Finding Experts on the Semantic Desktop

Gianluca Demartini and Claudia Niederée

L3S Research Center
Leibniz Universität Hannover
Appelstrasse 9a, 30167 Hannover, Germany
{demartini,niederee}@L3S.de

Abstract. Expert retrieval has attracted deep attention because of the huge economical impact it can have on enterprises. The classical dataset on which to perform this task is company intranet (i.e., personal pages, e-mails, documents). We propose a new system for finding experts in the user's desktop content. Looking at private documents and e-mails of the user, the system builds expert profiles for all the people named in the desktop. This allows the search system to focus on the user's topics of interest thus generating satisfactory results on topics well represented on the desktop. We show, with an artificial test collection, how the desktop content is appropriate for finding experts on the topic the user is interested in.

1 Introduction

Finding people who are expert on certain topics is a search task which has been mainly investigated in the enterprise context. Especially in big enterprises, topic areas can range very much also because of diverse and distributed data sources. This peculiarity of enterprise datasets can highly affect the quality of the results of the expert finding task [15, 16].

It is important to provide the enterprise managers with high quality expert recommendation. The managers need to build new project teams and to find people who can solve problems. Therefore, a high-precision tool for finding experts is needed. Moreover, not only managers need to find experts. In a highly collaborative environment where the willingness of sharing and helping other team members is present, all the employees should be able to find out to which colleague to ask for help in solving issues.

If we want to achieve high-quality results while searching for experts, considering the user's desktop content makes the search much more focused on the user's interests also because the desktop dataset will contain much more expertise evidence (on such topics) than the rest of the public enterprise intranet. Classic expert search systems [9, 30, 21, 25, 26, 17] work on the entire enterprise knowledge available. This means that they use shared repository, e-mails history, forums, wikis, databases, personal home pages, and all the data that an enterprise creates and stores. This makes the system to consider a huge variety of topics, for example, from accountability to IT specific issues. Our solution

focuses on using the user's desktop content as expertise evidence allowing the system to focus on the user's topics of interest thus providing high quality results for queries about such topics.

The system we propose is first indexing the desktop content also using metadata annotation that are produced by the Social Semantic Desktop system Nepomuk [19]. Our expert search system creates a vector space that includes the documents and the people that are present in the desktop content. After this step, when the desktop user issues a query of the type "*Find experts on the topic...*"+*keywords* the system shows a ranked list of people that the user can contact for getting help. Preliminary experiments show the high precision of the expert search results on topics which are covered by the desktop content. A limitation of our system is that it can return only people that are present on the user's desktop. Therefore, the performances are poor when the desktop content (i.e., number of items and people) is limited, as for example for new employees, or when the queries are different from the main topics represented in the desktop. The main contributions of the paper are:

- the description of how the Beagle++ system creates metadata regarding documents and people (Section 2.1).
- a new system for finding experts on a semantic desktop (Section 2.2).
- the description of possible test datasets: one composed of fictitious data and one containing real desktop content (Section 3).
- preliminary experimental results showing how a focused dataset leads to high-quality expert search results (Section 4).
- a review of the previous systems and formal models presented in the field of expert search and Personal Information Management (PIM) (Section 5).

2 System Architecture

2.1 Generating Metadata about People

In order to identify possible expert candidates and link them to desktop items, we used extractors from the Beagle++ Dekstop Search Engine[1] [2] [13, 8]. These extractors identify documents and e-mails authors by analysing the structure and the content of each file. For storing the produced metadata (see Figure 1) we employ the RDF repository developed in the Nepomuk project [19] based on Sesame[3] for storing, querying, and reasoning about RDF and RDF Schema, as well as on Lucene[4], which is integrated with the Sesame framework via the LuceneSail [27], for full-text search.

An additional step is the entity linkage applied to the identified candidates. For example, a person in e-mails is described by an e-mail address, whereas in a publication by the author's name. Other causes for the appearance of different

[1] `http://beagle2.kbs.uni-hannover.de`
[2] `http://www.youtube.com/watch?v=Ui4GDkcR7-U`
[3] `http://www.openrdf.org`
[4] `http://lucene.apache.org`

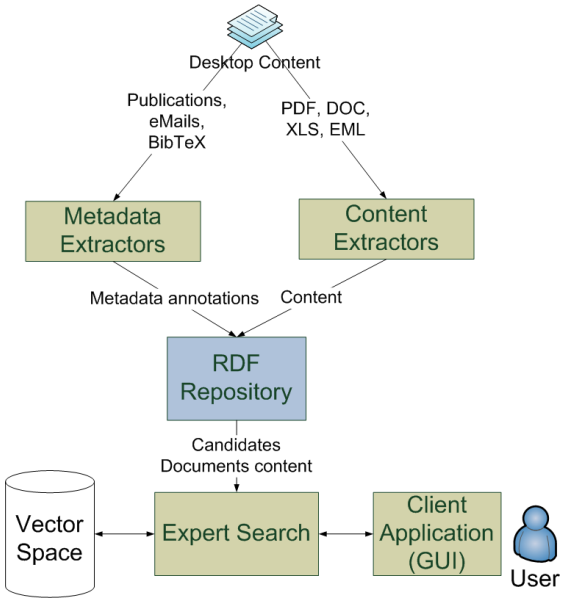

Fig. 1. An overview of how the desktop content is extracted and given in input to the expert search component for indexing. A client application is providing a user interface to the expert search service.

references to the same entity are misspellings, the use of abbreviations, initials, or the actual change of the entity over time (e.g., the e-mail address of a person might change). Again, we exploit a component of the Beagle++ search system for producing information about the linkage.

At this point, we obtained a repository describing desktop items content and metadata. In the next section we explain how we can exploit this data and metadata for finding experts in the semantic desktop content.

2.2 Leveraging Metadata for People Search

In the Nepomuk system, the service of Expert Recommendation[5] aims at providing the user with a list of experts (i.e., people) on a given topic. The experts are selected among a list of persons referral to in the desktop. In order to do so, the component needs to extract, out of the RDF repository, some information about the content of documents and e-mails and also a list of expert candidates (see Figure 1).

Thanks to the Beagle++ system, relations between people and documents are identified and stored in the repository. Entity Linkage identifies references pointing to the same entity by gathering clues as, for example, a person in e-mails described by an e-mail address, whereas in a publication by the author's name. In Beagle++, searching using a person's surname retrieves publications

[5] `http://dev.nepomuk.semanticdesktop.org/wiki/ExpertRecommender`

in which her surname appears as part of an author field as well as e-mails in which her e-mail address appears as part of the sender or receiver fields. This is obtained linking together the objects that refer to the same real world entities [20].

The expert search system we propose can leverage on the extracted relations between documents and people as well as on the linkage information between different representations (e.g., surname and e-mail address). The first step is to create an inverted index for documents: a vector representation of each publication, e-mail, and text-based resources on the desktop is created. Then, for each expert candidate referral to in the desktop, her position into the vector space is computed by linear combination of the resources related with her, using the relation strength as weight. At this point, each candidate expert is placed into the space and a query vector, together with a similarity measure (e.g., cosine similarity), can be used to retrieve a ranked list of experts. The fact that documents are indexed before candidates implies that the dimensions of the vector space are defined by the set of terms present in the desktop collection. This means that the topics of expertise that represents the candidates are those inferred from the documents.

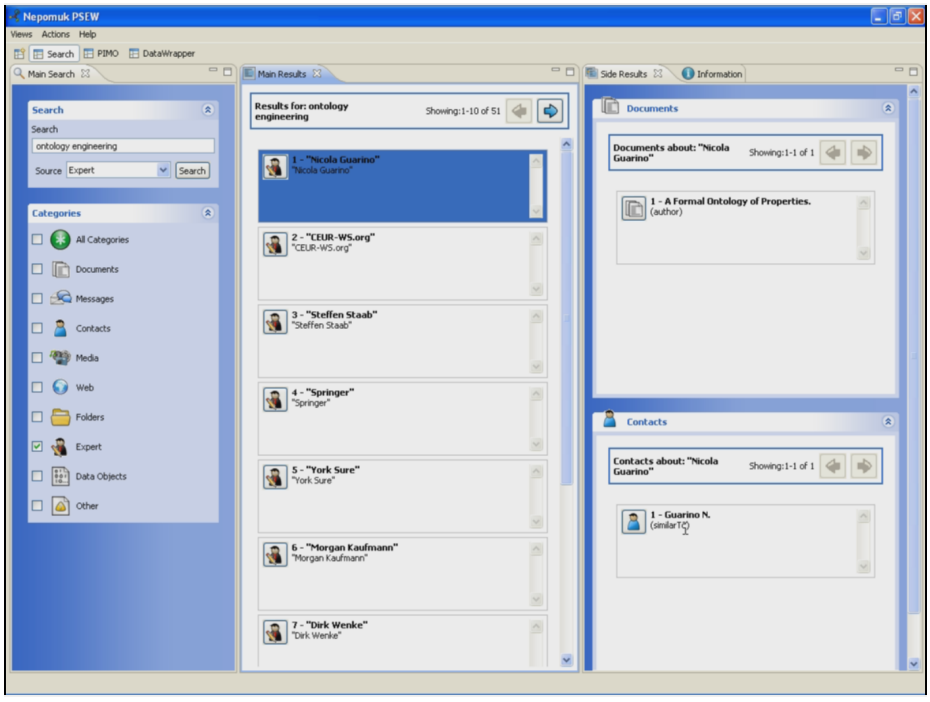

Fig. 2. A client application for searching experts on the semantic desktop.

A client application can then use the Nepomuk Expert Recommendation service (which implements the system described in this paper) by providing a keyword query taken from the user. A screenshot of a possible client application is shown in Figure 2. In the top-left corner the user can provide a keyword query and the choice of looking for experts. In the central panel a ranked list of people is presented as result of the query. In the right pane, resources related to the selected expert are shown.

3 Desktop Search Evaluation Datasets

Evaluation of desktop search algorithms effectiveness is a difficult task because of the lack of standard test collections. The main problem of building such test collection is the privacy concerns that data providers might have while sharing personal data. The privacy issue is major as it impedes the diffusion of personal desktop data among researches. Some solutions for overcoming these problems have been presented in previous work [11, 12].

In this section we describe two possible datasets for evaluating the effectiveness of finding experts using desktop content as evidence of expertise. One is a fictitious desktop dataset representing two hypothetical personas. This dataset has been manually created in the context of the Nepomuk project with the goal of providing a publicly available desktop dataset with no privacy concerns. As at present, the access to the actual data is still restricted. The second one is a set of real desktop data provided by 14 employees of a research center.

3.1 Fictitious Data

In order to obtain reproducible and comparable experimental results there is a need for a common test collection. That is, a set of resources, queries, and relevance assessments that are publicly available. In the case of PIM the privacy issue of sharing personal data has to be faced. For solving this issue the team working on the Nepomuk project has created a collection of desktop items (i.e., documents, e-mails, contacts, calendar items, ...) for some imaginary personas representing hypothetical desktop users. In this paper we describe two desktop collections built in this context.

The first persona is called Claudia Stern[6]. She is a project manager and her interests are mainly about ontologies, knowledge management, and information retrieval. Her desktop contains 56 publications about her interests, 36 e-mails, 19 Word documents about project meetings and deliverables, 12 slides presentations, 17 calendar items, 2 contacts, and an activity log collected while a travel was being arranged (i.e., flight booking, hotel reservation, search for shopping places) containing 122 actions. These resources have been indexed using the Beagle++ system obtaining a total of 22588 RDF triples which have been stored in the RDF repository.

The second persona is called Dirk Hagemann[7]. He works for the project that Claudia manages and his interests are similar to those of Claudia . His desktop

[6] `http://dev.nepomuk.semanticdesktop.org/wiki/Claudia`
[7] `http://dev.nepomuk.semanticdesktop.org/wiki/Dirk`

contains 42 publications, 9 e-mails, 19 Word documents, and 7 text files. These resources have been indexed using the Beagle++ system obtaining a total of 11914 RDF triples which have been stored in the RDF repository.

3.2 Real Data

For evaluating the retrieval effectiveness of a personal information retrieval system, a test collection that accurately represents the desktop characteristics is needed. However, given highly personal data that users usually have on their desktops, currently there are no desktop data collections publicly available. Therefore, we created for experimental purposes our internal desktop data collection. More detail can be found in [11].

The collection that we created - and which are currently using for evaluation experiments - is composed of data gathered from the PCs of 14 different users. The participant pool consists of PhD students, PostDocs and Professors in our research group. The data has been collected from the desktop contents present on the users' PCs in November 2006.

Privacy Preservation In order to face the privacy issues related to providing our personal data to other people, a written agreement has been signed by each of the 14 providers of data, metadata and activities. The document is written with implication that every data contributor is also a possible experimenter. The text is reported in the following:

> **L3S Desktop Data Collection**
> **Privacy Guarantees**
>
> – I will not redistribute the data you provided me to people outside L3S. Anybody from L3S whom I give access to the data will be required to sign this privacy statement.
> – The data you provided me will be automatically processed. I will not look at it manually (e.g. reading the e-mails from a specific person). During the experiment, if I want to look at one specific data item or a group of files/data items, I will ask permission to the owner of the data to look at it. In this context, if I discover possibly sensitive data items, I will remove them from the collection.
> – Permissions of all files and directories will be set such that only the *l3s-experiments-group* and the super-user has access to these files, and that all those will be required to sign this privacy statement.

Currently Available Data The desktop items that we gathered from our 14 colleagues, include e-mails (sent and received), publications (saved from e-mail attachments, saved from the Web, authored / co-authored), address books and calendar appointments. A distribution of the desktop items collected from each user can be seen in Table 1:

User#	E-mails	Publications	Addressbooks	Calendars
1	109	0	1	0
2	12456	0	0	0
3	4532	1054	1	1
4	834	237	0	0
5	3890	261	1	0
6	2013	112	0	0
7	218	28	0	0
8	222	95	1	0
9	0	274	1	1
10	1035	31	1	0
11	1116	157	1	0
12	1767	2799	0	0
13	1168	686	0	0
14	49	452	0	0
Total	29409	6186	7	2
Avg	2101	442	0.5	0.1

Table 1. Resource distribution over the users in the L3S Desktop Data Collection.

A total number of 48,068 desktop items (some of the users provided a dump of their desktop data, including all kinds of documents, not just e-mails, publications, address books or calendars) has been collected, representing 8.1GB of data. On average, each user provided 3,433 items.

In order to emulate a standard test collection, all participants provided a set of queries that reflects typical activities they would perform on their desktop. In addition, each user was asked to contribute their activity logs, related to the period until the point at which the data were provided. All participants defined their *own* queries, related to their activities, and performed search over the reduced images of their desktops, as mentioned above.

4 Preliminary Experiments

We used the Dirk and Claudia datasets (see Section 3.1) in order to perform some initial evaluation of our system for finding experts. We created some queries that match the personas interests imagining which kind of experts they would need to find.

The expert search queries on the Dirk's desktop are:

- ontology engineering
- pagerank
- religion

The expert search queries on the Claudia's desktop are:

- ontology engineering
- ranking in information retrieval
- document search

We issued the same query (i.e., "ontology engineering") on the two datasets in order to compare the results. Dirk and Claudia have the same interest for ontologies but the Dirk desktop contains less data than Claudia's. Table 2 shows the top 5 results on the two datasets. We can see that the results are similar as Dirk and Claudia also share some publications on their desktops. While all the top 5 retrieved people have been working on the topic, the ranking might be improved. For example, the candidate "Dirk Wenke" has less experience than "Nicola Guarino" or "Rudi Studer" on the topic. The explanation of this result is that only local evidence of expertise is used. The quality might be improved by looking at evidence on the web (e.g., DBLP[8] pages).

	Dirk	Claudia
1	Steffen Staab	Steffen Staab
2	York Sure	Riichiro Mizoguchi
3	Rudi Studer	Dirk Wenke
4	Dirk Wenke	York Sure
5	Nicola Guarino	Rudi Studer

Table 2. Top 5 results for the query "ontology engineering".

On Dirk's data we issues the query "pagerank" meaning the famous link based algorithm proposed by Brin and Page in [7]. The top 5 results are presented in Table 3. We can see, again, that all the retrieved candidates have some experience on the topic, but the ordering is not good enough. The authors of the algorithm are placed fourth and fifth while they should be at the top of the list. The first three retrieved candidates have been working on the P2P version of the algorithm.

	Dirk	Claudia	Claudia
	pagerank	ranking in information retrieval	document search
1	Karthikeyan Sankaralingam	Sergey Brin	Jon Kleinberg
2	Simha Sethumadhavan	Karl Aberer	Karl Aberer
3	James C. Browne	Lawrence Page	Eli Upfal
4	Sergey Brin	Jon Kleinberg	Sergey Brin
5	Lawrence Page	Eli Upfal	Monika Henzinger

Table 3. Top 5 results for the query "pagerank" on Dirk's desktop. Top 5 results for the queries "ranking in information retrieval" and "document search" on Claudia's desktop.

The query "religion" on Dirk's desktop, as expected, returned no results. This can be explained because there is no evidence of expertise on such topic in this dataset.

[8] `http://www.informatik.uni-trier.de/~ley/db/`

Finally, we discuss the last two queries on Claudia's dataset. We created queries on very similar topics (i.e., "ranking in information retrieval" and "document search") in order to compare the results. The results are shown in Table 3. We can see that the top 5 results are similar but the ranking. In this case it is hard to say which the best ranking should be as all the retrieved candidates have strong experience on the topic and deciding who is the most expert is highly subjective.

In conclusion, we have seen that the effectiveness of finding experts using the desktop content highly depends on the available resources. If the user queries for experts on topics well represented on her desktop, then the results can be satisfactory. If the query is off-topic then the results can be poor or even be missing. Moreover, further improvements are needed on the ranking function used. A novel measure replacing the cosine similarity used in this experiments might be used.

5 Discussion of Related Work

In this section we describe and discuss the previous work in the field of Expert Search and PIM. We show how existing systems have been designed, which formal models have been proposed, which PIM systems can be extended with expert search functionalities.

5.1 Expert Search Systems

Several expert search systems have been proposed in the last years. These systems use different information sources and features like social network information [9], co-occurrences of terms and changes in the competencies of people over time [30], rule-based models and FOAF[9] data [21]. For the web, a different context from the enterprise search one, one of the approaches proposed [29] focuses on scenarios like Java Online Communities where experts help newcomers or collaborate with each other, and investigated several algorithms that build on answer-reply interaction patterns, using PageRank and HITS authority models as well as additional algorithms exploiting link information in this context. We are not aware of any system for finding experts on the desktop.

The Enterprise PeopleFinder [25, 26] also known as P@noptic Expert [17] first builds a candidate profile attaching all documents related to that candidate in one big document giving different weights to the documents based on their type.

An interesting distinction has been made between *expert finding* and *expert profiling* in [4]. The former approach aims at first retrieving the documents relevant to the query and then extract the experts from them. The latter first builds a profile for each candidate and then matches the query with the profiles without considering the documents anymore [5].

[9] `http://www.foaf-project.org`

5.2 Expert Search Models

All systems mentioned up to now use different ad-hoc techniques but do not formally define retrieval models for experts. Some first steps in this direction have been made: probabilistic models [18] and language models [1–3] have been proposed. Another model for expert search proposed in [23] views this task as a voting problem. The documents associated to a candidate are viewed as votes for this candidate's expertise. In [24] the same authors extended the model including relevance feedback techniques, which is an orthogonal issue. More recently, focus has been put on finding high quality relationships between documents and people and evidence of expertise [28, 22, 6].

5.3 Personal Information Management Systems

A lot of research have been also done in the field of PIM. The most relevant area is the one of desktop search. Finding items on the desktop is not the same task as finding documents on the web. Several commercial systems have been proposed (e.g., Google, Yahoo!, Microsoft). Our expert finding system builds on top of the Beagle++ system: a semantic desktop search engine [8]. Beagle++ exploits the implicit semantic information residing at the desktop level in order to enhance desktop search. Moreover, it creates metadata annotations, thanks to its extractors, that can be reused by our expert finding system.

One important issue in the field of PIM is the evaluation of retrieval effectiveness. Retrieval systems are usually evaluated using standard testbeds (e.g., TREC[10]). In PIM such testbeds are not available mainly because of the privacy issues of sharing personal data. A way to overcome this problem is to create small collections internally to each research group [11].

The Nepomuk project aims at developing a framework for the Social Semantic Desktop. Our expert finding system is integrated in the Nepomuk system providing the user of the semantic desktop this additional search functionality.

If we want to find experts on the desktop, then a crucial task is to extract people names out of full text. Many techniques have been proposed and can be reused for this step. Possible solutions to the problem of measuring similarity between two named entities are presented in [14], how to pre-process a document collection in order to extract names from documents such as e-mail has been proposed in [10].

6 Conclusions and Future Work

In this paper we presented a system for finding experts on the semantic desktop. The approach works as follow. The desktop content is first indexed: metadata is extracted and an RDF repository is built with information about persons and documents. Then, a vector space containing candidate experts and documents is created by exploiting the relations existing between them. Once the documents as well as the candidates are placed into the vector space, a query vector can be placed into the space and a ranked list of experts can be obtained using a

[10] `http://trec.nist.gov`

similarity measure. We used two artificial datasets for performing preliminary experiments. The results show that search results are good for topics that are well represented in the desktop content and poor for others. Effectiveness might be improved by exploiting external evidence of expertise as, for example, web pages. The Beagle++ system indexes visited web pages and, therefore, it could include information from the web also leveraging on semantic technologies such as microforomats or RDFa. Moreover, evidence of expertise contained in both the enterprise intranet and the desktop could be combined in order to generate better results. As future work, we aim at performing a user study using the collection made of data from real user desktops (see Section 3.2) with the goal of evaluating the effectiveness of the expert finding system presented in this paper.

Acknowledgements. We thank the anonymous reviewers for their valuable comments. This work is partially supported by the EU Large-scale Integrating Project OKKAM[11] - Enabling a Web of Entities (contract no. ICT-215032) and by the NEPOMUK[12] project funded by the European Commission under the 6th Framework Programme (IST Contract No. 027705).

References

1. L. Azzopardi, K. Balog, and M. de Rijke. Language modeling approaches for enterprise tasks. *The Fourteenth Text REtrieval Conference (TREC 2005)*, 2006.
2. K. Balog, L. Azzopardi, and M. de Rijke. Formal models for expert finding in enterprise corpora. *Proceedings of the 29th SIGIR conference*, pages 43–50, 2006.
3. K. Balog and M. de Rijke. Finding experts and their Details in e-mail corpora. *Proceedings of the 15th international conference on World Wide Web*, pages 1035–1036, 2006.
4. K. Balog and M. de Rijke. Searching for people in the personal work space. *International Workshop on Intelligent Information Access (IIIA-2006)*, 2006.
5. K. Balog and M. de Rijke. Determining Expert Profiles (With an Application to Expert Finding). *Proceedings of IJCAI-2007*, pages 2657–2662, 2007.
6. K. Balog and M. de Rijke. Associating people and documents. In *ECIR*, pages 296–308, 2008.
7. S. Brin and L. Page. The anatomy of a large-scale hypertextual Web search engine. *Computer Networks and ISDN Systems*, 30(1-7):107–117, 1998.
8. I. Brunkhorst, P. A. Chirita, S. Costache, J. Gaugaz, E. Ioannou, T. Iofciu, E. Minack, W. Nejdl, and R. Paiu. The beagle++ toolbox: Towards an extendable desktop search architecture. *Proceedings of the Semantic Desktop and Social Semantic Collaboration Workshop (SemDesk 2006)*, November 2006.
9. C. Campbell, P. Maglio, A. Cozzi, and B. Dom. Expertise identification using email communications. *Proceedings of the 12th ACM Conference on Information and Knowledge Management (CIKM'03)*, pages 528–531, 2003.
10. V. Carvalho and W. Cohen. Learning to Extract Signature and Reply Lines from Email. *Proceedings of the Conference on Email and Anti-Spam*, 2004.
11. S. Chernov, G. Demartini, E. Herder, M. Kopycki, and W. Nejdl. Evaluating personal information management using an activity logs enriched desktop dataset. In *Proceedings of 3rd Personal Information Management Workshop (PIM 2008)*, 2008.

[11] `http://fp7.okkam.org`
[12] `http://www.nepomuk.semanticdesktop.org`

12. S. Chernov, P. Serdyukov, P.-A. Chirita, G. Demartini, and W. Nejdl. Building a desktop search test-bed. In *ECIR '07: Proceedings of the 29th European Conference on Information Retrieval*, pages 686–690, 2007.
13. P.-A. Chirita, S. Costache, W. Nejdl, and R. Paiu. Beagle^{++}: Semantically enhanced searching and ranking on the desktop. In *ESWC*, pages 348–362, 2006.
14. W. Cohen, P. Ravikumar, and S. Fienberg. A comparison of string distance metrics for name-matching tasks. *Proceedings of the IJCAI-2003 Workshop on Information Integration on the Web (IIWeb-03)*, 2003.
15. N. Craswell, A. de Vries, and I. Soboroff. Overview of the TREC 2005 Enterprise Track.
16. N. Craswell, A. de Vries, and I. Soboroff. Overview of the TREC-2006 Enterprise Track.
17. N. Craswell, D. Hawking, A. Vercoustre, and P. Wilkins. P@noptic Expert: Searching for Experts not just for Documents. *Ausweb*, 2001.
18. H. Fang and C. Zhai. Probabilistic Models for Expert Finding. *Proceedings of 29th European Conference on Information Retrieval (ECIR'07)*, pages 418–430, 2007.
19. T. Groza, S. Handschuh, K. Moller, G. Grimnes, L. Sauermann, E. Minack, M. Jazayeri, C. Mesnage, G. Reif, and R. Gudjonsdottir. The NEPOMUK Project-On the way to the Social Semantic Desktop. *Proceedings of I-Semantics'07*, pages 201–211.
20. E. Ioannou, C. Niederée, and W. Nejdl. Probabilistic entity linkage for heterogeneous information spaces. In *CAiSE*, 2008.
21. J. Li, H. Boley, V. C. Bhavsar, and J. Mei. Expert finding for eCollaboration using FOAF with RuleML rules. *Montreal Conference on eTechnologies (MCTECH)*, 2006.
22. C. Macdonald, D. Hannah, and I. Ounis. High quality expertise evidence for expert search. In *ECIR*, pages 283–295, 2008.
23. C. Macdonald and I. Ounis. Voting for Candidates: Adapting Data Fusion Techniques for an Expert Search Task. *Proceedings of the 15th ACM Conference on Information and Knowledge Management (CIKM'06)*, pages 387–396, 2006.
24. C. Macdonald and I. Ounis. Using Relevance Feedback in Expert Search. *Proceedings of 29th European Conference on Information Retrieval (ECIR'07)*, pages 431–443, 2007.
25. A. McLean, A. Vercoustre, and M. Wu. Enterprise PeopleFinder: Combining Evidence from Web Pages and Corporate Data. *Proceedings of Australian Document Computing Symposium*, 2003.
26. A. McLean, M. Wu, and A. Vercoustre. Combining Structured Corporate Data and Document Content to Improve Expertise Finding. *Arxiv preprint cs/0509005*, 2005.
27. E. Minack, L. Sauermann, G. Grimnes, C. Fluit, and J. Broekstra. The Sesame LuceneSail: RDF Queries with Full-text Search. Technical report, NEPOMUK 2008-1, February 2008.
28. P. Serdyukov and D. Hiemstra. Modeling documents as mixtures of persons for expert finding. In *ECIR*, pages 309–320, 2008.
29. J. Zhang, M. Ackerman, and L. Adamic. Expertise Networks in Online Communities: Structure and Algorithms. *Proceedings of the 16th international conference on World Wide Web*, pages 221–230, 2007.
30. J. Zhu, A. Gonçalves, V. Uren, E. Motta, and R. Pacheco. Mining Web Data for Competency Management. *Web Intelligence '05*, pages 94–100, 2005.