinates (UGC). We can then investigate the accuracy of our system given various levels of spatial granularity of the grid. To universally represent locations we bind them to a cell using their latitudes and longitudes, considering 0 - 2 digits of the decimal part (UGC are represented in Flickr as decimals). For example, each pair of co-ordinates, ignoring the decimal part and considering only *degree* units, defines a unique *location*: a rectangle, or a cell of the earth grid, with latitude side of about 111 kilometers long and variable longitude side. The length of longitude side depends on latitude and varies from 0 kilometers at the poles (but 60 starting from inhabited places) to 111 kilometers at the equator (Toyama et al., 2003). In this paper we consider locations (cells) of roughly 1, 5, 10, 50 and 100 kilometers long over latitude (see Figure 6.5).

Alternatively, one can tie a geographic location to a known semantic location as defined in Gazetteers. There locations are named and either defined by a bounding box (GeoPlanet[5]), or just by a point on the map (GeoNames). In the first case, bounding boxes are often overlapping and do not cover all regions where people may appear. In the second case, it is unclear how to define boundaries in an unsupervised way. Moreover, developing methods dependent on specific symbolic representation of geographical information

[5]geoplanet.yahoo.com

may lead to difficulties in maintenance of such a framework, including the need to rebuild models with every new update of underlying topology. At the same time, when the most probable geographic location of an image is determined, mapping this location onto a specific geographical ontology (*reverse geo-coding*) is almost straightforward.

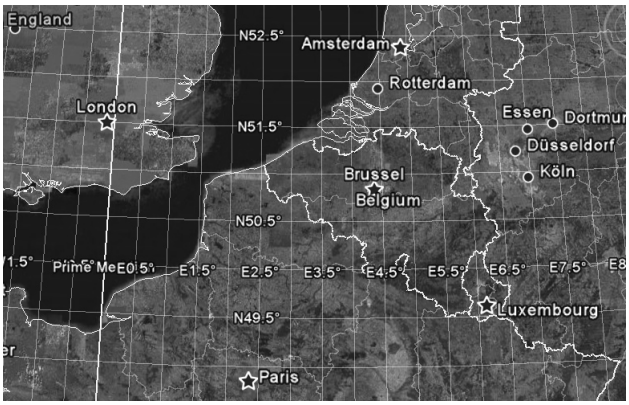



**Figure 6.5**: Western Europe divided into 50 km cells.

After defining the boundaries of locations, we need to represent them by some features useful for Flickr images classification. Note that most Flickr images are described only with tags, with no additional text attached: detailed descriptions and user comments are often missing and titles are often assigned automatically and hence meaningless (e.g. *pho1232.jpg*). To consider only the most reliable and most concise descriptions of photos, we classify Flickr images using their tags only. Each photo has associated 1 or more textual tags which is used to derive a language model that represents a location. We assume that the order of tags is not important for placing images on a map and adopt a bag-of-tags approach to sample representative tags for a given location. In order to preserve the unique semantics of each tag set we do not apply any stemming or stop-word filtering. However, we use the standard tag normalisation automatically provided by Flickr: all terms in compound tags are concatenated and all special characters are stripped.

In web retrieval it is common to assume various relations among documents defined by hyper-links. In our case we consider that the grid structure underlying our collection of locations implies a spatial relationship. Based on these observations, we represent all locations in an undirected graph, where the link between a pair of locations (grid cells) exists only if they are situated close enough on the grid. For the sake of simplicity, we use cell-based

distance and consider that any cell has 8 cells situated within 1-cell distance from it, 24 cells situated within 2-cell distance etc. Those locations that are found within a predefined distance can be then linked and hence considered as *neighbours*. In our case representing locations as *pseudo-documents* implies not only spatial, but also semantic similarity. This fact should not be neglected when localising a tag set: it is easy to expect that linked locations will have high probability to be represented by similar tags and that locations relevant to a classified image will also be close in the graph.

### 6.2.3 Modeling locations

In this section we describe the baseline approach for determining the location of a photo, given a set of tags. By estimating a language model through analysis the terms people use to describe images taken at a particular location, we can predict the most likely position where this photo was taken. In other words, we are interested in obtaining a ranking list of locations $L$, which is ordered by the descending probability that a given tag set $T$ belonging to an image is taken within the bounds of $L$:

$$P(L|T) = \frac{P(T|L)P(L)}{P(T)} \tag{6.7}$$

A location in our framework is represented by a multinomial probability distribution over the vocabulary of tags. Since we do not have any prior information about locations and tags that would otherwise influence the ranking, we consider that $P(L)$ is distributed uniformly and $P(T)$ does not influence the ranking. The locations are then ranked by the probability to generate the tag set of the image. Assuming that each tag $t_i$ in the tag set $T$ is generated independently, the tag set likelihood can be expressed as:

$$P(T|L) = \prod_{i=1}^{|T|} P(t_i|L) \tag{6.8}$$

$$P(t|L) = \frac{|L|}{|L|+\lambda} P(t|L)_{ML} + \frac{\lambda}{|L|+\lambda} P(t|G)_{ML} \tag{6.9}$$

where $P(t|L)_{ML}$ and $P(t|G)_{ML}$ are maximum likelihood estimates of tag generation probabilities for the location and for the global language models, $|L|$ is the size of the location $L$ in tags and $\lambda$ is the parameter of Dirichlet smoothing (Zhai and Lafferty, 2002). Dirichlet smoothing outperformed other kinds of smoothing in our preliminary experiments, seemingly due to its capability to decrease the influence of model lengths on ranking: other

smoothing models imply the preference for smaller models, while in our case we have a large number of "tiny" models (i.e. locations containing only a few images) due to sparseness of our data. This becomes even more critical, when the granularity of the earth grid is small.

The task of image classification using short user-generated metadata (tagsets) have a lot in common with the task of document retrieval if only to consider that users implicitly experience some information need while describing their images with tags. As in document retrieval we try to realize what document the user had in mind when typing a query, in spatial Flickr image classification we should suppose that user always has in mind some location, when typing tags for an image, and present the user with the most relevant one.

### Tag-based smoothing with neighbors

Motivation to smooth from neigbourhoods of locations comes from the need to overcome data sparseness and from understanding that some tags indicate an area that exceeds the bounds of specific location. For example, even if we use very large 100 km cells, some tags specify a country or continent, which may be larger than 100 km. Second, some geographical objects and related tags can be situated in several locations due to the way the grid is placed on the earth surface (for example Rio de Janeiro is situated in 4 neighboring 100 km cells). Smoothing of document class models with models for broader categories is known to be effective for hierarchical document classification (McCallum et al., 1998). However, we do not consider larger cells models for smoothing, since in that case we would ignore the areas lying outside the larger cells bounds (if smaller cells are situated very close to their borders).

The first way to use spatial neighbourhood is to consider that each tag found within a specific location is generated by either the location's language model, or by language models of neighboring locations:

$$P(t|L) = \mu\frac{|L| \cdot P(t|L)_{ML}}{|L| + \lambda} + (1 - \mu)P(t|NB(L)) + \frac{\lambda \cdot P(t|G)_{ML}}{|L| + \lambda} \quad (6.10)$$

$$P(t|NB(L)) = \sum_{L' \in NB(L)} \frac{|L'|}{|L'| + \lambda} \frac{P(t|L')_{ML}}{(2d+1)^2 - 1} \quad (6.11)$$

where $NB(L)$ consists of all locations included into the neighbourhood of location $L$, $d$ is the minimal distance (in grid cells) between locations to be connected in our earth grid graph and $\mu$ is the smoothing coefficient on the probability that the term is generated from the initial location's language model.

### Smoothing cell relevance probabilities

It is reasonable to assume that "good" locations come from "good" neighbourhoods. This means that some relevance should be propagated through the links between close locations. Similar techniques have shown themselves to be effective for the web retrieval (Shakery and Zhai, 2006) or expert finding (see Chapter 4). While, relevance propagation on document and entity graphs is traditionally modeled with random walks (Shakery and Zhai, 2006), we do not expect very distant nodes to have high influence on the relevance of the specific location. For these reasons and also because of computational efficiency requirements, we apply a simple *weighted in-degree* approach: the probability to generate the tag set of a certain location is augmented with the probabilities of neighbouring locations:

$$P(T|L) = \alpha P(T|L) + (1-\alpha) \sum_{L' \in NB(L)} \frac{P(T|L')}{(2d+1)^2 - 1} \qquad (6.12)$$

Note that we are still able to include indirectly adjacent locations by setting parameter $d > 1$:

So far we have regarded our grid graph as undirected, which means that probabilities from all neighboring locations are used in Equation 6.12. It is known from document retrieval (Richardson and Domingos, 2001; Shakery and Zhai, 2006) that it is more efficient to propagate relevance in the hyper-link graph in the direction of more relevant documents. We propose to propagate relevance only from those locations that have lower scores than the location to be smoothed. The motivation is that it is safer to support those documents that have already enough probability to be relevant, than to make highly relevant documents support poor ones. In the case of location retrieval, we may think of the following similar motivation. In the cases when smoothing from nearby locations helps, it succeeds not to select the best location within a certain neighbourhood, which is already efficiently selected by the initial retrieval step, but to select the right "local winner" among those winners from different parts of the globe. Thus, relevance propagation in the direction of "local losers" is not motivated. In graph-related terms, we make our graph query-dependent: edges between cells become directed (from lower scored to higher scored cells) and hence not all of them are used for calculating weighted in-degree.

### Boosting geo-related tags

It is known that users often annotate images with tags that can be recognised as location-specific from a first glance: names of places (e.g. cities or

countries), points-of-interest (e.g. monuments, stadiums, hotels, or bars) or events known to happen in certain locations (e.g. festivals, sport competitions). While the geo-specificity of these tags is captured by our models, it is possible to conclude that some of these tags should be more popular near certain locations even without analysis of their spatial distribution. We introduce preliminary knowledge about tags into our models using a simple boosting approach, similar to the one recently used to boost expansion terms (Cao et al., 2008):

$$P(t|L)_{ML} = P(t|L)_{ML}(1 + \beta P(Loc|t))/Z \tag{6.13}$$

where $P(Loc|t)$ is a probability of the tag $t$ to be location-specific, $\beta$ is a boosting coefficient and $Z$ is a normalisation coefficient.

For the research described in this paper, we use the list of toponyms limited to English names of populated locations, which is taken from the GeoNames database to decide whether the tag is location-specific. For all tags that are in this list the $P(Loc|t)$ equals 1.0 and otherwise equals 0. More sophisticated approaches to tag classification such as Overell et al. (2009) also fit the described method. Although the suggested boosting method depends on external sources, we can assume that most popular gazetteers more or less agree on official names of populated places and such information should be regarded as *common knowledge*. However, gazetteers might disagree significantly in how they define location centroids, bounding boxes and geographic ontologies. Using that information would make our conclusions less generalizable, and so we decided to avoid it.

#### Spatial ambiguity-aware smoothing

It is obvious that some tags are specific for more than one location: either because their scope exceeds the bounds of a single cell, or due to their ambiguity (for example *bath* and *Bath, UK* or because they are instances that are typically spotted at a few specific locations, such as *Elephants* (Weinberger et al., 2008). It is intuitive to trust highly spatially ambiguous tags less than tags that have a single geographical focus. Since we know the co-ordinates of all tag instances in the training data, we are able to characterise spatial ambiguity of a tag by the standard deviation of its latitudes and longitudes $\sigma_{lat}, \sigma_{lon}$. To include this factor into our model we let the smoothing coefficient $\lambda$ in Equation 6.9 be tag-specific and proportional to the ambiguity of a tag:

$$\lambda(t) = \lambda + \gamma(\sigma_{lat}(t) + \sigma_{lon}(t)) \tag{6.14}$$

where $\gamma$ is a weight coefficient to control the influence of ambiguity level on smoothing. s an expected effect, the individually generated probabilities of ambiguous tags will be less decisive for finding the most probable location for a tag set. This is especially important to prevent over-boosting of very ambiguous toponyms at the previous step (e.g. San Francisco refers to 28 populated places in Geonames database).

### 6.2.4 Experimental setup

To prepare a realistic dataset for our studies, we randomly sampled a set of 397000 geo-tagged Flickr images with the associated tags. The only filtering we performed was due to remove the effect of bulk uploads: users often describe a set of images taken at the same time and location with the same tagset (e.g. photos of the same object). To avoid overpresentation of accidental users and tags in the data, we considered only one image with a unique tagset per user per 1km cell. This reduced the set of photos to 140,000, which is still considerably larger than data sets used recently for data mining in Flickr (Rattenbury et al., 2007; Kennedy and Naaman, 2008).

To better understand the data, we geo-referenced all images with the GeoNames gazetteer. Figure 6.6 (left) indicates that a third of photos originate from the U.S., and an equal number of photos relate to Europe. Overall, the collection contains photos from about 180 different countries which makes the data set significantly more representative compared to data used recently, which are focused on the U.S. (Kennedy and Naaman, 2008).

The next step was to separate the data into three parts: data for building location models, data for parameters training and test data. After a series of preliminary experiments we came to the conclusion that the most realistic way to conduct experiments is to separate the data by users. In any other case (e.g. in case of random sampling or separation by image upload date), user-specific tags play a decisive role for finding locations (what would be a re-finding of one of the previous locations of the user). Despite that we expect the results to be better for the users that have their own data in the collection already, it was more important to find out whether we are able to make good predictions for unseen users and their tagsets. Therefore, to separate the data we just sorted the initial image set by unique user ids and considered roughly 120000 ($\approx$85%) images to build models, 10000 ($\approx$7%) to train parameters and 10000 ($\approx$7%) to test our methods.

The main metric that we use for the evaluation and for tuning parameters on training data is location accuracy **Acc**, which simply calculates the percentage of correct predictions over all test examples. However, since we consider *location recommendation* as the primary task to benefit from our lo-

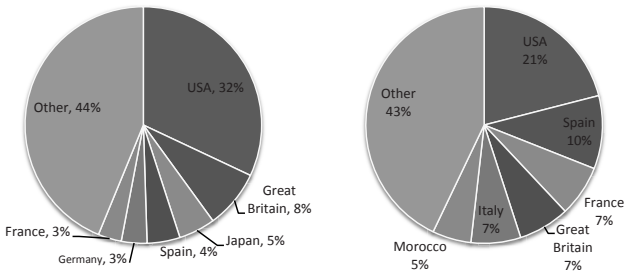



**Figure 6.6**: Images over countries: all images (left) and mapped correctly within 100 km (right)

.

cation prediction techniques, we also analyse additional "relaxed" measures of prediction quality:

- **Reciprocal Rank (MRR)** measures the ability of the system to find the actual location of a photo among its top recommendations.

- **Parent Accuracy (PAcc)** determines whether the predicted location belongs to the same parent with the correct location (for instance, 100 km cells are parents for 50 km cells, 50 km cells - for 10 km cells, etc.).

- **Accuracy within K cells (Acc@K)** computes whether the actual location is within a $K$-cell distance from the predicted location.

It is important to understand that the task of location prediction is nor purely classification task, since for many tagsets (e.g. *southrussia*) it makes no sense to predict the location even at the 100 square kilometers level. However, the ability to properly determine the most probable area where the image (photo) was taken is what users really expect from the system. That is why **PA** and **Acc@K** measures are even more representative than **Acc** measure. Moreover, since we expect the data to be very sparse, it is obvious that we will not have models (containing sufficient number of images) for all possible cells on the earth grid (especially, when we use 1 - 10 kilometers divisions). Prediction of nearby cell then becomes the only way to at least partly fulfill the purpose of the system and this ability is evaluated by **Acc@K** measures.

| **Division** | Acc | MRR | Acc@1 | Acc@2 | Acc@3 | PAcc |
|---|---|---|---|---|---|---|
| 1km | 0.067 | 0.073 | 0.125 | 0.152 | 0.170 | 0.122 |
| 5km | 0.140 | 0.155 | 0.226 | 0.248 | 0.259 | 0.177 |
| 10km | 0.181 | 0.197 | 0.261 | 0.278 | 0.291 | 0.247 |
| 50km | 0.256 | 0.277 | 0.332 | 0.354 | 0.378 | 0.289 |
| 100km | 0.288 | 0.309 | 0.370 | 0.410 | 0.435 | - |

**Table 6.3**: Performance of the baseline LM method

### 6.2.5 Results

We evaluated the following methods and their combinations: baseline language model **LM**, tag-based smoothing **TS**, cell-based smoothing **CS**, cell-based smoothing with score propagation in the direction of higher relevance **CSR**, toponym based boosting **TB** and ambiguity-aware tag specific smoothing **AS**.

All parameters for the tested methods (see Equations 6.9, 6.10, 6.12, 6.13, 6.14) are optimised on the held-out data using line search strategy with maximising accuracy (**Acc** in the table) until no improvement is observed. After optimising $\lambda$ for the baseline retrieval model, the other parameters are optimised independently.

Table 6.3 details the performance of our methods for the different evaluation measures at five different levels of spatial granularity (1, 5, 10, 50, and 100 km). Focusing first on the results for our baseline **LM** method, we observe that the accuracy increases, when increasing the grid size, from 0.067 to 0.288, which is consistent with our expectations. Additional performance improvement is observed when analysing the relaxed accuracy measures to include the direct neighbours of the predicted location. Moreover, as demonstrated by **Acc@2** measure, it is more efficient to locate images within 5km by using 1km cells and 2-cell distance, than by using 5km cells (same observation can be made for 10km and 50km cells).

Further, due to space constraints, we do not show the performance of advanced methods on 5km and 50km grid divisions. However, we testify that results in these cases generally agree with results for 1km, 10km and 100km cells.

Focusing on the effect of the three neighbourhood smoothing extensions **TS**, **CS** and **CSR**, we find some marginal improvements, with the CSR method outperforming the other two smoothing extensions independent of the chosen grid size, as shown in Table 6.4 for the grid size of 1, 10, and 100 km. Smoothing was only done with the immediate neighbours ($d = 1$ in Equations 6.11, 6.12), using larger neighbourhoods did not turn out to be

| **Method** | Acc | MRR | Acc@1 | Acc@2 | Acc@3 | PAcc |
|---|---|---|---|---|---|---|
| | 1 km | | | | | |
| LM | 0.067 | 0.073 | 0.125 | 0.152 | 0.170 | 0.122 |
| +TS | 0.068 | 0.074 | 0.128 | 0.160 | 0.180 | 0.129 |
| +CS | 0.066 | 0.073 | 0.13 | 0.158 | 0.179 | 0.126 |
| +CSR | **0.070** | **0.075** | **0.141** | **0.176** | **0.197** | **0.140** |
| | 10 km | | | | | |
| LM | 0.181 | 0.197 | 0.261 | 0.278 | 0.291 | 0.247 |
| +TS | 0.181 | 0.197 | 0.260 | 0.278 | 0.291 | 0.245 |
| +CS | 0.183 | 0.195 | 0.266 | 0.285 | 0.297 | 0.252 |
| +CSR | **0.187** | **0.201** | **0.271** | **0.288** | **0.301** | **0.255** |
| | 100 km | | | | | |
| LM | 0.288 | 0.309 | 0.370 | 0.410 | 0.435 | - |
| +TS | 0.290 | 0.311 | 0.371 | 0.409 | 0.437 | - |
| +CS | 0.289 | 0.310 | 0.387 | 0.430 | 0.456 | - |
| +CSR | **0.296** | **0.314** | **0.390** | **0.443** | **0.470** | - |

**Table 6.4**: Performance of neighbourhood smoothing

beneficial.

Finally, we have tested different combinations of **LM**, **TB**, **AS** and **CSR** methods. Table 6.5 shows the baseline and the best performing methods from Table 6.4 for clarity. We first observe that all methods improve over baseline **LM** for all measures. Second, among all improvements applied alone, **TB** method shows the best performance. However, the combinations of two methods produce even better results and the maximum performance is reached by using all three methods together (except for **Acc** and **MRR** measures for 1 km grid division). To sum up, the proposed techniques equally improve the performance of **Acc** and **MRR** measures for all cell sizes and are especially effective for 1 km cells according to the rest of measures (up to 31% improvement). The results for accuracy-based measures were not tested for statistical significance because they have a binary outcome (correct or incorrect). The final improvements for MRR are significant at $p < 0.001$ for 1 km and 10 km cells, and at $p < 0.005$ for 100 km cells for the paired t-test.

To study the potential of the proposed improvements, we also measured highly relaxed accuracies with **K** up to 50. Classification into regions of larger bounds is principally easier and, as shown on Figure 6.7, the performance of the baseline method increases accordingly. Despite that fact, the benefit from the advanced methods stays the same which probably means that they avoid especially coarse errors (from improper predicted continents to countries).

We conducted an error analysis of our best method to know the bound-

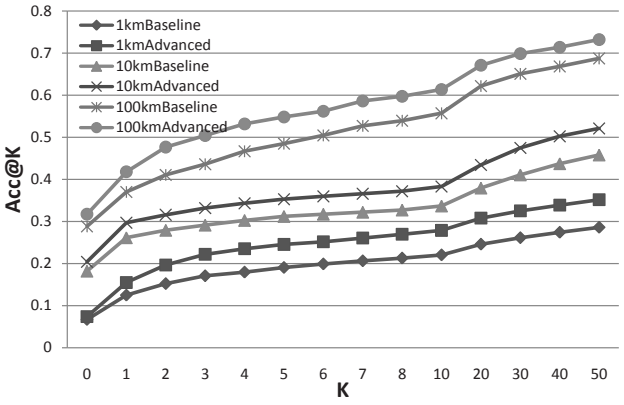



**Figure 6.7**: Accuracies at different distances $K$


**Figure 6.8**: Images with *palma* tag falsely mapped near Palma de Mallorca, Spain

aries of its performance. There are two main sources of errors: caused by sparsity and noisiness of location models and arising from ambiguity and incompleteness of tag sets used for mapping. In the first case, the right location is either not represented in the data (from 70% test cases for 1 km to 7% for 100 km cells), or poorly represented with tags specific for this location only (e.g. containing no toponyms). There are several types of images that are difficult to localize in the second case: (1) images with tags specific to too many locations (e.g. *beach coast rocks lovers*; (2) images with toponyms, but with no tags disambiguating them (e.g. *michigan cats dogs*); (3) images with a tag falsely indicating the reference to a location (e.g. *madrid toronto* taken in Madrid, but mapped to Toronto, or *paris hilton* picturing a poster in New York); (4) images containing a tag specific to a region larger than a chosen grid cell size (e.g. *alaska snow* for 100 km cells or *montmartre paris* for 1 km cells). We suppose that first three types of errors can be eliminated in the future by taking some contextual or user-specific evidence: for instance, tags of recently uploaded images or the location of user IP. Cases like

the one shown in Figure 6.8 may be resolved with additional image content analysis. The latter type of error is more difficult to avoid since some parts of large locations will be always richer in region-specific tags due to being more popular among users. However, the suggestion of the most popular part of it instead of the region as a whole or its centroid is a sensible strategy conceivably correlating with user satisfaction. It is interesting to notice that the dependence of mapping performance on the quality of test tag sets is reflected in distribution of accurately mapped images over countries. As we see on Figure 6.6 (right), photos taken in very popular tourist destinations, such as France, Italy or Morocco, are represented better among correct mappings than in the entire data set, seemingly because tourists almost always describe their photos with location specific tags.

There are several ways in which we would like to extend our mapping approach in the future. First, it is necessary to automatically define an appropriate grid division for a tag set. It is important to minimise interactions between users and the system by showing a map view at the optimal zoom level: probably covering more than one cell, if there are several relevant cells near each other. Second, it seems promising to study the utility of additional evidence coming from a user profile, uploads history, social network or IP address. Finally, images used to build location models can be distinguished by using common (e.g. noise ratio) or Flickr-specific (number of views, interestingness) quality measures.

## 6.3 Summary

In this chapter we described our solutions for two novel tasks resembling expert finding to a considerable degree: entity ranking in Wikipedia and placing Flickr images on a map. While it was possible to apply the state-of-art document retrieval method to the first problem, category expansion and random walk based methods for relevance propagation helped us to improve the baseline performance. As a solution to a completely different entity ranking task, we presented and evaluated generic methods for automatically placing photos uploaded to Flickr on the World map. We showed that we can effectively estimate language models of locations through analysis of the terms people use to describe images taken within their bounds. This left us with an extensible baseline, for which we have shown that we can further increase the accuracy of our predictions by incorporating ambiguity-aware smoothing, cell-based smoothing with score propagation in the direction of highly relevant neighbours, and using an external knowledge base.

| Method | Acc | MRR | Acc@1 | Acc@2 | Acc@3 | PAcc |
|---|---|---|---|---|---|---|
| | | | 1 km | | | |
| LM | 0.067 | 0.073 | 0.125 | 0.152 | 0.170 | 0.122 |
| +CSR | 0.070 | 0.075 | 0.141 | 0.176 | 0.197 | 0.140 |
| +TB | **0.074** (+10%) | **0.078** (+7%) | 0.141 | 0.175 | 0.198 | 0.142 |
| +TB+CSR | 0.074 | 0.078 | 0.147 | 0.188 | 0.212 | 0.148 |
| +AS | 0.061 | 0.068 | 0.132 | 0.167 | 0.186 | 0.124 |
| +AS+TB | 0.070 | 0.074 | 0.143 | 0.183 | 0.207 | 0.139 |
| +AS+CSR | 0.062 | 0.068 | 0.143 | 0.181 | 0.202 | 0.136 |
| +AS+TB+CSR | 0.069 | 0.073 | **0.155** (+24%) | **0.197** (+30%) | **0.222** (+31%) | **0.149** (+22%) |
| | | | 10 km | | | |
| LM | 0.181 | 0.197 | 0.261 | 0.278 | 0.291 | 0.247 |
| +CSR | 0.187 | 0.201 | 0.271 | 0.288 | 0.301 | 0.255 |
| +TB | 0.198 | 0.209 | 0.283 | 0.303 | 0.316 | 0.269 |
| +TB+CSR | 0.198 | 0.210 | 0.286 | 0.305 | 0.319 | 0.269 |
| +AS | 0.190 | 0.205 | 0.275 | 0.292 | 0.306 | 0.260 |
| +AS+TB | 0.204 | 0.213 | 0.295 | 0.314 | 0.329 | 0.279 |
| +AS+CSR | 0.194 | 0.206 | 0.285 | 0.303 | 0.317 | 0.267 |
| +AS+TB+CSR | **0.204** (+13%) | **0.213** (+8%) | **0.297** (+14%) | **0.316** (+14%) | **0.332**(+14%) | **0.280**(+13%) |
| | | | 100 km | | | |
| LM | 0.288 | 0.309 | 0.370 | 0.410 | 0.435 | - |
| +CSR | 0.296 | 0.314 | 0.390 | 0.443 | 0.470 | - |
| +TB | 0.306 | 0.322 | 0.398 | 0.446 | 0.475 | - |
| +TB+CSR | 0.310 | 0.324 | 0.409 | 0.465 | 0.494 | - |
| +AS | 0.302 | 0.321 | 0.386 | 0.427 | 0.452 | - |
| +AS+TB | 0.314 | 0.328 | 0.406 | 0.453 | 0.481 | - |
| +AS+CSR | 0.309 | 0.324 | 0.405 | 0.461 | 0.488 | - |
| +AS+TB+CSR | **0.317** (+10%) | **0.329**(+6%) | **0.418** (+13%) | **0.477** (+16%) | **0.504** (+16%) | - |

**Table 6.5**: Performance of combinations of methods

# 7

# Conclusions

This chapter summarizes our findings by giving an overview of techniques we developed and observations that we made while answering the research questions posed in Section 1.2. We also highlight future research directions and describe potential ways to extend and generalize the results presented in this thesis.

## 7.1 Contributions

This section outlines the lessons that we learned while following our research objectives and answering related research questions. Our final conclusions together with the methods we described in this thesis represent our contributions to research progress in enterprise search, expert finding and related areas.

#### RO1: Going beyond independence of terms and experts

In many cases, the main source of expertise evidence in the enterprise are the documents that contain mentions of employees. Despite that such mentions may potentially refer to different persons, especially if their names are ambiguous, normally, if a full name or an e-mail address are mentioned, the relation of a person to a document can be established with high certainty. Documents thus represent a unique layer that connects topics expressed in words to actual people by using them together in the same context. This observation immediately leads to the assumption that the strength of relation of a person to query terms, measured by the frequency of their co-occurrence, is also a measure of personal expertise with respect to the topic of the query. State-of-the-art expert finding methods armed with this principle provide elegant ways to model the co-occurrence, but regard individual occurrences

of persons and terms in documents as independent events. Although, such default view of text generation allows for acceptable performance, it was still tempting to explore alternatives and answer the following research questions:

> *Does the assumption about dependence of terms and persons in a document lead to better performance of expert finding methods measuring the degree of their co-occurrence? How to model this dependence and estimate its strength? How to use the assumption of dependence to infer personal expertise?*

In order to answer these questions in Chapter 3, we proposed a novel expert finding method. It is based on a model assuming that terms appear in documents after persons are already mentioned, since these persons define the topics of documents and "generate" their terms. The driving motivation behind our approach was to break the assumption of independent generation of persons and terms by document language models. Finally, we described and evaluated three variations of our method. The first variation was based on learning the language model of a person from the data by maximizing its likelihood. We calculated the probabilities that a document relates to a person by giving scores to occurrences of persons specific to the document part where the person is found. In the second variation, we learned these probabilities directly from the data along with personal language models. In the third variation, we additionally utilized a non-uniformly distributed prior probability that a person is an expert. Thus we strengthened the importance of the frequency of its appearance in topical documents to be confident that expertise evidence coming from a personal language model is acquired using sufficient amount of data. Note that we considered that persons generate only relevant documents, since we used only top ranked documents as our data sample for each query. We evaluated all three versions of our approach and observed that the latter two of them outperform the baseline method assuming the independent occurrence of terms and persons given a document. This observation led us to the first important conclusion.

**Conclusion 1.** The assumption of dependence between persons and terms in a document allows to build models that are more effective than those assuming independence. It might be beneficial to consider that the occurrence of terms in a document depends on persons that are mentioned in this document.

While we assumed above that persons are responsible for terms in a document, there are more ways to measure the degree of their relation. For example, we can imagine that their order in the document gives us a clue about non-randomness of their co-occurrence. We explored the utility of this

heuristic by proposing and evaluating an order-based expert finding method. This method considered that a person is related to the (relevant) content of a document to a degree indicated by the position of his/her identifier given positions of query terms. After successful evaluation of the proposed method we arrived at the following conclusion.

**Conclusion 2.** There might be dependency between persons and query terms that is indicated by their order in the document. In this case, the position of a person in respect to positions of query terms should be considered when measuring the strength of the relation between a person and the relevance expressed in the document.

### RO2: Going beyond the scope of directly related documents

Even though it is intuitive to search for expertise evidence in the documents where persons are mentioned, other less obvious sources of evidence should not be necessarily ignored. A person and the set of documents where he/she is mentioned is not a closed world in many cases (see Figure 4.1 in Section 4). If we look at persons and the documents mentioning them as a graph, we may notice that it contains a lot of implicit relations between nodes represented by paths consisting of more than one directed edge. For example, there may be indirect links among persons due to their co-occurrence in the same documents and indirect links among documents when they mention the same person. Direct links among persons due to their professional relations and direct links among documents due to hyperlinks or citations might make such a graph even denser. These observations suggest that remote expertise evidence coming from those documents which are not directly connected to persons may have some potential to improve expert finding performance. To formalize and test our intuition, we answered the following research questions.

> *What sources of expertise evidence in the organization, besides those documents that mention the person, can be used for estimation of his/her expertise? Should we stop after the first step of relevance probability propagation from retrieved documents to directly related candidate experts? How to model the multi-step relevance propagation in a graph of documents and persons?*

As an answer to these questions, we suggested three random walk based models that are able to propagate relevance from documents to candidates even when they are not in direct containment relations. In all three cases we assume that the further evidence is located from a person, the less reliable it is. However, we allow all sources of evidence to add to our trust in a person

as an expert. All proposed methods assume that there is a random surfer (a user) searching for expertise, who always starts the walk from relevant documents, but not necessarily stops after the first step. The smooth and fair propagation of relevance is established by compulsory concentration of the walk around relevant documents, when chances to visit persons depend on how far they are located from the most relevant documents. Their probabilities of being visited at some point in the future are used as measures of their expertise. All three methods differ by how they define that point in the future and how they concentrate the walk around relevant documents. The first method, based on a finite random walk, assumes that a random surfer makes always a predefined number of steps, but stays at documents with a probability proportional to their relevance. The second method, based on an infinite random walk, models a surfer as a tireless seeker, regularly returning to one of the relevant documents with some probability. The third method, using an absorbing random walk, calculates the probability that a surfer visits the candidate in the future (or after a certain number of steps) at least once. We evaluated all three random walk based methods and found each of them more effective than the baseline one-step relevance propagation method using only direct expertise evidence for candidates. These observations let us conclude the following.

**Conclusion 3**. There might be other sources of expertise evidence for a person, besides documents explicitly mentioning that person, namely, implicitly related documents. It is possible to utilize these sources by propagating document relevance in multiple steps in a graph of documents and persons. By using this indirect expertise evidence we are able to improve our chances to predict personal expertise.

Using user models to calculate probabilities that the candidate is an expert can be additionally motivated by the fact that the search for expertise also usually starts from issuing a query to a document search engine, reading one of top ranked documents and finding a person, hopefully an expert, inside. However, users do not always stop after finding that person, since there is no guarantee that they found an expert. Real-world and hence in many cases not fully automated search for expertise is a multi-step process involving reading several documents and meeting several persons by following all kinds of document and social path-ways (Ackerman et al., 2002). That imaginary, yet sensible and realistic activity, is exactly what we attempted to model. The success of our models implies the following.

**Conclusion 4**. If to model expert finding in the enterprise, it is beneficial to assume that the search for expertise is a multi-step process of consulting documents and people.

### RO3: Going beyond the scope of the organization

Previously, we discovered that it is beneficial to utilize additional sources of expertise evidence implicitly related to a candidate. However, the idea to search for indirect evidence never implied that we should limit ourselves only to the sources of such evidence existing within an enterprise. Moreover, since expert finding methods measure the intensity of relevant social activity, it is reasonable to highly reward the activity outside the enterprise due to its effect on the public image of the company. So, as the next step in our research, we tried to find the answers to the following questions.

> *What information sources outside of the organization are useful for finding experts? What measures can be used to get high-quality estimates of expertise from these sources? Is there any benefit in combining evidence acquired from these sources with evidence found in the organization?*

It was promising to look for the additional expertise evidence in publicly available sources on the Web. Using several content-specific APIs of major search engines, we were able to acquire expertise evidence of six types: global, regional and document-specific web search based evidence, as well as news, blog and academic search based evidence. In almost all cases when we used only one type of external evidence, expert finding performed better than when we relied solely on the evidence originating from the enterprise. This observation helped us confirm the following.

**Conclusion 5**. There are sources of information on the Web that can be used to search for expertise evidence. Generic web content, as well as public sources of specific types, like news, blogs or on-line libraries can be used to compare expertise of employees within an organization. Expertise evidence found on the Web may have even better quality than the evidence found within the enterprise.

However, our major goal was more than just testing the predicting power of several types of expertise evidence. We noticed that experts are people who are often not only knowledgeable, but also authoritative and recognizable. Thus, we can expect experts to be frequently mentioned in relevant documents found in more than one type of information source. To validate this principle, we aggregated ranks that have been assigned to the candidates when we used each type of evidence separately. Combining rankings built on different types of evidence increased our performance by far and enabled us to conclude the following.

**Conclusion 6**. People are more likely experts if ranked high according to more than one type of expertise evidence. Thus, it might be important

to aggregate different kinds of global and organizational evidence for the maximum performance of expert finding.

We need to admit that using web evidence was much more beneficial than using sophisticated methods to extract evidence from organizational documents mentioning the person (see Section 3) and other indirect sources of evidence in the enterprise (see Section 4). However, the assumption that "omnipresence" is mandatory for experts should hold only in those companies that encourage their employees for public activity and knowledge sharing (e.g. research institutes). In many other cases, we do not expect web evidence to have such a dramatic influence on expert finding performance. Other proposed methods operating in the scope of a single organization would then become more decisive.

### RO4: Going beyond the scope of the expert finding task

Expert finding belongs to the class of information retrieval tasks whose primary challenge is to estimate relevance of entities with incomplete, non-reliable or missing descriptions. That is the reason why solutions proposed in this thesis should be applicable for a variety of tasks that can be formulated as entity ranking on an abstract level. In order to support the claim that our methods have a scientific and practical value not only for one specific application, we explored related areas and tried to answer the following research questions.

> *Do other applications benefit from the principles used to develop expert finding algorithms? To what degree should solutions be adapted for related tasks?*

The first task we approached was entity ranking in Wikipedia (see Section 6.1). Despite the fact that entities are described by Wikipedia pages, a large part of them have only limited descriptions. Our random walk based approaches made it possible to utilize relevance evidence coming from pages explicitly and implicitly linked to an entity in the Wikipedia hyperlink graph. As our experiments demonstrated, we were able to outperform methods ranking entities only by relevance probabilities of their pages. Another novel application, which we described in this thesis was placing Flickr images on a map using only their user-generated descriptions (see Section 6.2). While, mapping images does not look like an information retrieval task at a first glance, in fact, it can be easily formulated as such. In that case, entities (locations) are described by directly related documents (images found within their bounds) and correct or nearest (relevant) locations should be found for

a tagset (query). Using aggregated indirect relevance evidence coming from spatially nearest neighbors of locations was helpful.

However, we found that it is not always possible to directly transfer solutions from one area to another and immediately reach maximum performance. Expert finding and similar entity ranking tasks are not exceptions in this regard. Entity ranking in Wikipedia heavily depends on category expansion techniques and placing Flickr images on a map benefits from considering toponymy and spatial ambiguity of tags. All these techniques are very task-specific. However, all applications covered in this thesis benefited from the principle of indirect evidence propagation in a graph of entities. To sum up these observations, we formulated the following conclusion.

**Conclusion 7**. Methods and principles developed for the expert finding can be applied to a number of related tasks, particularly to rank entities in Wikipedia and to place objects on a map using their user-generated descriptions. However, there is always the need for a task specific solution if performance is a critical issue.

## 7.2 Directions for Future Work

Various research directions can be followed in the future. Certainly, the person-centric modeling approach from Chapter 3 can be extended up to higher complexity. We can analyze not only top documents, but the entire collection and regard the personal language model as a mixture of sub-persons (clusters) representing different fields of his/her expertise. These inside experts can be used differently across documents and their probability of use may even depend on the set of other persons appearing in the document. A document can be also represented not only as a mixture of persons, but also as a mixture of global latent topics, which in turn appear to be mixtures of persons, accumulating knowledge in the corresponding fields. As an alternative, we can even suppose that terms and persons are independent given such a latent topic, which generates both these kinds of entities. It is also reasonable to find a better use of specific data formats. Particularly, we can consider that persons in the email document appear non-independently: the occurrence of persons in *to* and *cc* fields should depend on the author of the email, who is selecting them for communication. It is also promising to take document links into account: for instance, by regarding emails relating to one thread as a single document. It might be also beneficial to have a unified model combining all features describing co-occurrence into account: the proximity of a candidate to query terms, the order of their occurrence, the frequency of a candidate's occurrence etc. It is also promising to make the

document-candidate weights depend not only on the positions of candidate experts with respect to query terms positions in a document, but also on their absolute positions in the document and on its type.

The random walk based principle of expertise evidence dissemination presented in Chapter 4 is also worthwhile to be researched deeper. For example, a random surfer could move from persons to documents with a probability proportional not only to their association degree, but also to their relevance. Additionally, 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.

It is also promising to explore the usefulness of other possible sources of global expertise evidence (see Section 5.3) and to apply more sophisticated measures of result set quality. First, various normalization and smoothing techniques could be applied to the URL quality measures that we used. Second, we can try to acquire click/visit popularities of pages. Not only major search engines with their huge query logs are able to analyze such statistics. Web sites like `Alexa.com` and `Compete.com` provide a unique opportunity (also through APIs) to inquire about the total number of visits and overall time spent at the domain by web surfers. Other source-specific quality measures should not be overlooked: e.g. blog features (number of subscribers) when using Blog search based evidence or publication features (publisher's authority or citation index) when using academic evidence. However, the most promising direction is applying machine learning mechanisms to find out which quality features of a web result item are the most important and how to combine them into a powerful expertise prediction model. It is also clear that we need a more efficient strategy of evidence acquisition. Sending queries for each person and a query to every web search service is not practical, resource consuming and causes too much latency. The round-robin strategy used in this work may be improved by asking evidence for less promising persons from each next evidence source after rank aggregation at each step. Focused web crawling using all candidates and the organization's names should not be disregarded as a possible option.

A number of related entity ranking tasks are worth to be researched as

well. Problems similar to expert finding arise in the scope of many popular web applications. Possible examples are routing incomplete articles to editors in Wikipedia, or routing questions to potential answerers in question answering portals, like Yahoo! Answers. Expert finding in on-line social networks should be foreseen as a key stage of their evolution. Moreover, proactive recommendation of contacts, recruiters, jobs and companies based on explicit and implicit queries built from user profiles is already in high demand and provided by such services as `Linkedin`.

# Bibliography

*Overview of the INEX 2007 Entity Ranking Track*, Berlin, Heidelberg, 2008. Springer-Verlag.

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

L. A. Adamic, J. Zhang, E. Bakshy, and M. S. Ackerman. Knowledge sharing and Yahoo! answers: everyone knows something. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 665–674, New York, NY, USA, 2008. ACM.

N. Agarwal, H. Liu, L. Tang, and P. S. Yu. Identifying the influential bloggers in a community. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 207–218, New York, NY, USA, 2008. ACM.

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, New York, NY, USA, 2008. ACM.

S. Ahern, M. Naaman, R. Nair, and J. Yang. World Explorer: Visualizing aggregate data from unstructured text in geo-referenced collections. In *JCDL '07*, 2007.

M. Ames and M. Naaman. Why we tag: motivations for annotation in mobile and online media. In *CHI '07*, pages 971–980, New York, NY, USA, 2007. ACM.

E. Amitay, N. Har'El, R. Sivan, and A. Soffer. Web-a-where: geotagging web content. In *SIGIR '04*, pages 273–280, New York, NY, USA, 2004. ACM.

M. Arrington. War of the people search. `www.techcrunch.com/2007/05/09/war-of-the-people-search/`, 2007.

J. A. Aslam and M. Montague. Models for metasearch. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 276–284, New York, NY, USA, 2001. ACM.

L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak. Spatial variation in search engine queries. In *WWW '08*, 2008.

P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. The CSIRO enterprise search test collection. *SIGIR Forum*, 41(2):42–45, 2007a.

P. Bailey, N. Craswell, I. Soboroff, and A. P. de Vries. Overview of the TREC Enterprise 2007 Track. In *The Sixteenth Text Retrieval Conference (TREC 2007) Proceedings*. NIST, November 2007b.

K. Balog and M. de Rijke. Finding similar experts. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 821–822, New York, NY, USA, 2007a. ACM.

K. Balog and M. de Rijke. Non-local evidence for expert finding. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 489–498, New York, NY, USA, 2008. ACM.

K. Balog and M. de Rijke. Combining candidate and document models for expert search. In *Seventeenth Text REtrieval Conference (TREC 2008)*. NIST, NIST, February 2009.

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, New York, NY, USA, 2006. ACM.

K. Balog and M. de Rijke. Determining expert profiles (with an application to expert finding). In *Proceedings IJCAI-2007)*, pages 2657–2662, 2007b.

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, New York, NY, USA, 2006. ACM.

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, New York, NY, USA, 2007. ACM.

K. Balog, M. de Rijke, and W. Weerkamp. Bloggers as experts. In *SIGIR '08*, 2008a.

K. Balog, E. Meij, W. Weerkamp, J. He, and M. de Rijke. The University of Amsterdam at TREC 2008: Blog, Enterprise, and Relevance Feedback. In *TREC 2008 Working Notes*, pages 126–135, November 2008b.

J. Bar-Ilan. Which h-index? - a comparison of WoS, Scopus and Google Scholar. *Scientometrics*, 74(2):257–271, February 2008.

P. Baumard. *Tacit Knowledge in Organizations*. Sage Publications, Inc., Thousand Oaks, CA, USA, 2001.

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.

D. Bennett and D. Taylor. Unethical practices in authorship of scientific papers. *Emergency Medicine*, 15(3):263–270, 2003.

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, New York, NY, USA, 2008. ACM.

M. Buffa. Intranet wikis. In *Proceedings of Intraweb workshop, WWW'06*, 2006.

I. V. Cadez, S. Gaffney, and P. Smyth. A general probabilistic framework for clustering individuals and objects. In *KDD '00: Proceedings of the sixth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 140–149, New York, NY, USA, 2000. ACM.

J. P. Callan, Z. Lu, and W. B. Croft. Searching distributed collections with inference networks. In *Proceedings of ACM SIGIR'95*, pages 12–20, 1995.

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, New York, NY, USA, 2003. ACM.

G. Cao, J.-Y. Nie, J. Gao, and S. Robertson. Selecting good expansion terms for pseudo-relevance feedback. In *SIGIR '08*, pages 243–250, New York, NY, USA, 2008. ACM.

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.

V. R. Carvalho and W. W. Cohen. Ranking users for intelligent message addressing. In Macdonald et al. (2008b), pages 321–333.

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 *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, 2006.

T. Cheng, X. Yan, and K. C.-C. Chang. Entityrank: Searching entities directly and holistically. In *Proceedings of the 33rd International Conference on Very Large Data Bases, University of Vienna, Austria, September 23-27, 2007*, pages 387–398, 2007.

S. Chernov. Task detection for activity-based desktop search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 894–894, New York, NY, USA, 2008. ACM.

S. Chernov, P. Serdyukov, M. Bender, S. Michel, G. Weikum, and C. Zimmer. Database selection and result merging in p2p web search. In G. Moro, S. Bergamaschi, S. Joseph, J.-H. Morin, and A. M. Ouksel, editors, *DBISP2P*, volume 4125 of *Lecture Notes in Computer Science*, pages 26–37. Springer, 2005.

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, New York, NY, USA, 2005. ACM.

N. Craswell. *Methods for Distributed Information Retrieval*. PhD thesis, ANU, 2000. `http://es.csiro.au/pubs/craswell_thesis00.pdf`.

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.

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.

N. Craswell, A. de Vries, and I. Soboroff. Overview of the TREC-2005 Enterprise Track. In *Proceedings of TREC-2005*, Gaithersburg, USA, 2005a.

N. Craswell, S. Robertson, H. Zaragoza, and M. Taylor. Relevance weighting for query independent evidence. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 416–423, New York, NY, USA, 2005b. ACM.

F. Crestani. Application of spreading activation techniques in information retrieval. *Artif. Intell. Rev.*, 11(6):453–482, 1997.

F. Crestani, M. Lalmas, C. J. V. Rijsbergen, and I. Campbell. "Is this document relevant? : Probably": a survey of probabilistic models in information retrieval. *ACM Comput. Surv.*, 30(4):528–552, 1998.

R. Datta, D. Joshi, J. Li, and J. Z. Wang. Image retrieval: Ideas, influences, and trends of the new age. *ACM Comput. Surv.*, 40(2):1–60, 2008.

T. Davenport. Knowledge Management at Microsoft. White paper. January 1997.

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

G. Demartini. Finding Experts Using Wikipedia. In Zhdanova et al. (2007), pages 33–41.

G. Demartini and C. Niederee. Finding experts on the semantic desktop. In *Personal Identification and Collaborations: Knowledge Mediation and Extraction (PICKME08)*, 2008.

A. Dempster, N.M.Laird, and D.B.Rubin. Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39(1):1–38, 1977.

J. Ding, L. Gravano, and N. Shivakumar. Computing geographical scopes of web resources. In *VLDB '00*, pages 545–556, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

B. Dom and D. Paranjpe. A bayesian technique for estimating the credibility of question answerers. In *SDM*, pages 399–409. SIAM, 2008.

D. Dreilinger and A. E. Howe. Experiences with selecting search engines using metasearch. *ACM Trans. Inf. Syst.*, 15(3):195–222, 1997.

H. Duan, Q. Zhou, Z. Lu, O. Jin, S. Bao, Y. Cao, and Y. Yu. Research on Enterprise Track of TREC 2007 at SJTU APEX Lab. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

S. T. Dumais and J. Nielsen. Automating the assignment of submitted manuscripts to reviewers. In *SIGIR '92: Proceedings of the 15th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 233–244, New York, NY, USA, 1992. ACM.

L. Efimova and J. Grudin. Crossing boundaries: A case study of employee blogging. In *HICSS '07: Proceedings of the 40th Annual Hawaii International Conference on System Sciences*, page 86, Washington, DC, USA, 2007. IEEE Computer Society.

J. L. Elsas, J. Arguello, J. Callan, and J. G. Carbonell. Retrieval and feedback models for blog feed search. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 347–354, New York, NY, USA, 2008. ACM.

G. Erkan and D. R. Radev. Lexrank: Graph-based lexical centrality as salience in text summarization. *J. Artif. Intell. Res. (JAIR)*, 22:457–479, 2004.

R. Fagin, R. Kumar, and D. Sivakumar. Comparing top k lists. *SIAM J. Discrete Math.*, 17(1):134–160, 2003.

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

C. Firestone, P. Kelly, and R. Adler. Next-Generation Media: The Global Shift. Report, Aspen Institute, 2007.

J. Fogarty, S. E. Hudson, C. G. Atkeson, D. Avrahami, J. Forlizzi, S. Kiesler, J. C. Lee, and J. Yang. Predicting human interruptibility with sensors. *ACM Trans. Comput.-Hum. Interact.*, 12(1):119–146, 2005.

J. Gao, M. Zhou, J.-Y. Nie, H. He, and W. Chen. Resolving query translation ambiguity using a decaying co-occurrence model and syntactic dependence relations. In *SIGIR '02*, pages 183–190, 2002.

T. Golta. The 2008 recruiting landscape. Five recruiting gurus' 2008 predictions. White paper. January 2008.

L. Gravano. *Querying multiple document collections across the Internet*. PhD thesis, Stanford, CA, USA, 1998.

D. Gruhl, D. N. Meredith, J. H. Pieper, A. Cozzi, and S. Dill. The web beyond popularity: a really simple system for web scale RSS. In *WWW '06: Proceedings of the 15th international conference on World Wide Web*, pages 183–192, New York, NY, USA, 2006. ACM.

GuideWireGroup. Blogging in the enterprise. White paper. January 2005.

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.

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

D. Hawking and S. Robertson. On collection size and retrieval effectiveness. *Inf. Retr.*, 6(1):99–105, 2003.

J. Hays and A. A. Efros. im2gps: estimating geographic information from a single image. In *Proceedings of the IEEE Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2008.

B. He, C. Macdonald, I. Ounis, J. Peng, and R. L. Santos. University of glasgow at trec 2008: Experiments in blog, enterprise, and relevance feedback tracks with terrier. In *Seventeenth Text REtrieval Conference (TREC 2008)*. NIST, NIST, February 2009.

M. A. Hearst. Clustering versus faceted categories for information exploration. *Commun. ACM*, 49(4):59–61, 2006.

M. Hertzum and A. M. Pejtersen. The information-seeking practices of engineers: searching for documents as well as for people. *Inf. Process. Manage.*, 36(5):761–778, 2000.

S. Hettich and M. J. Pazzani. Mining for proposal reviewers: lessons learned at the national science foundation. In *KDD '06: Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 862–871, New York, NY, USA, 2006. ACM.

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

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.

K. Hofmann, K. Balog, T. Bogers, and M. de Rijke. Integrating contextual factors into topic-centric retrieval models for finding similar experts. In *SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval (fCHER)*, Singapore, July 2008.

A. Hogan and A. Harth. The expertfinder corpus 2007 for the benchmarking and development of expertfinding systems. In *Proceedings of the 1st International ExpertFinder Workshop*, Berlin, 1 2007.

J. Huh, L. Jones, T. Erickson, W. A. Kellogg, R. K. E. Bellamy, and J. C. Thomas. Blogcentral: the role of internal blogs at work. In *CHI '07: CHI '07 extended abstracts on Human factors in computing systems*, pages 2447–2452, New York, NY, USA, 2007. ACM.

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

G. Jeh and J. Widom. SimRank: a measure of structural-context similarity. In *KDD '02: Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, New York, NY, USA, 2002. ACM.

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, New York, NY, USA, 2003. ACM.

J. Jiang, W. Lu, and D. Liu. Csir at trec 2007. In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2007.

C. Johansson, P. A. V. Hall, and M. Coquard. ¨talk to paula and peter - they are experienced¨ the experience engine in a nutshell. In *SEKE '99: Proceedings of the 11th International Conference on Software Engineering*

*and Knowledge Engineering, Learning Software Organizations, Methodology and Applications*, pages 171–185, London, UK, 2000. Springer-Verlag.

M. Jones, A. Schuckman, and K. Watson. *The Ethics of Pre-Employment Screening Through the Use of the Internet*, chapter 4. Ethica Publishing, 2007.

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, New York, NY, USA, 2007. ACM.

M. Karimzadehgan, C. Zhai, and G. Belford. Multi-aspect expertise matching for review assignment. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1113–1122, New York, NY, USA, 2008. ACM.

L. Kennedy and M. Naaman. Generating diverse and representative image search results for landmarks. In *WWW '08*, 2008.

R. King. Social networks: Execs use them too. *BusinessWeek*, September 2006.

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

F. Knabe and D. Tunkelang. Enterprise information access and the user experience. *IT Professional*, 9(1):21–28, 2007.

E. Kolek and D. Saunders. Online disclosure: An empirical examination of undergraduate Facebook profiles. *NASPA Journal*, 45(1), 2008.

J. Koren, Y. Zhang, and X. Liu. Personalized interactive faceted search. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 477–486, New York, NY, USA, 2008. ACM.

W. Kraaij, T. Westerveld, and D. Hiemstra. The importance of prior probabilities for entry page search. In *SIGIR '02: Proceedings of the 25th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 27–34, New York, NY, USA, 2002. ACM.

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, New York, NY, USA, 2006. ACM.

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, New York, NY, USA, 2001. ACM.

V. Lavrenko and W. B. Croft. Relevance based language models. In *SIGIR '01: Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 120–127, New York, NY, USA, 2001. ACM Press.

J. Leidner. *Toponym Resolution in Text: Annotation, Evaluation and Applications of Spatial Grounding of Place Names*. PhD thesis, University of Edinburgh, 2007.

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

J. Levine. Business gets social: Corporate usage of Web 2.0 explodes. ChangeWave, January 2008.

D. Liben-Nowell and J. Kleinberg. The link prediction problem for social networks. In *CIKM '03: Proceedings of the twelfth international conference on Information and knowledge management*, pages 556–559, New York, NY, USA, 2003. ACM.

X. Liu, J. Bollen, M. L. Nelson, and H. V. de Sompel. Co-authorship networks in the digital library research community. *Information Processing and Management*, 41(6):1462 – 1480, 2005a. Special Issue on Infometrics.

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, New York, NY, USA, 2005b. ACM Press.

Y.-T. Liu, T.-Y. Liu, T. Qin, Z.-M. Ma, and H. Li. Supervised rank aggregation. In *WWW '07: Proceedings of the 16th international conference on World Wide Web*, pages 481–490, New York, NY, USA, 2007. ACM.

L. Lloyd, P. Kaulgud, and S. Skiena. Newspapers vs. blogs: Who gets the scoop? In *Proceedings of the AAAI Symp. Computational Approaches to Analysing Weblogs (AAAI-CAAW 2006)*, page 8, Stanford, 2006.

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

C. Macdonald and I. Ounis. Using relevance feedback in expert search. In *ECIR 2007*, pages 431–443, 2007a.

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.

C. Macdonald and I. Ounis. Expertise drift and query expansion in expert search. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 341–350, New York, NY, USA, 2007b. ACM.

C. Macdonald, D. Hannah., and I. Ounis. High quality expertise evidence for expert search. In *Proceedings of 30th European Conference on Information Retrieval (ECIR08)*, 2008a.

C. Macdonald, I. Ounis, V. Plachouras, I. Ruthven, and R. W. White, editors. *Advances in Information Retrieval , 30th European Conference on IR Research, ECIR 2008, Glasgow, UK, March 30-April 3, 2008. Proceedings*, volume 4956 of *Lecture Notes in Computer Science*, 2008b. Springer.

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

A. McCallum, R. Rosenfeld, T. M. Mitchell, and A. Y. Ng. Improving text classification by shrinkage in a hierarchy of classes. In *ICML '98*, pages 359–367, San Francisco, CA, USA, 1998. Morgan Kaufmann Publishers Inc.

F. McCown and M. L. Nelson. Agreeing to disagree: search engines and their public interfaces. In *JCDL '07: Proceedings of the 7th ACM/IEEE joint conference on Digital libraries*, pages 309–318, New York, NY, USA, 2007. ACM.

D. W. McDonald and M. S. Ackerman. Just talk to me: a field study of expertise location. In *CSCW '98: Proceedings of the 1998 ACM conference on Computer supported cooperative work*, pages 315–324, New York, NY, USA, 1998. ACM Press.

Q. Mei, C. Liu, H. Su, and C. Zhai. A probabilistic approach to spatiotemporal theme pattern mining on weblogs. In *WWW '06*, 2006.

D. Metzler and W. B. Croft. Latent concept expansion using markov random fields. In *SIGIR '07*, pages 311–318, 2007.

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

M. Naaman, A. Paepcke, and H. Garcia-Molina. From where to what: metadata sharing for digital photographs with geographic coordinates. *On The Move to Meaningful Internet Systems 2003: CoopIS, DOA, and ODBASE*, pages 196–217, 2003.

M. Naaman, S. Harada, Q. Wang, H. Garcia-Molina, and A. Paepcke. Context data in geo-referenced digital photo collections. In *MULTIMEDIA '04*, pages 196–203, New York, NY, USA, 2004. ACM.

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, New York, NY, USA, 2007. ACM.

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, New York, NY, USA, 2001. ACM.

I. Ounis, C. Macdonald, and I. Soboroff. On the trec blog track. In *Proceedings of the Second International Conference on Weblogs and Social Media (ICWSM 2008)*. AAAI Press, 2008.

S. Overell, B. Sigurbjornsson, and R. van Zwol. Classifying tags using open content resources. In *WSDM '09*, 2009.

L. Owens. The Forrester Wave: Enterprise Search, Q2 2008. Report, Forrester, 2008.

L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford University, 1998.

B. Pang and L. Lee. Opinion mining and sentiment analysis. *Found. Trends Inf. Retr.*, 2(1-2):1–135, 2008.

D. Petkova and W. B. Croft. Hierarchical language models for expert finding in enterprise corpora. In *ICTAI '06: Proceedings of the 18th IEEE International Conference on Tools with Artificial Intelligence*, pages 599–608, 2006.

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, New York, NY, USA, 2007. ACM.

J. Raper, G. Gartner, H. Karimi, and C. Rizos. Applications of location-based services: A selected review. *Journal of Location Based Services*, 1 (2), 2007.

T. Rattenbury, N. Good, and M. Naaman. Towards automatic extraction of event and place semantics from flickr tags. In *SIGIR '07*, 2007.

Recommind. Businesses In Dire Need Of Expertise Location, Recommind Survey Reveals. 2009. URL `http://www.recommind.com/node/533`.

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.

S. E. Robertson, S. Walker, and M. Beaulieu. Experimentation as a way of life: Okapi at TREC. *Information Processing and Management*, 36(1):95 – 108, 2000.

H. Rode. *From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search*. Phd thesis, University of Twente, 2008.

E. Selberg and O. Etzioni. Multi-Service Search and Comparison Using the Meta-Crawler. In *Proceedings of WWW4*, Boston MA, December 1995.

J. Seo and W. B. Croft. Blog site search using resource selection. In *CIKM '08: Proceeding of the 17th ACM conference on Information and knowledge management*, pages 1053–1062, New York, NY, USA, 2008. ACM.

P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In Macdonald et al. (2008b), pages 309–320.

P. Serdyukov and D. Hiemstra. Being Omnipresent to be Almighty: The importance of Global Web evidence for organizational expert finding. In *SIGIR 2008 Workshop on Future Challenges in Expertise Retrieval (fCHER)*, 2008b.

P. Serdyukov, S. Chernov, and W. Nejdl. Enhancing expert search through query modeling. In *Advances in Information Retrieval, 29th European Conference on IR Research, ECIR 2007, Rome, Italy, April 2-5, 2007, Proceedings*, pages 737–740, 2007a.

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, New York, NY, USA, 2007b. ACM.

P. Serdyukov, H. Rode, and D. Hiemstra. University of Twente at the TREC 2007 Enterprise Track: Modeling relevance propagation for the expert search task. In *Proceedings of the 16th Text REtrieval Conference (TREC 2007)*, 2007c.

P. Serdyukov, L. Feng, A. H. van Bunningen, S. Evers, H. van Heerde, P. M. G. Apers, M. M. Fokkinga, and D. Hiemstra. The Right Expert at the Right Time and Place. In T. Yamaguchi, editor, *PAKM*, volume 5345 of *Lecture Notes in Computer Science*, pages 38–49. Springer, 2008a.

P. Serdyukov, H. Rode, and D. Hiemstra. 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*, 2008b.

P. Serdyukov, H. Rode, and D. Hiemstra. 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*, 2008c.

P. Serdyukov, H. Rode, and D. Hiemstra. Modeling multi-step relevance propagation for expert finding. In *CIKM '08*, pages 1133–1142, New York, NY, USA, 2008d. ACM.

P. Serdyukov, R. Aly, and D. Hiemstra. University of Twente at the TREC 2008 Enterprise Track: Using the Global Web as an expertise evidence source. In *Proceedings of the 16th Text REtrieval Conference (TREC 2008)*, 2009a.

P. Serdyukov, V. Murdock, and R. van Zwol. Placing Flickr Photos on a Map. In *SIGIR '09: Proceedings of the 32th annual international ACM SIGIR conference on Research and development in information retrieval*, 2009b.

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, New York, NY, USA, 2006. ACM Press.

D. Shen, T. Walkery, Z. Zhengy, Q. Yangz, and Y. Li. Personal name classification in web queries. In *WSDM '08: Proceedings of the international conference on Web search and web data mining*, pages 149–158, New York, NY, USA, 2008. ACM.

C. Sherman. *Google Power: Unleash the Full Potential of Google. The art of googling people*, chapter 12. Barnes and Noble, 2005.

L. Si and J. Callan. Relevant document distribution estimation method for resource selection. In *SIGIR '03: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 298–305, New York, NY, USA, 2003. ACM.

B. Sigurbjornsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *proceedings of the 17th International World Wide Web Conference (WWW 2008)*, Beijing, China, April 2008.

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.

X. Song, Y. Chi, K. Hino, and B. Tseng. Identifying opinion leaders in the blogosphere. In *CIKM '07: Proceedings of the sixteenth ACM conference on Conference on information and knowledge management*, pages 971–974, New York, NY, USA, 2007. ACM.

L. Streeter and K. Lochbaum. An expert/expert-locating system based on automatic representation of semantic structure. *Artificial Intelligence Applications, 1988., Proceedings of the Fourth Conference on*, pages 345–350, Mar 1988.

M. Thelwall and L. Hasler. Blog search engines. *Online Information Review*, 31(4):467–479, 2007. doi: 10.1108/14684520710780421.

C. Torniai, S. Battle, and S. Cayzer. Sharing, discovering and browsing geotagged pictures on the web. Technical report, Digital Media Systems Laboratory, HP Laboratories Bristol, 2007.

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, New York, NY, USA, 2004. ACM.

K. Toyama, R. Logan, and A. Roseway. Geographic location tags on digital images. In *MULTIMEDIA '03*, pages 156–166, New York, NY, USA, 2003. ACM.

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.

S. Vadrevu, Y. Zhang, B. Tseng, G. Sun, and X. Li. Identifying regional sensitive queries in web search. In *Proceedings of WWW '08*, 2008.

E. M. Voorhees, N. K. Gupta, and B. Johnson-Laird. Learning collection fusion strategies. In *Proceedings of ACM SIGIR'95*, pages 172–179, 1995.

C. Wang, J. Wang, X. Xie, and W.-Y. Ma. Mining geographic knowledge using location aware topic model. In *GIR '07*, 2007.

K. Weinberger, M. Slaney, and R. van Zwol. Resolving tag ambiguity. In *Proceedings of the 16th International ACM Conference on Multimedia (MM 2008)*, Vancouver, Canada, November 2008.

T. Westerveld. Correlating topic rankings and person rankings to find experts. In *Proceedings of the 15th Text REtrieval Conference (TREC 2006)*, November 2006.

R. W. White, M. Richardson, M. Bilenko, and A. P. Heath. Enhancing web search by promoting multiple search engine use. In *SIGIR '08: Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 43–50, New York, NY, USA, 2008. ACM.

F. Wu, R. Hoffmann, and D. S. Weld. Information extraction from Wikipedia: moving down the long tail. In *KDD '08: Proceeding of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 731–739, New York, NY, USA, 2008. ACM.

M. Wu, F. Scholer, M. Shokouhi, S. Puglisi, and H. Ali. RMIT University at the TREC 2007 Enterprise Track. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

Y. Wu and D. W. Oard. Indexing emails and email threads for retrieval. In *SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 665–666, New York, NY, USA, 2005. ACM.

J. Xu and W. B. Croft. Cluster-based language models for distributed retrieval. In *SIGIR '99: Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 254–261, New York, NY, USA, 1999. ACM Press.

Y. Yue, T. Finley, F. Radlinski, and T. Joachims. A support vector method for optimizing average precision. In *SIGIR '07: Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 271–278, New York, NY, USA, 2007. ACM.

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.

T. Zesch and I. Gurevych. Analysis of the Wikipedia Category Graph for NLP Applications. In *Proceedings of the TextGraphs-2 Workshop (NAACL-HLT 2007)*, pages 1–8, 2007.

C. Zhai and J. Lafferty. Two-stage language models for information retrieval. In *SIGIR '02*, pages 49–56, New York, NY, USA, 2002. ACM.

C. Zhai and J. Lafferty. Model-based feedback in the language modeling approach to information retrieval. In *CIKM '01: Proceedings of the tenth international conference on Information and knowledge management*, pages 403–410, 2001.

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.

A. V. Zhdanova, L. J. B. Nixon, M. Mochol, and J. G. Breslin, editors. *Proceedings of the 2nd International ISWC+ASWC Workshop on Finding Experts on the Web with Semantics, Busan, Korea, November 12, 2007*, volume 290 of *CEUR Workshop Proceedings*, 2007. CEUR-WS.org.

J. Zhu, D. Song, and S. Rueger. The Open University at TREC 2007 Enterprise Track. In *Proceedings of the 15th Text REtrieval Conference (TREC 2007)*, 2007.

Z. Zhuang, C. Brunk, and C. L. Giles. Modeling and visualizing geosensitive queries based on user clicks. In *LocWeb '08*, 2008.

C.-N. Ziegler and M. Skubacz. Towards automated reputation and brand monitoring on the web. In *WI '06: Proceedings of the 2006 IEEE/WIC/ACM International Conference on Web Intelligence*, pages 1066–1072, Washington, DC, USA, 2006. IEEE Computer Society.

J. Zobel. Collection selection via lexicon inspection. In P. Bruza, editor, *Proceedings of the Australian Document Computing Symposium*, pages 74–80, Apr. 1997.

W. Zong, D. Wu, A. Sun, E.-P. Lim, and D. H.-L. Goh. On assigning place names to geography related web pages. In *JCDL '05*, pages 354–362, New York, NY, USA, 2005. ACM.

# Abstract

The automatic search for knowledgeable people in the scope of an organization is a key function which makes modern Enterprise search systems commercially successful and socially demanded. A number of effective approaches to expert finding were recently proposed in academic publications. Although, most of them use reasonably defined measures of personal expertise, they often limit themselves to rather unrealistic and sometimes oversimplified principles. In this thesis, we explore several ways to go beyond state-of-the-art assumptions used in research on expert finding and propose several novel solutions for this and related tasks.

First, we describe measures of expertise that do not assume independent occurrence of terms and persons in a document what makes them perform better than the measures based on independence of all entities in a document. One of these measures makes persons central to the process of terms generation in a document. Another one assumes that the position of the person's mention in a document with respect to the positions of query terms indicates the relation of the person to the document's relevant content. Second, we find the ways to use not only direct expertise evidence for a person concentrated within the document space of the person's current employer and only within those organizational documents that mention the person. We successfully utilize the predicting potential of additional indirect expertise evidence publicly available on the Web and in the organizational documents implicitly related to a person. Finally, besides the expert finding methods we proposed, we also demonstrate solutions for the tasks from related domains. In one case, we use several algorithms of multi-step relevance propagation to search for typed entities in Wikipedia. In another case, we suggest generic methods for placing photos uploaded to Flickr on the World map using language models of locations built entirely on the annotations provided by users with a few task specific extensions.

# SIKS Dissertatiereeks

1998-1 Johan van den Akker (CWI)
DEGAS - An Active, Temporal Database of Autonomous Objects
1998-2 Floris Wiesman (UM)
Information Retrieval by Graphically Browsing Meta-Information
1998-3 Ans Steuten (TUD)
A Contribution to the Linguistic Analysis of Business Conversations within the Language/Action Perspective
1998-4 Dennis Breuker (UM)
Memory versus Search in Games
1998-5 E.W.Oskamp (RUL)
Computerondersteuning bij Straftoemeting
1999-1 Mark Sloof (VU)
Physiology of Quality Change Modelling;
Automated modelling of Quality Change of Agricultural Products
1999-2 Rob Potharst (EUR)
Classification using decision trees and neural nets
1999-3 Don Beal (UM)
The Nature of Minimax Search
1999-4 Jacques Penders (UM)
The practical Art of Moving Physical Objects
1999-5 Aldo de Moor (KUB)
Empowering Communities: A Method for the Legitimate User-Driven Specification of Network Information Systems
1999-6 Niek J.E. Wijngaards (VU)
Re-design of compositional systems
1999-7 David Spelt (UT)
Verification support for object database design
1999-8 Jacques H.J. Lenting (UM)
Informed Gambling: Conception and Analysis of a Multi-Agent Mechanism for Discrete Reallocation.
2000-1 Frank Niessink (VU)
Perspectives on Improving Software Maintenance
2000-2 Koen Holtman (TUE)
Prototyping of CMS Storage Management
2000-3 Carolien M.T. Metselaar (UVA)
Sociaal-organisatorische gevolgen van kennistechnologie;
een procesbenadering en actorperspectief.
2000-4 Geert de Haan (VU)
ETAG, A Formal Model of Competence Knowledge for User Interface Design
2000-5 Ruud van der Pol (UM)
Knowledge-based Query Formulation in Information Retrieval.

2000-6 Rogier van Eijk (UU)
Programming Languages for Agent Communication
2000-7 Niels Peek (UU)
Decision-theoretic Planning of Clinical Patient Management
2000-8 Veerle Coup (EUR)
Sensitivity Analyis of Decision-Theoretic Networks
2000-9 Florian Waas (CWI)
Principles of Probabilistic Query Optimization
2000-10 Niels Nes (CWI)
Image Database Management System Design Considerations,
Algorithms and Architecture
2000-11 Jonas Karlsson (CWI)
Scalable Distributed Data Structures for Database Management
2001-1 Silja Renooij (UU)
Qualitative Approaches to Quantifying Probabilistic Networks
2001-2 Koen Hindriks (UU)
Agent Programming Languages: Programming with Mental Models
2001-3 Maarten van Someren (UvA)
Learning as problem solving
2001-4 Evgueni Smirnov (UM)
Conjunctive and Disjunctive Version Spaces with
Instance-Based Boundary Sets
2001-5 Jacco van Ossenbruggen (VU)
Processing Structured Hypermedia: A Matter of Style
2001-6 Martijn van Welie (VU)
Task-based User Interface Design
2001-7 Bastiaan Schonhage (VU)
Diva: Architectural Perspectives on Information Visualization
2001-8 Pascal van Eck (VU)
A Compositional Semantic Structure for Multi-Agent Systems Dynamics.
2001-9 Pieter Jan 't Hoen (RUL)
Towards Distributed Development of Large Object-Oriented Models,
Views of Packages as Classes
2001-10 Maarten Sierhuis (UvA)
Modeling and Simulating Work Practice
BRAHMS: a multiagent modeling and simulation language
for work practice analysis and design
2001-11 Tom M. van Engers (VUA)
Knowledge Management:
The Role of Mental Models in Business Systems Design
2002-01 Nico Lassing (VU)
Architecture-Level Modifiability Analysis
2002-02 Roelof van Zwol (UT)
Modelling and searching web-based document collections
2002-03 Henk Ernst Blok (UT)
Database Optimization Aspects for Information Retrieval
2002-04 Juan Roberto Castelo Valdueza (UU)
The Discrete Acyclic Digraph Markov Model in Data Mining
2002-05 Radu Serban (VU)
The Private Cyberspace Modeling Electronic Environments
inhabited by Privacy-concerned Agents
2002-06 Laurens Mommers (UL)
Applied legal epistemology;
Building a knowledge-based ontology of the legal domain
2002-07 Peter Boncz (CWI)
Monet: A Next-Generation DBMS Kernel For Query-Intensive Applications
2002-08 Jaap Gordijn (VU)
Value Based Requirements Engineering: Exploring Innovative
E-Commerce Ideas
2002-09 Willem-Jan van den Heuvel(KUB)
Integrating Modern Business Applications with Objectified Legacy Systems
2002-10 Brian Sheppard (UM)

Towards Perfect Play of Scrabble

2002-11 Wouter C.A. Wijngaards (VU)
Agent Based Modelling of Dynamics: Biological and Organisational Applications

2002-12 Albrecht Schmidt (Uva)
Processing XML in Database Systems

2002-13 Hongjing Wu (TUE)
A Reference Architecture for Adaptive Hypermedia Applications

2002-14 Wieke de Vries (UU)
Agent Interaction: Abstract Approaches to Modelling, Programming and Verifying Multi-Agent Systems

2002-15 Rik Eshuis (UT)
Semantics and Verification of UML Activity Diagrams for Workflow Modelling

2002-16 Pieter van Langen (VU)
The Anatomy of Design: Foundations, Models and Applications

2002-17 Stefan Manegold (UVA)
Understanding, Modeling, and Improving Main-Memory Database Performance

2003-01 Heiner Stuckenschmidt (VU)
Ontology-Based Information Sharing in Weakly Structured Environments

2003-02 Jan Broersen (VU)
Modal Action Logics for Reasoning About Reactive Systems

2003-03 Martijn Schuemie (TUD)
Human-Computer Interaction and Presence in Virtual Reality Exposure Therapy

2003-04 Milan Petkovic (UT)
Content-Based Video Retrieval Supported by Database Technology

2003-05 Jos Lehmann (UVA)
Causation in Artificial Intelligence and Law - A modelling approach

2003-06 Boris van Schooten (UT)
Development and specification of virtual environments

2003-07 Machiel Jansen (UvA)
Formal Explorations of Knowledge Intensive Tasks

2003-08 Yongping Ran (UM)
Repair Based Scheduling

2003-09 Rens Kortmann (UM)
The resolution of visually guided behaviour

2003-10 Andreas Lincke (UvT)
Electronic Business Negotiation: Some experimental studies on the interaction between medium, innovation context and culture

2003-11 Simon Keizer (UT)
Reasoning under Uncertainty in Natural Language Dialogue using Bayesian Networks

2003-12 Roeland Ordelman (UT)
Dutch speech recognition in multimedia information retrieval

2003-13 Jeroen Donkers (UM)
Nosce Hostem - Searching with Opponent Models

2003-14 Stijn Hoppenbrouwers (KUN)
Freezing Language: Conceptualisation Processes across ICT-Supported Organisations

2003-15 Mathijs de Weerdt (TUD)
Plan Merging in Multi-Agent Systems

2003-16 Menzo Windhouwer (CWI)
Feature Grammar Systems - Incremental Maintenance of Indexes to Digital Media Warehouses

2003-17 David Jansen (UT)
Extensions of Statecharts with Probability, Time, and Stochastic Timing

2003-18 Levente Kocsis (UM)
Learning Search Decisions

2004-01 Virginia Dignum (UU)
A Model for Organizational Interaction: Based on Agents, Founded in Logic

2004-02 Lai Xu (UvT)
Monitoring Multi-party Contracts for E-business

2004-03 Perry Groot (VU)
A Theoretical and Empirical Analysis of Approximation in Symbolic Problem Solving

2004-04 Chris van Aart (UVA)
Organizational Principles for Multi-Agent Architectures

2004-05 Viara Popova (EUR)
Knowledge discovery and monotonicity
2004-06 Bart-Jan Hommes (TUD)
The Evaluation of Business Process Modeling Techniques
2004-07 Elise Boltjes (UM)
Voorbeeldig onderwijs; voorbeeldgestuurd onderwijs, een opstap naar abstract denken, vooral voor meisjes
2004-08 Joop Verbeek(UM)
Politie en de Nieuwe Internationale Informatiemarkt, Grensregionale politiële gegevensuitwisseling en digitale expertise
2004-09 Martin Caminada (VU)
For the Sake of the Argument; explorations into argument-based reasoning
2004-10 Suzanne Kabel (UVA)
Knowledge-rich indexing of learning-objects
2004-11 Michel Klein (VU)
Change Management for Distributed Ontologies
2004-12 The Duy Bui (UT)
Creating emotions and facial expressions for embodied agents
2004-13 Wojciech Jamroga (UT)
Using Multiple Models of Reality: On Agents who Know how to Play
2004-14 Paul Harrenstein (UU)
Logic in Conflict. Logical Explorations in Strategic Equilibrium
2004-15 Arno Knobbe (UU)
Multi-Relational Data Mining
2004-16 Federico Divina (VU)
Hybrid Genetic Relational Search for Inductive Learning
2004-17 Mark Winands (UM)
Informed Search in Complex Games
2004-18 Vania Bessa Machado (UvA)
Supporting the Construction of Qualitative Knowledge Models
2004-19 Thijs Westerveld (UT)
Using generative probabilistic models for multimedia retrieval
2004-20 Madelon Evers (Nyenrode)
Learning from Design: facilitating multidisciplinary design teams
2005-01 Floor Verdenius (UVA)
Methodological Aspects of Designing Induction-Based Applications
2005-02 Erik van der Werf (UM))
AI techniques for the game of Go
2005-03 Franc Grootjen (RUN)
A Pragmatic Approach to the Conceptualisation of Language
2005-04 Nirvana Meratnia (UT)
Towards Database Support for Moving Object data
2005-05 Gabriel Infante-Lopez (UVA)
Two-Level Probabilistic Grammars for Natural Language Parsing
2005-06 Pieter Spronck (UM)
Adaptive Game AI
2005-07 Flavius Frasincar (TUE)
Hypermedia Presentation Generation for Semantic Web Information Systems
2005-08 Richard Vdovjak (TUE)
A Model-driven Approach for Building Distributed Ontology-based Web Applications
2005-09 Jeen Broekstra (VU)
Storage, Querying and Inferencing for Semantic Web Languages
2005-10 Anders Bouwer (UVA)
Explaining Behaviour: Using Qualitative Simulation in Interactive Learning Environments
2005-11 Elth Ogston (VU)
Agent Based Matchmaking and Clustering - A Decentralized Approach to Search
2005-12 Csaba Boer (EUR)
Distributed Simulation in Industry
2005-13 Fred Hamburg (UL)
Een Computermodel voor het Ondersteunen van Euthanasiebeslissingen
2005-14 Borys Omelayenko (VU)
Web-Service configuration on the Semantic Web; Exploring how semantics meets pragmatics

2005-15 Tibor Bosse (VU)
Analysis of the Dynamics of Cognitive Processes
2005-16 Joris Graaumans (UU)
Usability of XML Query Languages
2005-17 Boris Shishkov (TUD)
Software Specification Based on Re-usable Business Components
2005-18 Danielle Sent (UU)
Test-selection strategies for probabilistic networks
2005-19 Michel van Dartel (UM)
Situated Representation
2005-20 Cristina Coteanu (UL)
Cyber Consumer Law, State of the Art and Perspectives
2005-21 Wijnand Derks (UT)
Improving Concurrency and Recovery in Database Systems by
Exploiting Application Semantics
2006-01 Samuil Angelov (TUE)
Foundations of B2B Electronic Contracting
2006-02 Cristina Chisalita (VU)
Contextual issues in the design and use of information technology in organizations
2006-03 Noor Christoph (UVA)
The role of metacognitive skills in learning to solve problems
2006-04 Marta Sabou (VU)
Building Web Service Ontologies
2006-05 Cees Pierik (UU)
Validation Techniques for Object-Oriented Proof Outlines
2006-06 Ziv Baida (VU)
Software-aided Service Bundling - Intelligent Methods & Tools
for Graphical Service Modeling
2006-07 Marko Smiljanic (UT)
XML schema matching – balancing efficiency and effectiveness by means of clustering
2006-08 Eelco Herder (UT)
Forward, Back and Home Again - Analyzing User Behavior on the Web
2006-09 Mohamed Wahdan (UM)
Automatic Formulation of the Auditor's Opinion
2006-10 Ronny Siebes (VU)
Semantic Routing in Peer-to-Peer Systems
2006-11 Joeri van Ruth (UT)
Flattening Queries over Nested Data Types
2006-12 Bert Bongers (VU)
Interactivation - Towards an e-cology of people, our technological environment, and the arts
2006-13 Henk-Jan Lebbink (UU)
Dialogue and Decision Games for Information Exchanging Agents
2006-14 Johan Hoorn (VU)
Software Requirements: Update, Upgrade, Redesign - towards a Theory of Requirements Change
2006-15 Rainer Malik (UU)
CONAN: Text Mining in the Biomedical Domain
2006-16 Carsten Riggelsen (UU)
Approximation Methods for Efficient Learning of Bayesian Networks
2006-17 Stacey Nagata (UU)
User Assistance for Multitasking with Interruptions on a Mobile Device
2006-18 Valentin Zhizhkun (UVA)
Graph transformation for Natural Language Processing
2006-19 Birna van Riemsdijk (UU)
Cognitive Agent Programming: A Semantic Approach
2006-20 Marina Velikova (UvT)
Monotone models for prediction in data mining
2006-21 Bas van Gils (RUN)
Aptness on the Web
2006-22 Paul de Vrieze (RUN)
Fundaments of Adaptive Personalisation
2006-23 Ion Juvina (UU)
Development of Cognitive Model for Navigating on the Web

2006-24 Laura Hollink (VU)
Semantic Annotation for Retrieval of Visual Resources
2006-25 Madalina Drugan (UU)
Conditional log-likelihood MDL and Evolutionary MCMC
2006-26 Vojkan Mihajlovic (UT)
Score Region Algebra: A Flexible Framework for Structured Information Retrieval
2006-27 Stefano Bocconi (CWI)
Vox Populi: generating video documentaries from semantically annotated media repositories
2006-28 Borkur Sigurbjornsson (UVA)
Focused Information Access using XML Element Retrieval
2007-01 Kees Leune (UvT)
Access Control and Service-Oriented Architectures
2007-02 Wouter Teepe (RUG)
Reconciling Information Exchange and Confidentiality: A Formal Approach
2007-03 Peter Mika (VU)
Social Networks and the Semantic Web
2007-04 Jurriaan van Diggelen (UU)
Achieving Semantic Interoperability in Multi-agent Systems: a dialogue-based approach
2007-05 Bart Schermer (UL)
Software Agents, Surveillance, and the Right to Privacy:
a Legislative Framework for Agent-enabled Surveillance
2007-06 Gilad Mishne (UVA)
Applied Text Analytics for Blogs
2007-07 Natasa Jovanovic' (UT)
To Whom It May Concern - Addressee Identification in Face-to-Face Meetings
2007-08 Mark Hoogendoorn (VU)
Modeling of Change in Multi-Agent Organizations
2007-09 David Mobach (VU)
Agent-Based Mediated Service Negotiation
2007-10 Huib Aldewereld (UU)
Autonomy vs. Conformity: an Institutional Perspective on Norms and Protocols
2007-11 Natalia Stash (TUE)
Incorporating Cognitive/Learning Styles in a
General-Purpose Adaptive Hypermedia System
2007-12 Marcel van Gerven (RUN)
Bayesian Networks for Clinical Decision Support: A Rational Approach to
Dynamic Decision-Making under Uncertainty
2007-13 Rutger Rienks (UT)
Meetings in Smart Environments; Implications of Progressing Technology
2007-14 Niek Bergboer (UM)
Context-Based Image Analysis
2007-15 Joyca Lacroix (UM)
NIM: a Situated Computational Memory Model
2007-16 Davide Grossi (UU)
Designing Invisible Handcuffs. Formal investigations in Institutions
and Organizations for Multi-agent Systems
2007-17 Theodore Charitos (UU)
Reasoning with Dynamic Networks in Practice
2007-18 Bart Orriens (UvT)
On the development an management of adaptive business collaborations
2007-19 David Levy (UM)
Intimate relationships with artificial partners
2007-20 Slinger Jansen (UU)
Customer Configuration Updating in a Software Supply Network
2007-21 Karianne Vermaas (UU)
Fast diffusion and broadening use: A research on residential adoption and usage of
broadband internet in the Netherlands between 2001 and 2005
2007-22 Zlatko Zlatev (UT)
Goal-oriented design of value and process models from patterns
2007-23 Peter Barna (TUE)
Specification of Application Logic in Web Information Systems
2007-24 Georgina Ramirez Camps (CWI)

Structural Features in XML Retrieval
2007-25 Joost Schalken (VU)
Empirical Investigations in Software Process Improvement
2008-01 Katalin Boer-Sorban (EUR)
Agent-Based Simulation of Financial Markets: A modular,continuous-time approach
2008-02 Alexei Sharpanskykh (VU)
On Computer-Aided Methods for Modeling and Analysis of Organizations
2008-03 Vera Hollink (UVA)
Optimizing hierarchical menus: a usage-based approach
2008-04 Ander de Keijzer (UT)
Management of Uncertain Data - towards unattended integration
2008-05 Bela Mutschler (UT)
Modeling and simulating causal dependencies on process-aware information systems
from a cost perspective
2008-06 Arjen Hommersom (RUN)
On the Application of Formal Methods to Clinical Guidelines, an Artificial Intelligence Perspective
2008-07 Peter van Rosmalen (OU)
Supporting the tutor in the design and support of adaptive e-learning
2008-08 Janneke Bolt (UU)
Bayesian Networks: Aspects of Approximate Inference
2008-09 Christof van Nimwegen (UU)
The paradox of the guided user: assistance can be counter-effective
2008-10 Wauter Bosma (UT)
Discourse oriented summarization
2008-11 Vera Kartseva (VU)
Designing Controls for Network Organizations: A Value-Based Approach
2008-12 Jozsef Farkas (RUN)
A Semiotically Oriented Cognitive Model of Knowledge Representation
2008-13 Caterina Carraciolo (UVA)
Topic Driven Access to Scientific Handbooks 2008-14 Arthur van Bunningen (UT)
Context-Aware Querying; Better Answers with Less Effort 2008-15 Martijn van Otterlo (UT)
The Logic of Adaptive Behavior: Knowledge Representation and Algorithms for
the Markov Decision Process Framework in First-Order Domains.
2008-16 Henriette van Vugt (VU)
Embodied agents from a user's perspective
2008-17 Martin Op 't Land (TUD)
Applying Architecture and Ontology to the Splitting and Allying of Enterprises
2008-18 Guido de Croon (UM)
Adaptive Active Vision
2008-19 Henning Rode (UT)
From Document to Entity Retrieval: Improving Precision and Performance of Focused Text Search
2008-20 Rex Arendsen (UVA)
Geen bericht, goed bericht. Een onderzoek naar de effecten van de introductie
van elektronisch berichtenverkeer met de overheid op de administratieve
lasten van bedrijven
2008-21 Krisztian Balog (UVA)
People Search in the Enterprise
2008-22 Henk Koning (UU)
Communication of IT-Architecture
2008-23 Stefan Visscher (UU)
Bayesian network models for the management of ventilator-associated pneumonia
2008-24 Zharko Aleksovski (VU)
Using background knowledge in ontology matching
2008-25 Geert Jonker (UU)
Efficient and Equitable Exchange in Air Traffic Management Plan Repair
using Spender-signed Currency
2008-26 Marijn Huijbregts (UT)
Segmentation, Diarization and Speech Transcription: Surprise Data Unraveled
2008-27 Hubert Vogten (OU)
Design and Implementation Strategies for IMS Learning Design
2008-28 Ildiko Flesch (RUN)
On the Use of Independence Relations in Bayesian Networks

2008-29 Dennis Reidsma (UT)
Annotations and Subjective Machines - Of Annotators, Embodied Agents, Users, and Other Humans
2008-30 Wouter van Atteveldt (VU)
Semantic Network Analysis: Techniques for Extracting, Representing and Querying Media Content
2008-31 Loes Braun (UM)
Pro-Active Medical Information Retrieval
2008-32 Trung H. Bui (UT)
Toward Affective Dialogue Management using Partially Observable Markov Decision Processes
2008-33 Frank Terpstra (UVA)
Scientific Workflow Design; theoretical and practical issues
2008-34 Jeroen de Knijf (UU)
Studies in Frequent Tree Mining
2008-35 Ben Torben Nielsen (UvT)
Dendritic morphologies: function shapes structure
2009-01 Rasa Jurgelenaite (RUN)
Symmetric Causal Independence Models
2009-02 Willem Robert van Hage (VU)
Evaluating Ontology-Alignment Techniques
2009-03 Hans Stol (UvT)
A Framework for Evidence-based Policy Making Using IT
2009-04 Josephine Nabukenya (RUN)
Improving the Quality of Organisational Policy Making using Collaboration Engineering
2009-05 Sietse Overbeek (RUN)
Bridging Supply and Demand for Knowledge Intensive Tasks - Based on Knowledge, Cognition, and Quality
2009-06 Muhammad Subianto (UU)
Understanding Classification
2009-07 Ronald Poppe (UT)
Discriminative Vision-Based Recovery and Recognition of Human Motion
2009-08 Volker Nannen (VU)
Evolutionary Agent-Based Policy Analysis in Dynamic Environments
2009-09 Benjamin Kanagwa (RUN)
Design, Discovery and Construction of Service-oriented Systems
2009-10 Jan Wielemaker (UVA)
Logic programming for knowledge-intensive interactive applications
2009-11 Alexander Boer (UVA)
Legal Theory, Sources of Law & the Semantic Web
2009-12 Peter Massuthe (TUE, Humboldt-Universitaet zu Berlin)
Operating Guidelines for Services
2009-13 Steven de Jong (UM)
Fairness in Multi-Agent Systems
2009-14 Maksym Korotkiy (VU)
From ontology-enabled services to service-enabled ontologies (making ontologies work in e-science with ONTO-SOA)
2009-15 Rinke Hoekstra (UVA)
Ontology Representation - Design Patterns and Ontologies that Make Sense
2009-16 Fritz Reul (UvT)
New Architectures in Computer Chess
2009-17 Laurens van der Maaten (UvT)
Feature Extraction from Visual Data
2009-18 Fabian Groffen (CWI)
Armada, An Evolving Database System
2009-19 Valentin Robu (CWI)
Modeling Preferences, Strategic Reasoning and Collaboration in Agent-Mediated Electronic Markets
2009-20 Bob van der Vecht (UU)
Adjustable Autonomy: Controling Influences on Decision Making
2009-21 Stijn Vanderlooy (UM). Ranking and Reliable Classification
2009-22 Pavel Serdyukov (UT). Search for expertise: going beyond direct evidence