Searching for People in the Personal Work Space

Krisztian Balog Maarten de Rijke
ISLA, University of Amsterdam
Kruislaan 403, 1098 SJ Amsterdam, The Netherlands
kbalog,mdr@science.uva.nl

***Abstract*— We use a personal work space setting to define several search tasks related to people. We propose models for accomplishing those tasks, and present results of recent and ongoing experiments, propose various further tasks and raise a number of questions.**

I. Introduction

In organizational settings, the role of computer-based collaborative systems has grown recently. An organization's internal and external website, e-mail and database records, agendas and address books are all sources of information, to which people are connected in their work space. This *personal work space* covers the electronic data held by the organization. Within this setting it is natural to look not only for documents, but for entities: answers, services, objects, people, ... Our interest is in one type of entity in particular: *people*. There is much interest in "people search," both from a practical point of view and from the research community, as is witnessed by numerous recent publications on finding experts and on the recent introduction of an expert finding task at TREC [14].

Our aim in this paper is to sketch a research agenda for *people search*. We define several tasks within the general area of people search, where we restrict ourselves to work place scenarios. We then argue that probabilistic language models (LM) provide a particularly convenient and natural way of modeling the tasks we consider. Finally, we briefly discuss issues related to evaluation, both in the setting of tasks assessed at platforms such as TREC, and in connection with possible future evaluation efforts.[1]

II. People Search Tasks

The scenario that we assume is that of an organization with textual content in electronic form, heterogeneous document repositories, containing, amongst others a mixture of different document types (intranet, internet, discussion, internal documents, memos, etc.). Below, we argue that for searching in such an environment, people are an important retrieval cue [5]. Personal information reflects the social milieu in which we organize our work. People are a critical organizing element for workplace information. Below, we identify a number of tasks that flesh out this idea in different directions.

A. Expert finding

Some of the most valuable knowledge in an enterprise resides in the minds of its employees. Enterprises must combine digital information with the knowledge and experience of employees. Expert finding addresses the task of finding the right person with the appropriate skills and knowledge: "Who are the experts on topic X?"

Initial approaches to expert finding employed a database housing the skills and knowledge of each individual in the organization [12], [6]. More recently there has been a move to automatically extract such representations from heterogeneous document collections such as those found within a corporate intranet [4]. However, until recently much of the work performed in this area has been performed in industry with only sketchy solutions, and without formal evaluations.

The output of such an expert finding system should not just be a list of people ranked by their expertise to a topic at hand, but it should also include evidence that supports the decision. A simple way of providing that evidence is to make the relevant documents available, documents created or authored by the candidate (she claims herself to be an expert), and those in which the candidate is mentioned (others say that she is an expert).

B. Expert profiling

The next natural task is to turn the expert finding task around: to profile an individual is to produce a record of the types and areas of skills and knowledge of that individual, together with an identification of levels of 'competency' in each: "What does expert Y know?"

The expert profiling task is naturally decomposed into two stages. The first covers the discovery and identification of possible *knowledge areas*. The second step contains the measurement of the person's competency in each of these areas. Users of a real-word application should (i) understand how to interpret the ratings attached, and (ii) be convinced that these values can be trusted. The latter again requires that the output is supported by evidence, and if needed, this evidence is readily available (e.g., the system can show, or redirect to, relevant documents).

When we seek to determine a person's expertise profile, it is natural to supplement it with further personal details (address, telephone, fax number, e-mail address, affiliation, etc.) and biographical information. However, we use the term *expert profiling* strictly to the description of a person's knowledge, and use the terms *contact detail mining* and *biography finding*

[1] The paper covers a long term research agenda, parts of which have been carried out at the time of writing [1], [3], [2], while others are ongoing or on our to-do list.

for these additional tasks, respectively. Some preliminary work on mining contact details of experts can be found in [3], while biography finding is the topic of a number of recent publications in computational linguistics (see e.g., [10]).

C. Relationship finding

The expert finding and profiling tasks look at individuals and their areas expertise. In the *relationship finding task* we look beyond individuals and are interested in the connections between, and spheres of influence of, people within an organization. We formulate the following subtasks within the area of relationship finding:

- **Connection finding.** In our context, connections between members of the organization cover all types of relations where people work together. Practically, this could mean collaboration, co-operation, co-authorship, chief-employee relationship, etc. In short, given an individual X, we want to know "Who is related to X?" and, moreover, we want to know the particular type of relationship.
- **Collaboration finding.** The purpose of this task is to find people that collaborate with each other on a given field, area or topic. This is a special connection finding task, where we restrict our search to a specific topic. This task, then, seeks to provide answers to the following type of question: "Who has worked/has been working together with X, on the topic Y?"
- **Reputation analysis.** We define an individual's reputation as the collection of opinions that are held about him or her: "What do others say about X?" In reputation analysis, the first step is to identify such opinionated text segments. A next step could be to assign weights to these opinions, based on the reputation of the opinion's owner.

Recognizing when support for a relationship is lacking and determining whether the lack is because the relationship does not exist or is being hidden/missed is a major concern. Users of the system's output need sufficient information to establish confidence in any support given.

III. Modeling People Search

Now that we have sketched key people search tasks, we outline proposals for modeling the tasks. Our approach is based on probabilistic language modeling, which provides a natural setting for capturing the multi-faceted nature of the tasks at hand. We devote most attention to the expert finding task; our proposals for the other tasks share many of the same intuitions and modeling decisions.

A. Expert finding

In the expert finding task, we are given a topic q and a list of candidate experts, and the task is to rank the candidates with respect to their expertise on q. We model the task as follows:

> *what is the probability of a candidate ca being an expert given the query topic q?*

That is, we determine the probability $p(ca|q)$, and rank candidates ca according to this probability. The top candidates are deemed the most probable experts for the given query. Instead of computing this probability directly, we apply Bayes' Theorem, and obtain

$$p(ca|q) = \frac{p(q|ca)p(ca)}{p(q)}, \qquad (1)$$

where $p(ca)$ is the probability of a candidate and $p(q)$ is the probability of a query. Thus, the ranking of candidates is proportional to the probability of the query given the candidate $p(q|ca)$.

We assume that the candidate and the query are conditionally independent from each other. The computation of $p(q|ca)$ uses the following process: we first find documents that are relevant to the query and then score over all documents associated with that candidate. That is

$$p(q|ca) = \textstyle\sum_d p(q|d)p(d|ca), \qquad (2)$$

where $p(d|ca)$ expresses the strength of the association between candidate ca and document d. To determine $p(q|d)$, the probability of a query given a document, we use a standard language modeling technique for IR approach [13], [11], and infer a document model θ_d for each document d [2].

To estimate $p(d|ca)$ we have to solve an *information extraction* task, viz. the task of identifying candidates in a corpus of text; this may involve several types of extraction mechanisms, and (when working with structured or semi-strucured documents such as emails) the strength of the asssociations may depend on the field in which the candidate expert occurs—e.g., in email messages an occurrence of a candidate expert's name in the `cc` field may indicate a strong relation between the candidate and the topic of the email message [3].

B. Expert profiling

As mentioned previously, expert profiling involves two steps: (i) identifying knowledge areas, and (ii) measuring a candidate expert's competency in these areas. Problem (i) is a fairly well researched area, known as *topic detection*. Here, we assume that a list of possible knowledge areas is given, and restrict ourselves to (ii), and state the problem as follows:

> *what are the skills of the candidate ca given the knowledge area ka?*

We estimate $score(ca, ka)$, the candidate's competency for each of the knowledge areas ka, and rank areas according to this score. The top ranked areas are regarded as the skills (or profile) of the candidate ca.

We estimate the candidate's skill scores with $p(ka|ca)$, the probability of a knowledge area ka given the candidate ca. In order to calculate $p(ka|ca)$, we assume that the candidate and the topical area ka are conditionally independent from each other. We first find documents that are relevant to the knowledge area and then score over all documents associated with that candidate. Formally:

$$p(ka|ca) = \textstyle\sum_d p(ka|d)p(d|ca). \qquad (3)$$

This is in fact identical to Eq. 2, if we use the knowledge area as a query topic ($q = ka$). This also means that the relation between the expert finding and expert profiling tasks is captured via Bayes' rule (see Eq. 1). Thus, both tasks involve the estimation of $p(q|ca)$ (or $p(ka|ca)$).

Note that $score$ reflects the candidate's absolute knowledge on the given knowledge area, which is now estimated by "reversing" the expert search model.

One potential issue here are people that are often cited or that take part in many discussions—they may end up being "experts" in everything. One way out is to give each candidate N skillpoints to assign across, say, 50 knowledge areas. A candidate could "earn" all points in one area or spread evenly across many. By receiving more points on a particular area a candidate is more likely to be the "top" expert for the area.

C. Relationship finding

How can relationship finding be modeled? Unlike general web search, significant information extraction efforts are a natural part of solutions for people search tasks, ranging from named entity recognition and classification to topic detection and relation extraction. In particular, we propose to model connection finding and collaboration finding as a mixture of a language modeling approach to snippet retrieval and a (generative) noisy channel approach to determining the likelihood that a retrieved snippet "generates" a relationship between two candidates ca_1 and ca_2 [8]; for collaboration finding this likelihood needs to be conditioned on the knowledge area. Finally, this proposal naturally extends to reputation analysis, where the type of relationship sought is that of expressing an opinion about a candidate at hand.

IV. Evaluating People Search

For the evaluation of people search tasks in the work space, we need a document collection with special characteristics. Organizations have a mixture of document types: web pages, email archives, database records, shared agendas, etc. Because of security and privacy issues, creating such a test collection is non-trivial. The W3C corpus is a crawl of the World Wide Web Consortium's web site (w3c.org), created and made available for the TREC 2005 Enterprise Track [14]. It contains six different subcollections: main web site, personal home pages, mailing lists, CVS archive, Wiki, and 'other.' The corpus contains $330,000$ documents (5.7 GB). In addition, a list of 1092 people (W3C members) was made available.

We use this data collection for evaluating our methods. But the W3C corpus provides more than just the data from which to extract evidence for expertise, profiles, and relationships. Specifically, by using the so-called "working groups" of the W3C as the topics for which experts are being sought, and the groups' members as experts, we obtain ground truth with minimal effort.

The TREC Enterprise Track also provided 50 topics and relevance assessments for the expert finding task. Results achieved by our approaches are top 5 results—note that, unlike some of the other top 5 performing approaches, our models are unsupervised; no manual efforts were made to increase the performance. Moreover, unlike some other systems we do not make any assumptions with regard to the data collection and the topics. In particular, we do not resort to a special treatment of some of the documents (such as e.g., discussion lists), and we do not utilize the fact that the test topics were names of W3C working groups [2], [3].

The same method based on working groups can also be used to assess expert profiling, taking the working group titles to be the relevant knowledge areas to which candidate experts can be assigned. Evaluation results are not available yet.

What is needed to evaluate relationship finding? Here we can take cuez from the "other questions" assessed within the TREC QA task, the relationship finding task at TREC, and the question answering using Wikipedia pilot being launched at CLEF in 2006. Briefly, human generated ground truth is required, where text segments are assessed for relevancy, and, if found relevant, evidence in support of a relationship having been found—we are still attempting to determine a viable solution that requires at most a very small amount of human-generated assessment.

V. Conclusions

We outlined a long-term agenda for people search in the work space setting. We defined several tasks (expert finding, expert profiling, and relationship finding), and sketched models for some of them, all based on language modeling techniques. We believe that people search tasks are a fruitful domain for the combination of annotations with unstructured search.

References

[1] L. Azzopardi, K. Balog, and M. de Rijke, "Language modeling approaches for enterprise tasks," in *Proceedings TREC 2005*, 2006.

[2] K. Balog, L. Azzopardi, and M. de Rijke, "Formal models for expert finding in enterprise corpora," in *Proceedings SIGIR '06*, 2006.

[3] K. Balog and M. de Rijke, "Finding experts and their details in e-mail corpora," in *Proceedings WWW 2006*, 2006.

[4] N. Craswell, D. Hawking, A. M. Vercoustre, and P. Wilkins, "P@noptic expert: Searching for experts not just for documents," in *Ausweb*, 2001, uRL: http://es.csiro.au/pubs/craswell_ausweb01.pdf.

[5] E. Cutrell, S. Dumais, and J. Teevan, "Searching to eliminate personal information management," *Communications of the ACM*, vol. 49, no. 1, pp. 58–64, 2006.

[6] T. H. Davenport and L. Prusak, *Working Knowledge: How Organizations Manage What They Know*. Boston, MA: Harvard Business School Press, 1998.

[7] S. Dumais, E. Cutrell, J. Cadiz, G. Jancke, R. Sarin, and D. Robbins, "Stuff I've seen: a system for personal information retrieval and re-use," in *Proceedings SIGIR '03*, 2003, pp. 72–79.

[8] A. Echihabi and D. Marcu, "A noisy-channel approach to question answering," in *Proceedings ACL'03*, 2003.

[9] R. Fagin, R. Kumar, K. McCurley, J. Novak, D. Sivakumar, J. Tomlin, and D. Williamson, "Searching the workplace web," in *Proceedings WWW 2003*, 2003, pp. 366–375.

[10] E. Filatova and J. Prager, "Tell me what you do and I'll tell you what you are: Learning occupation-related activities for biographies," in *Proceedings HLT-NAACL*, 2005, pp. 49–56.

[11] D. Hiemstra, "Using language models for information retrieval," Ph.D. dissertation, University of Twente, 2001.

[12] M. E. Maron, S. Curry, and P. Thompson, "An inductive search system: Theory, design and implementation," *IEEE Transaction on Systems, Man and Cybernetics*, vol. 16, no. 1, pp. 21–28, 1986.

[13] J. M. Ponte and W. B. Croft, "A language modeling approach to information retrieval," in *Proceedings SIGIR '98*, 1998, pp. 275–281.

[14] "Enterprise track," 2005, URL: http://www.ins.cwi.nl/projects/trec-ent/wiki/.

[15] "The W3C test collection," 2005, URL: http://research.microsoft.com/users/nickcr/w3c-summary.html.