ARTICLE IN PRESS

Information Processing and Management xxx (2011) xxx–xxx

Contents lists available at ScienceDirect

Information Processing and Management

journal homepage: www.elsevier.com/locate/infoproman

Mimicking Web search engines for expert search

Rodrygo L.T. Santos *, Craig Macdonald, Iadh Ounis

School of Computing Science, University of Glasgow, G12 8QQ Glasgow, UK

ARTICLE INFO

Article history:
Received 15 June 2010
Received in revised form 12 November 2010
Accepted 17 November 2010
Available online xxxx

Keywords:
Expert search
Web search engines

ABSTRACT

Many enterprise employees may publish content outside their corporate intranet, making the Web a valuable source for identifying company experts. In this article, we thoroughly investigate the usefulness of Web search engines (WSEs) for expert search. In particular, we claim that the ranking of documentary expertise evidence provided by a WSE should also give an indication of the importance of such evidence. To investigate this, we mimic the rankings of seven different WSEs by trying to reproduce their underlying ranking mechanisms in order to search for candidate experts in the TREC CERC collection. Experimental results show that our approach is effective for expert search, and can significantly improve an intranet-based expert search engine. Moreover, when the mimicking of WSEs is further improved by training, expert search performance is also generally enhanced. Finally, we show that WSEs can be mimicked as effectively using only titles and snippets instead of the full content of WSEs' results, while drastically reducing network costs.

1. Introduction

With the advent of the vast pools of information and documents in large enterprise organisations, collaborative users regularly have the need to find not only documents, but also people with whom they share common interests, or who have specific knowledge in a required area. In an expert search task, the user's goal is to identify people who have relevant expertise to a given topic of interest. This can be accomplished, for instance, by inspecting the organisation's intranet for documents relevant to this topic. Indeed, modern expert search engines usually work as a two-layer retrieval process, with an initial search for relevant expertise evidence for a topic of interest from a document collection, and a subsequent ranking of candidate experts associated to the documentary evidence retrieved in the initial step (Bailey, Craswell, de Vries, & Soboroff, 2007; Balog et al., 2008; Craswell, de Vries, & Soboroff, 2005; Soboroff, de Vries, & Craswell, 2006).

An intranet corpus, however, generally reflects the view of the organisation that it serves—content generation often tends to be autocratic or bureaucratic rather than democratic (Fagin et al., 2003). Such a centralised publishing behaviour may pose additional difficulties for finding experts in a particular subject area among all employees within the organisation (Mukherjee & Mao, 2004). On the other hand, nowadays, many employees may publish content outside the domains of their organisation. For instance, employees may contribute articles to conference proceedings or journals; alternatively, they may participate in public forums or submit posts or comments to blogs. This means that there may be sufficient expertise evidence available on the Web for searching for company experts.

A natural source for looking for experts on the Web is to query existing Web search engines (WSEs) for evidence supporting the expertise of a set of candidates on a particular topic of interest. Unfortunately, however, WSEs are not tailored expert search systems, and hence we cannot measure their performance at an expert search task directly. An alternative would be to use the WSE result listing as the first layer in an expert search system. However, this is difficult because we cannot make

* Corresponding author.
E-mail address: rodrygo@dcs.gla.ac.uk (R.L.T. Santos).

0306-4573/$ - see front matter © 2010 Elsevier Ltd. All rights reserved.
doi:10.1016/j.ipm.2010.11.009

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

WSEs directly target subsets of the Web (e.g., those containing expertise evidence for a set of candidates) through their public interfaces. Moreover, WSEs do not provide relevance scores for their retrieved documents, which are often necessary to generate an effective ranking of candidates (Balog, Azzopardi, & de Rijke, 2006; Macdonald & Ounis, 2006). Therefore, in order to leverage the full potential of WSEs for expert search, we propose to simulate their use as expert search systems themselves.

In this article, our major contributions are two-fold:

- Firstly, we introduce a novel approach for expert search that builds upon the assumed quality of WSEs' ranking mechanisms by explicitly trying to mimic their rankings of suitable documentary expertise evidence.
- Secondly, we conduct a comprehensive experimental investigation on the effectiveness of this approach for expert search in the context of the Expert Search task of the TREC 2007 and 2008 Enterprise tracks.

In contrast with previous approaches, we not only make use of the expertise evidence identified by WSEs, but also of how important this evidence is considered to be based on how it is ranked by individual WSEs. In particular, we experiment by querying seven different WSEs to gather expertise evidence for a set of candidate experts from the TREC CERC collection in the subject areas of each of the TREC 2007 and 2008 Enterprise track topics (Bailey et al., 2007; Balog et al., 2008). This evidence is then indexed locally and scored using four distinct document ranking approaches so as to try to mimic each WSE's underlying scoring mechanism, building what we call a *pseudo*-WSE. By assigning scores to individual documents, we can estimate the extent to which a given documentary evidence is on-topic, and how much it contributes to the relevance of a given candidate. By training the pseudo-WSEs in order to refine the mimicking of the corresponding WSEs, this estimation can be further improved. Finally, this enriched expertise evidence is aggregated into a ranking of candidate experts using an effective expert search model.

Our results using TREC data attest the effectiveness of employing mimicked WSEs' document rankings in replacement to or in combination with existing intranet evidence for expert search. Moreover, by training the pseudo-WSEs to better reproduce the original WSEs' rankings, we find that additional improvements can be attained, often significantly. Finally, by using the titles and snippets extracted from the WSEs' results instead of downloading the corresponding full documents, we show that higher retrieval performances can be achieved while drastically reducing the network overhead incurred by our approach, was it to be deployed in a typical enterprise setting.

The remainder of this article is organised as follows. In Section 2, we describe related work. In Section 3, we show how expertise evidence from WSEs' document rankings can be mined and effectively mimicked by pseudo-WSEs. In Section 4, we describe the experimental setup used in our investigations. In Section 5, we assess the impact of different mimicking strategies for reproducing the document rankings of different WSEs. In Section 6, we evaluate our approach for leveraging expertise evidence from different WSEs for expert search. In Section 7, we evaluate the effectiveness of combining this external evidence with intranet expertise evidence. In Section 8, we investigate the usefulness of titles and snippets as an alternative to downloading the full content of the WSEs' retrieved results. Finally, in Section 9, we present our concluding remarks and directions for future work.

2. Related work

In the following, we overview existing approaches to expert search. Motivated by our proposed approach to leverage Web search engines for expert search, we describe an expert search framework that is agnostic to the underlying document ranking component. We then discuss how external evidence has been previously used in general search tasks and how it can be effectively used within this general framework.

2.1. Expert search in the enterprise

Expert search has emerged as a knowledge management task for finding people with relevant expertise in a given topic (Yimam-seid & Kobsa, 2003). While several enterprise organisations have developed their own expert search systems, the problem has also gathered considerable interest from the information retrieval (IR) research community, with the advent of the expert search task within the TREC 2005 Enterprise track (Craswell et al., 2005). This task ran until TREC 2008, providing a standard testbed for the evaluation of expert search approaches (Bailey et al., 2007; Balog et al., 2008; Craswell et al., 2005; Soboroff et al., 2006).

Modern expert search systems work around the notion of profiles. The profile of a candidate contains documentary evidence representing the expertise of this candidate (e.g., documents, articles, project reports, or e-mails authored by or that contain some sort of identification of the candidate). The profiles of all candidates within an organisation can be used to rank these candidates in response to a query according to their expertise to the topic of the query. To this end, a common approach relies on probabilistic models. For instance, Balog et al. (2006) proposed two generative language models for expert search. While their Model 1 directly builds a language model from a candidate's profile, their Model 2 estimates document language models and their association to a candidate. In the same vein, Fang and Zhai (2007) proposed candidate and topic generation models for expert search. Serdyukov and Hiemstra (2008b) also proposed a generative model to represent the documents retrieved for a query as a mixture of candidate language models. Recent works have also focused on discriminative

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

probabilistic models. In particular, Fang, Si, and Mathur (2010b) proposed a learning framework for expert search and derived effective discriminative models for finding experts.

Different from the aforementioned probabilistic approaches, Macdonald and Ounis (2006) proposed an expert search approach inspired by social choice theory and data fusion techniques. In their Voting Model, documents retrieved for a given query that belong to the profile of a candidate are considered as votes for the relevant expertise of that candidate to the topic of the query. In particular, this approach is not limited to probabilistic document retrieval models. In fact, the Voting Model is completely agnostic to the underlying document ranking component. More importantly for our study, it can operate on top of any document search engine, which makes it particularly suitable for experimenting with rankings produced by WSEs.

The Voting Model defines many voting techniques, which convert a single ranking of documents into a single ranking of candidates. Note that this is different from traditional data fusion techniques, which combine multiple rankings of documents (Macdonald & Ounis, 2006). In this article, we use one of the most effective voting techniques of the Voting Model, namely expCombMNZ (Macdonald & Ounis, 2008), which takes into account the number of voting documents associated to a candidate as well as the scores of these documents, transformed by an exponential function. Applying the exponential function has two effects: it removes the logarithm present in many document weighting models and, in doing so, it places more emphasis on the highly scored documents. It is defined as:

$$score_{expCombMNZ}(C,Q) = |R(Q) \cap P(C)| \times \sum_{d \in R(Q) \cap P(C)} e^{score(d,Q)} \quad (1)$$

where $R(Q)$ corresponds to the set of documents retrieved for the query Q, $P(C)$ corresponds to the profile of candidate C, and $score(d,Q)$ corresponds to the score of document d with respect to the query Q, as given by the underlying document weighting model. Documents in both $R(Q)$ and $P(C)$ are usually taken from an intranet collection. In the next section, however, we describe related works on expert search that rely on expertise evidence gathered from outside the enterprise sphere. Moreover, we discuss how we can improve over these by not only considering this external expertise evidence as such, but also the estimation of its importance as determined by how it is ranked by WSEs.

2.2. External evidence in expert search

External evidence has been traditionally employed in IR for enriching local collections with the primary goal of enhancing the effectiveness of pseudo-relevance feedback techniques (Kwok & Chan, 1998). Collection enrichment is based on the premise that local collections may not contain sufficient relevant information for some queries or may present too noisy term statistics, in which case retrieving documents from a larger, higher-quality external corpus can provide better feedback documents for expanding such difficult queries. The queries expanded from the external pseudo-relevance feedback documents are then used to retrieve documents from the local collection. Collection enrichment has been also shown to be effective in document search within an enterprise setting (Bailey et al., 2007; Balog et al., 2008). Indeed, the sparsity of the vocabulary used in intranets makes collection enrichment a suitable alternative to traditional query expansion techniques (Hawking, 2004).

In the context of expert search within an organisation, exploiting external evidence can be particularly advantageous. Organisations create intranets to facilitate communication and access to information. However, intranet development differs substantially from the Web, which grows democratically. In contrast, an intranet collection is inherently limited in the sense that it is maintained by an assigned number of individuals and aims to reflect the view of a single entity, i.e., the organisation that it serves (Fagin et al., 2003; Mukherjee & Mao, 2004). In such controlled environments, there may be insufficient documentary evidence for ranking relevant candidates for a topic of interest. In this scenario, resorting to external resources can bring useful evidence for finding experts within the enterprise (Hawking, 2004).

In this article, we categorise related approaches on using external evidence for expert search into two main classes: candidate-centric and document-centric. Candidate-centric approaches mine expertise evidence around the name of a candidate, whereas document-centric approaches narrow the expertise mining by targeting occurrences of a candidate's name in the context of a set of topics of interest.

While candidate-centric approaches have the advantage of allowing expertise evidence to be mined off-line, they may not be able to target useful evidence from the Web. Indeed, except for candidates with very distinctive names, these approaches are likely to mine misleading evidence associated to homonyms of the actual candidates. Document-centric approaches, on the other hand, perform expertise mining in an on-demand basis, by targeting the relevant portions of the Web for an unseen expertise query, and incrementally build a local expertise base that can be used to answer future queries.

Two main candidate-centric approaches were proposed recently. Jiang, Han, and Lu (2008) investigated the usefulness of expertise evidence gathered from Google. A breakdown of the results retrieved by this WSE for a list of candidate names extracted from the TREC CERC collection confirmed that much more evidence can be retrieved from the Web than from the intranet collection. Moreover, they showed that this evidence can be effectively used in combination with the intranet evidence in order to retrieve candidate experts. Their approach was later adapted by Balog and de Rijke (2008) with an alternative model for aggregating the expertise evidence derived from result snippets retrieved using the Yahoo! WSE.

A document-centric approach was proposed by Serdyukov and Hiemstra (2008a) and later extended by Serdyukov, Aly, and Hiemstra (2008). In their approach, a query was submitted to six different WSEs and several expertise indicators were derived based on features extracted from the retrieved results. These comprised query-independent features—such as URL depth and domain size—and a simple query-dependent feature, namely, the frequency of the query terms in the URL, title,

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

or snippet of each of the retrieved results. Using any of these features, the ranking produced by the WSE and that produced on the local enterprise collection were then aggregated using the Borda count method from metasearch (Aslam & Montague, 2001). A document-centric approach was also used by He, Macdonald, Ounis, Peng, and Santos (2008) for gathering expertise evidence from the full content of results retrieved using the Yahoo! WSE, however with a limited success.

Table 1 summarises the main characteristics of all these approaches as well as our proposed one. These characteristics include the sources of external evidence, the features considered from each source, and the feature integration strategies deployed by each approach: *query-independent* (based on result counts or other query-independent features extracted from the WSEs' retrieved results), *query-dependent* (based on re-scoring the WSEs' retrieved results using standard retrieval approaches), or *mimicking* (based on fitting the retrieval approaches to replicate the WSEs' underlying ranking mechanism). Additionally, Table 1 organises all approaches as either candidate or document-centric.

As shown in Table 1, our approach can also be classified as document-centric. However, differently from the aforementioned approaches, we wish to fully exploit the capabilities of WSEs for tackling the expert search problem. Indeed, the major contribution of this article is a comprehensive investigation of the several dimensions involved in integrating expertise evidence from WSEs to an expert search system. Our aim is to answer the general question: *how well can WSEs do at expert search?* As WSEs are not tailored expert search systems, however, this question cannot be answered directly—e.g., WSEs do not rank specific portions of the Web containing evidence to support the expertise of a particular set of candidates on a topic of interest, nor they provide relevance scores to enable an improved usage of such evidence. Instead, in order to benefit from the full potential of WSEs for this task, we propose to simulate their use as expert search systems themselves. In particular, besides mining expertise evidence from WSEs, we make use of the valuable information provided by their underlying ranking mechanisms in order to answer a given expertise information need. By appropriately scoring the results retrieved by WSEs, we can generate the relevance scores not provided by them. More importantly, by explicitly trying to mimic their original rankings, we can build an improved expert search system upon their assumed highly effective document rankings.

3. Mimicking Web search engines for expert search

In this section, we describe our approach to expert search, centred on mimicking WSEs' document rankings for a given expertise information need. Following a general expert search flow, as discussed in Section 2.1, we first need to build a set of candidate profiles, each profile comprising documentary evidence of the expertise of a candidate for a given set of topics. An efficient alternative for mining Web information for a given set of topics is to use WSEs' Application Programming Interfaces (APIs) to directly query WSEs. Various WSEs provide programmatic APIs to allow developers to postulate queries and retrieve the associated rankings of URLs which would have been returned by the WSE as for a normal user.

However, if we were to issue an expert search query to a WSE directly, it is likely that no documents related to a particular candidate would be retrieved, primarily due to the large size of the Web. Moreover, the WSEs' APIs do not provide methods to only rank arbitrary subsets of the Web, i.e., those with documents relevant to assessing the expertise of the candidate to the topics of interest. Instead, by formulating appropriate queries, we can use WSEs' APIs for deriving an external profile of the expertise of a given candidate to a predefined set of topics. In this work, we follow the document-centric query formulation strategy suggested by Serdyukov and Hiemstra (2008a) in order to generate what we call *evidence identification queries*. In particular, each evidence identification query contains:

- the candidate's full name in quotes: e.g., *"john smith"*;
- the name of the organisation: e.g., *csiro*;
- the query terms without any quotations: e.g., *genetic modification*;
- a directive prohibiting any results from the actual organisation Web site: e.g., *-site:csiro.au*.

Table 1
Comparison between our approach and the related works of Jiang et al. (2008), Balog and de Rijke (2008), Serdyukov and Hiemstra (2008a), Serdyukov et al. (2008), and He et al. (2008), organised as either candidate or document-centric approaches.

		Candidate		Document			
		Jiang et al. (2008)	Balog and de Rijke (2008)	Serdyukov and Hiemstra (2008a)	Serdyukov et al. (2008)	He et al. (2008)	Ours
Sources	Google	✓					✓
	Yahoo!		✓	✓	✓	✓	✓
	Others			✓			✓
Features	Full-content	✓				✓	✓
	Snippets	✓	✓		✓		✓
	Counts			✓	✓	✓	✓
Integration	Query-independent			✓	✓	✓	✓
	Query-dependent	✓	✓		✓	✓	✓
	Mimicking						✓

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

The use of the name of the organisation helps in name disambiguation by preventing the matching of any content not related to the candidate expert in question (however, this will also prevent the matching of evidence for a candidate from a previous employer). The prohibitive *-site* directive, on the other hand, ensures that the acquired expertise evidence does not overlap with the intranet collection.

Given a WSE, we issue an evidence identification query through its API and obtain a list of search results. For each returned result, we download the corresponding full-text document (as identified by the result URL) and extract its relevant metadata, such as its associated title and descriptive snippet. This content comprises part of the profile of a candidate with respect to a specific topic, as given by the issued evidence identification query. This profile can be enriched on-demand by submitting additional queries involving the candidate's name and different topics.

After building the external candidate profiles, we proceed to ranking the candidates in the organisation. At this stage, our strategy differs from that of related approaches in the literature, as introduced in Section 2.2. In particular, we do not discard the valuable information provided by the WSEs' underlying ranking mechanism. In contrast, we propose an approach more in spirit with the Voting Model (described in Section 2.1), where the external evidence of each candidate's expertise is ranked in response to an expert search query (i.e., the original query without the candidate's name or organisation) in light of the estimated relevance of such evidence, as conveyed by the WSEs' ranking mechanism.

In order to enable our approach, we form *pseudo*-Web search engines (pseudo-WSEs), each of which corresponds to a WSE with indices restricted to the documents contained in the profiles of the candidates as obtained above. To facilitate the creation of the pseudo-WSEs, the documents in the profiles of all candidates are downloaded and indexed locally. Using this index, we can then apply a document retrieval approach in order to mimic the real WSE. In this way, a pseudo-WSE attempts to simulate the ranking mechanism of the corresponding WSE as if the latter was only permitted to retrieve from the documents previously identified in the profiles. Moreover, by using pseudo-WSEs, we can actually assign a score to each of their ranked documents, as WSEs' APIs do not provide this information. The generated ranking of documents can then be used as input to the Voting Model to produce a ranking of candidates. Additionally, as we have control over the document ranking approach applied by each pseudo-WSE, we can explore different ranking strategies in order to better mimic the rankings produced by the corresponding real WSEs.

4. Experimental setup

To ascertain how good WSEs are at expert search, we aim to answer the following research questions:

1. Can we accurately mimic the WSEs' rankings as pseudo-WSEs?
2. Can we effectively use the generated pseudo-WSEs for expert search?
3. Can we effectively combine the evidence mined from WSEs with the existing intranet evidence?
4. Can we leverage useful alternative evidence from WSEs instead of having to download their full results?

Sections 5–8 investigate each of these research questions in turn. In the remainder of this section, we describe the experimental setup aimed at supporting all these investigations.

4.1. Collection and topics

In this work, we experiment with the TREC CSIRO Enterprise Research Collection (CERC) (Bailey et al., 2007; Balog et al., 2008), a corpus of 370,715 documents crawled from the public domain of the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO) in March 2007. This collection was introduced in TREC 2007 as a more realistic representation of real-world enterprise search, in comparison to other existing enterprise test collections. In particular, its topic development phase involved actual members of CSIRO engaged in their work tasks (Bailey et al., 2007). Moreover, CERC comprises a timely corpus with respect to the Web evidence used in our experiments, as described in Section 4.3. In our experiments, we index both the TREC CERC collection and the acquired Web evidence using the Terrier IR platform (Ounis et al., 2006),[1] after applying Porter's stemmer and removing standard English stopwords.

The TREC CERC collection was used in the expert search task of the TREC 2007 and 2008 Enterprise track. During these two editions (Bailey et al., 2007; Balog et al., 2008), a total of 127 test topics was produced. In order to make our results comparable to those reported in the expert search literature, in our experiments, we split these topics into two sets: EX07, comprising the 50 TREC 2007 topics (numbered CE-001–CE-050), and EX08, comprising the 77 TREC 2008 topics (numbered CE-051–CE-127).

4.2. Baseline intranet ranking

Out of the 3,475 candidates experts identified by their corporate e-mail address in the CERC test collection (Macdonald, 2009), we build a set of unique candidates from the top 100 ones suggested by our intranet baseline expert search engine for

[1] http://www.terrier.org.

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

each of the 127 topics from the TREC 2007 and 2008 Enterprise track. This baseline uses the expCombMNZ voting technique (Eq. (1)) on top of a document ranking produced using the Divergence From Randomness DLH13 document weighting model (Amati, 2006; Macdonald, He, Plachouras, & Ounis, 2005). In particular, through an extensive experimentation, this baseline has been shown to deliver an effective expert search performance (Macdonald, 2009).

4.3. Web evidence acquisition

In order to improve our initial intranet baseline, for each candidate it retrieved for each of the EX07 and EX08 topics, we build an evidence identification query, as described in Section 3, including the candidate's name and the title field of the topic. We then submit these queries to major WSEs using their public APIs, which will allow Web documents specific to the query topic and to the candidate to be retrieved. We build 4834 evidence identification queries for the 50 EX07 topics and 7234 of such queries for the 77 EX08 topics. In total, 12,068 evidence identification queries are issued to each of the following WSEs:

1. *Google*: A general WSE, to identify Web documents relating the candidate to the query in question.
2. *Yahoo!*: Another general WSE, to provide comparative results.
3. *Google-pdf*: As Google, but only PDF documents are retrieved.
4. *Yahoo!-pdf*: Analogously, as Yahoo!, but only PDF documents are retrieved.
5. *Google Scholar*: To identify any academic publications by the candidate about the topic area.[2]
6. *Google Blogs*: To identify any blog posts linking the candidate to the query.
7. *Google News*: To identify any news stories linking the candidate to the query. A candidate cited or quoted in a news article is likely to be very authoritative in the topic of the article.

For each WSE, the evidence identification queries are issued and the search result listings obtained. From these, we extract a list of results associated to each candidate. A maximum of 24 results per query are extracted and the corresponding Web pages downloaded.[3,4] These pages form the external profiles of the candidates. Note that these profiles are query-biased, as only documents which are related to the query topic(s) are associated to each candidate.

Table 2 details the statistics of the results found and downloaded from the result lists provided by the WSEs for evidence identification queries based on both EX07 and EX08 topic sets.[5] For each WSE, we note the number of evidence identification queries which retrieved at least one result. As most WSEs use some form of conjunctive querying, in which all query terms must be found in a document for it to be retrieved, not retrieving documents for every evidence identification query is expected—indeed, not every candidate expert considered will have on-topic documents for every evidence identification query. We also report the number of documents retrieved, the number of candidates for whom some evidence was found (out of the 3475 in the CERC test collection), and the average number of documents identified per candidate. For example, in the first row of Table 2, we detail the statistics of the evidence identification queries based on the EX07 topics submitted to the Google WSE: 3464 out of the 4834 queries issued retrieved 1 or more documents; in total, 19,225 documents were retrieved; this provides expertise evidence for 1498 unique candidates from the CERC collection (about 43% of all candidates); this amounts to an average profile size of around 23 documents per candidate. From the table, we note that Google and Yahoo! produce the most evidence, while, as expected, restricting their searches to only PDF documents reduces the number of documents identified. Google Blogs and Google News produce little evidence, while Google Scholar provides evidence for around 50–60% of the candidates covered by Google.

4.4. Mimicking strategies

Besides acquiring expertise evidence from WSEs, we want to ensure that our pseudo-WSEs produce document rankings that are as accurate as possible. For such, we assume that the rankings produced by the real WSEs are of high quality. This is an acceptable assumption, even if purely on the basis that they have many people employed to ensure that their search results are of high quality. Therefore, we want to have each pseudo-WSE produce rankings that are as similar as possible to the real WSE that it is mimicking. However, the ranking strategies adopted by commercial search engines are a closely guarded secret: we cannot know which weighting model they apply, and which additional features are taken into account.

In order to investigate different alternatives for impersonating the WSEs' underlying ranking mechanisms, we apply four standard weighting models: BM25 (Robertson, Walker, Hancock-Beaulieu, Gatford, & Payne, 1995), Hiemstra's language modelling (LM) (Hiemstra, 2001), and two Divergence From Randomness (Amati, 2003) models, namely, PL2 and DLH13. For each of these models, we deploy two mimicking strategies, aimed at reproducing the WSEs' rankings: UNTRAINED and TRAINED.

[2] As Google Scholar does not provide an API, we opted to download and process its result pages for each query directly.
[3] The WSEs were queried and all of their retrieved results downloaded in August, 2008.
[4] As the Google API limits the number of results retrieved per request to 8, in order to reduce the network costs of subsequently downloading the corresponding full documents, we arbitrarily chose to issue a maximum of three requests (i.e., to retrieve up to 3 result pages) per evidence identification query.
[5] We strongly believe in the importance of reproducible experiments. Therefore, we will make all data acquired from the WSEs available.

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

Table 2
Statistics of the indices of external Web content used for expertise evidence.

	Source	# Queries	# Docs	# Cands	Avg. Prof.
2007	Google	3464	19,225	1498	22.86
	Yahoo!	2970	16,738	1391	18.95
	Google-pdf	2969	11,080	1355	18.27
	Yahoo!-pdf	2502	9450	1259	14.69
	Scholar	1576	2376	871	6.95
	Blogs	67	53	50	1.80
	News	32	33	26	1.81
2008	Google	5060	21,456	1551	28.85
	Yahoo!	3969	18,509	1372	23.51
	Google-pdf	4339	11,848	1399	23.20
	Yahoo!-pdf	3263	10,283	1220	18.71
	Scholar	1906	2325	792	8.68
	Blogs	65	56	41	2.51
	News	31	37	18	3.17

The UNTRAINED mimicking strategy applies the considered weighting models with their default parameter settings. In particular, we use the often suggested settings of BM25's $b = 0.75$ (Robertson et al., 1995), LM's $\lambda = 0.15$ (Hiemstra, 2001), and PL2's $c = 1.0$ (Amati, 2006). Note that the DLH13 weighting model has no parameters to train (Amati, 2003). Besides the UNTRAINED mimicking strategy, we investigate another strategy to improve the mimickings of BM25, LM, and PL2. In particular, our TRAINED mimicking strategy optimises the parameter settings of each of these models using the training queries that we have available. As discussed in Section 3, for each WSE, we have a list of the evidence identification queries that it answered and the ranking of documents produced for these queries. From Table 2, we can see that, for most WSEs, this amounts to over 2000 queries. We can then train each of our pseudo-WSEs to reproduce the corresponding WSE's ranking as accurately as possible, in effect treating the training process as a restricted learning-to-rank problem (Joachims, Li, Liu, & Zhai, 2007).

The next issue is how the effectiveness of the pseudo-WSEs should be ascertained during training. If we restrict the documents retrieved for a given query to the same documents that the real WSE retrieved, then all standard IR measures will give 1.0, as all and only relevant documents are retrieved. However, we are not interested in the precision and recall of our pseudo-WSEs but, instead, in the extent to which their rankings correlate with the real WSEs. In particular, we use three different measures to quantify the extent to which our pseudo-WSEs achieve the correct ranking of documents: nDCG, τ_{ap}, and ρ.

Discounted cumulative gain (DCG (Järvelin & Kekäläinen, 2002)) is a standard IR evaluation measure that uses graded (i.e., non-binary) relevance labels. It is computed by summing the relevance labels of all retrieved documents, penalised by the logarithm of the rank position of each document:

$$\mathrm{DCG} = \sum_{i=1}^{k} \frac{\mathrm{rel}(i)}{\log_2(i+1)} \tag{2}$$

where k is the total number of retrieved documents and $\mathrm{rel}(i)$ is the relevance label assigned to the ith retrieved document. In order to make the DCG values for different queries comparable, we employ its normalised version, nDCG (Järvelin & Kekäläinen, 2002):

$$\mathrm{nDCG} = \frac{\mathrm{DCG}}{\mathrm{iDCG}} \tag{3}$$

where DCG is given by Eq. (2) and iDCG is the ideal DCG value, obtained by producing a perfect ranking for the same query. While nDCG is normally applied with up to five levels of relevance, we apply it with up to 24 levels (i.e., the maximum number of results retrieved per evidence identification query), with the highest level denoting the top-ranked document for a given query. The final nDCG value is then calculated over the ranking of documents up to the number of relevant (retrieved) documents.

Average precision (AP) correlation (τ_{ap} (Yilmaz, Aslam, & Robertson, 2008)) is based on average precision and extends the traditional Kendall's τ rank correlation coefficient (Kendall, 1938) by performing a "rank-aware" correlation, thus distinguishing between errors towards higher ranks and those towards lower ranks. It is defined as:

$$\tau_{ap} = \frac{2}{k-1} \sum_{i=2}^{k} \left(\frac{\mathrm{con}(i)}{i-1} \right) - 1 \tag{4}$$

where $\mathrm{con}(i)$ is the number of concordant pairs of documents with respect to a reference ranking above rank i. In our experiments, the reference rankings are those obtained from the different WSEs.

Finally, we also use Spearman's rank correlation coefficient (Spearman's ρ (Wackerly, Mendenhall, & Scheaffer, 2002)), which is equivalent to computing a linear correlation using the ranks (rather than the scores) of the retrieved documents:

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

$$\rho = \frac{\sum_j (x_j - \bar{x})(y_j - \bar{y})}{\sqrt{\sum_j (x_j - \bar{x})^2 \sum_i (y_j - \bar{y})^2}}. \quad (5)$$

where x_j and y_j are the rank positions of document j according to the ranked lists x and y, respectively. In our experiments, x is produced by a WSE, while y is produced by the corresponding pseudo-WSE.

4.5. Evidence combination

Besides evaluating a pseudo-WSE in comparison to a baseline intranet ranking, we also evaluate the combination of these two rankings. In particular, to combine the results of the intranet and external expert search engines—denoted *int* and *ext*, respectively—we apply a weighted sum:

$$score\_cand_{final}(C, Q) = w_{int} \times score\_cand_{int}(C, Q) + w_{ext} \times score\_cand_{ext}(C, Q), \quad (6)$$

where $score\_cand_{\{int,ext\}}(C, Q)$ can be any voting technique—here, we use the expCombMNZ voting technique from Eq. (1)—and may be unbounded, while w_{int} and w_{ext} combine the roles of normalising candidate scores and weighting the importance of the intranet and external evidence. Moreover, by combining separate candidate rankings, we do not mix statistics of distinct collections. In our experiments, we combine the rankings produced by an expert search engine based on the intranet evidence with each of the rankings produced using pseudo-WSEs, in both their untrained and trained versions. In all cases, the intranet and external engines apply the same document weighting model and the same voting technique (i.e., expCombMNZ). In order to avoid temporal disparities between the Web evidence acquired for the EX07 and EX08 topic sets, we train the weights w_{int} and w_{ext} for each set independently, by performing a simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) through a five-fold cross validation.

5. Mimicking WSEs' document rankings

In this section, we address our first research question, regarding the accuracy of our mimicking strategies. In particular, we want to assess the extent to which our pseudo-WSEs can reproduce the document rankings provided by the different WSEs. In this investigation, we apply the aforementioned document weighting models (i.e., BM25, LM, PL2, and DLH13) for our pseudo-WSEs in their default and trained settings. Table 3 reports the increase in nDCG[6] achieved by the trained settings of each of these models according to the three considered measures (i.e., nDCG, τ_{ap}, and ρ) when compared to the model's untrained setting. Significance between the untrained (i.e., baseline) and trained settings is calculated using the Wilcoxon signed-rank matched-pairs test and denoted by one of five symbols: =, denoting no significant difference from the baseline; < and >, denoting significant decreases or increases compared to the baseline with $p < 0.05$; and $\ll$ and $\gg$, denoting significant decreases or increases compared to the baseline with $p < 0.01$. The best achieved mimicking of each WSE is shown in bold.

From Table 3, we note, as expected, that nDCG can be improved by training. Moreover, while there is no clearly preferred training measure, all three measures bring significant mimicking improvements across most settings, with the exception of Google Blogs and Google News, which provide the least evidence in Table 2. Indeed, significance is likely to occur for very small improvements on potentially large sets of queries. Examining the best settings, we note that, when no training is applied, DLH13 is the most effective weighting model in 5 out of 7 cases. Moreover, when training is applied, it remains best for three WSEs.[7] Among the other weighting models, LM seems to perform best overall, with and without training. Further improvements towards better reproducing the underlying WSEs' ranking mechanisms might be possible by the use of additional features by the corresponding pseudo-WSEs, such as anchor text or linkage information. However, this is made difficult because the real WSEs can leverage all anchor text identified for each document from the entire Web. In this sense, our pseudo-WSEs can never behave identically to the corresponding WSEs, due to their lack of knowledge of the whole Web surrounding the documents that they act on. Nevertheless, the very high values reported for nDCG in Table 3 attest the accuracy of our mimicking strategies. Recalling our first research question, this confirms the feasibility of building pseudo-WSEs that almost clone the original WSEs' document rankings, while providing our approach with scoring information for each ranked document.

6. Leveraging Web search results for expert search

In this section, we evaluate the expert search retrieval performance of using our pseudo-WSEs in comparison to a baseline expert search engine based on intranet evidence only. By doing so, we address our second research question. Additionally, we want to quantify the impact of further refining the mimicking of WSEs' rankings through training on the final expert search performance. As in Section 5, we build pseudo-WSEs using four document weighting models, in their untrained and trained versions. From these document rankings, up to rank 1000, we then apply the expCombMNZ voting technique (Eq. (1)).

[6] Results for τ_{ap} and ρ show similar trends.

[7] Note that this does not contradict the observed usefulness of training; instead, it attests the effectiveness of DLH13 compared to the other weighting models in their untrained settings.

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

Table 3
nDCG performances before and after training pseudo-WSEs to better mimic the corresponding WSEs using evidence identification queries for all EX07 and EX08 127 topics. nDCG, AP correlation (τ_{ap}), and Spearman's rank coefficient (ρ) are used for training. DLH13 has no parameters to train.

Source	Mimicking	BM25	LM	PL2	DLH13
Google	UNTRAINED	0.9337	0.9366	0.8917	0.9400
	TRAINED(nDCG)	0.9389$^{\gg}$	**0.9416**$^{\gg}$	0.9045$^{\gg}$	
	TRAINED(τ_{ap})	0.9385$^{\gg}$	0.9409$^{\gg}$	0.8917$^{=}$	
	TRAINED(ρ)	0.9387$^{\gg}$	0.9412$^{\gg}$	0.9045$^{\gg}$	
Yahoo!	UNTRAINED	0.9110	0.9152	0.9045	**0.9159**
	TRAINED(nDCG)	0.9132$^{\gg}$	0.9154$^{=}$	0.9084$^{\gg}$	
	TRAINED(τ_{ap})	0.9109$^{\ll}$	0.9153$^{=}$	0.9076$^{\gg}$	
	TRAINED(ρ)	0.9131$^{\gg}$	0.9155$^{=}$	0.9085$^{\gg}$	
Google-pdf	UNTRAINED	0.9436	0.9540	0.9069	0.9553
	TRAINED(nDCG)	0.9529$^{\gg}$	**0.9568**$^{\gg}$	0.9156$^{\gg}$	
	TRAINED(τ_{ap})	0.9529$^{\gg}$	0.9566$^{\gg}$	0.9154$^{\gg}$	
	TRAINED(ρ)	0.9529$^{\gg}$	**0.9568**$^{\gg}$	0.9155$^{\gg}$	
Yahoo!-pdf	UNTRAINED	0.9117	0.9172	0.9086	**0.9180**
	TRAINED(nDCG)	0.9123$^{>}$	0.9176$^{=}$	0.9164$^{\gg}$	
	TRAINED(τ_{ap})	0.9114$^{=}$	0.9174$^{=}$	0.9164$^{\gg}$	
	TRAINED(ρ)	0.9119$^{=}$	0.9176$^{=}$	0.9164$^{\gg}$	
Scholar	UNTRAINED	0.9427	0.9473	0.9302	**0.9483**
	TRAINED(nDCG)	0.9443$^{\gg}$	**0.9483**$^{\gg}$	0.9361$^{\gg}$	
	TRAINED(τ_{ap})	0.9441$^{\gg}$	0.9478$^{=}$	0.9358$^{\gg}$	
	TRAINED(ρ)	0.9441$^{\gg}$	0.9468$^{=}$	0.9334$^{\gg}$	
Blogs	UNTRAINED	0.9867	0.9933	0.9890	0.9926
	TRAINED(nDCG)	0.9909$^{=}$	**0.9944**$^{=}$	0.9920$^{=}$	
	TRAINED(τ_{ap})	0.9894$^{=}$	**0.9944**$^{=}$	0.9886$^{=}$	
	TRAINED(ρ)	0.9909$^{=}$	**0.9944**$^{=}$	0.9920$^{=}$	
News	UNTRAINED	0.9786	0.9785	0.9795	0.9753
	TRAINED(nDCG)	**0.9815**$^{=}$	0.9785$^{=}$	0.9751$^{=}$	
	TRAINED(τ_{ap})	0.9761$^{=}$	0.9728$^{=}$	0.9804$^{=}$	
	TRAINED(ρ)	0.9788$^{=}$	0.9785$^{=}$	0.9794$^{=}$	

Table 4 presents the results of our experiments using the TREC EX07 and EX08 topics (CE-001–CE-050 and CE-051–CE-127, respectively) in terms of standard mean average precision (MAP). Statistical significance between each of the default and trained settings of our pseudo-WSEs and a baseline expert search engine based on intranet evidence is denoted with one of the symbols described in Section 5. A second symbol denotes the significance of each trained setting with respect to the best performing among the default and trained settings, which is itself marked with an asterisk (*). For example, for the EX07 topics, both the untrained and trained settings of BM25 + Yahoo!-pdf significantly underperform when compared to the intranet baseline (MAP 0.3576), while three of them do not significantly differ from the best trained setting (τ_{ap}, MAP 0.2325). The best performance using each WSE is shown in bold. A row with the performance reported by Jiang et al. (2008) using external results retrieved by Google is also included. Note that a direct comparison to the works by Serdyukov et al. (2008) and Balog and de Rijke (2008) is not possible, as both use summaries (snippets) of search results instead of their full content, and the latter (i.e., (Balog & de Rijke, 2008)) does not report on their performance using evidence exclusively from outside the domain of the organisation under consideration. The performance reported by Serdyukov and Hiemstra (2008a) using titles and snippets is included as a reference value in the investigation described in Section 8.[8]

From the results in Table 4, we firstly note that some of the considered WSEs can be effectively applied for identifying relevant experts in the CERC collection. In particular, pseudo-WSEs mimicking Google and Yahoo! markedly outperform the intranet baseline for most settings for the EX07 topics, albeit not significantly. Moreover, these results show that, using exactly the same document ranking techniques, it can be more effective to mine the Web than the intranet of the actual organisation. This is somewhat expected as, given the size of the Web, it is possible that some prominent experts will have useful expertise evidence outside of their organisation's intranet. This is typical of research organisations such as CSIRO, as researchers write papers and give talks outside the organisations, which leads to their name and some evidence of their

[8] Our goal in providing published results in the literature is to offer the reader a reference value for our own results. Although care has been taken to ensure that such a reference value was produced in similar experimental conditions to our reported results, one has to bear in mind the dynamic nature of this particular research when directly comparing these results. Indeed, WSEs' results may vary depending on several factors, including the time and the location from, where these WSEs were queried.

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

Table 4
MAP results of applying the expCombMNZ voting technique to trained and untrained document rankings produced by each pseudo-WSE for the 127 topics from the EX07 and EX08 tasks.

Source	Mimicking	EX07 topics (CE-001–CE-050)				EX08 topics (CE-051–CE-127)			
		BM25	LM	PL2	DLH13	BM25	LM	PL2	DLH13
Intranet		0.3576	0.3366	0.3582	0.3560	0.3481	0.3365	0.3543	0.3656
Google	UNTRAINED	0.3805^{--}	0.3797^{--}	0.3133^{--}	0.3807^{-}	0.2293$^{\ll\ll}$	**0.3044**$^{-*}$	0.1819$^{\ll-}$	0.3010$^{<}$
	TRAINED(nDCG)	0.3873^{--}	0.3826^{--}	0.3104^{--}		0.2379$^{\ll-}$	0.2852$^{\ll}$	0.1935$^{\ll-}$	
	TRAINED(τ_{ap})	0.3874^{--}	0.3840^{-*}	0.3133^{--}		0.2362$^{\ll-}$	0.2925$^{-<}$	0.1819$^{\ll-}$	
	TRAINED(ρ)	**0.3881**$^{-*}$	0.3834^{--}	0.3151^{-*}		0.2381$^{\ll*}$	0.2892$^{-<}$	0.1950$^{\ll*}$	
Yahoo!	UNTRAINED	0.4017^{--}	0.3999^{--}	0.3210^{--}	**0.4116**$^{-}$	0.2502$^{\ll-}$	0.3263^{--}	0.2057$^{\ll-}$	0.3177^{-}
	TRAINED(nDCG)	0.4004^{--}	0.4065^{--}	0.3316^{--}		0.2530$^{\ll*}$	0.3265^{--}	0.2334$^{\ll-}$	
	TRAINED(τ_{ap})	0.4059^{-*}	0.4066^{-*}	0.3527^{-*}		0.2500$^{\ll-}$	**0.3270**$^{-*}$	0.2571$^{\ll*}$	
	TRAINED(ρ)	0.4016^{--}	0.4066^{-*}	0.3321^{--}		0.2520$^{\ll<}$	0.3269^{--}	0.2343$^{\ll-}$	
Google-pdf	UNTRAINED	0.2392$^{\ll*}$	0.3197^{-*}	0.1518$^{\ll\ll}$	**0.3250**$^{-}$	0.1671$^{\ll*}$	**0.2571**$^{<*}$	0.1270$^{\ll\ll}$	0.2544$^{\ll}$
	TRAINED(nDCG)	0.2330$^{\ll-}$	0.3154^{--}	0.1963$^{\ll-}$		0.1612$^{\ll-}$	0.2449$^{\ll-}$	0.1456$^{\ll-}$	
	TRAINED(τ_{ap})	0.2309$^{\ll-}$	0.3157^{--}	0.1994$^{\ll-}$		0.1607$^{\ll-}$	0.2436$^{\ll-}$	0.1483$^{\ll-}$	
	TRAINED(ρ)	0.2310$^{\ll-}$	0.3156^{--}	0.1996$^{\ll*}$		0.1609$^{\ll-}$	0.2446$^{\ll-}$	0.1491$^{\ll*}$	
Yahoo!-pdf	UNTRAINED	0.2314$^{\ll-}$	0.2906^{--}	0.2324$^{\ll-}$	0.3036^{-}	0.1640$^{\ll<}$	0.2738$^{-<}$	0.1447$^{\ll\ll}$	0.2748$^{\ll}$
	TRAINED(nDCG)	0.2323$^{\ll-}$	0.2958^{--}	0.2748$^{<*}$		0.1656$^{\ll-}$	**0.2808**$^{-*}$	0.1985$^{\ll*}$	
	TRAINED(τ_{ap})	0.2325$^{\ll*}$	**0.3049**$^{-*}$	0.2745$^{<-}$		0.1663$^{\ll*}$	0.2720$^{<-}$	0.1984$^{\ll-}$	
	TRAINED(ρ)	0.2315$^{\ll-}$	0.2970^{--}	0.2745$^{<-}$		0.1645$^{\ll-}$	**0.2808**$^{-*}$	0.1984$^{\ll-}$	
Scholar	UNTRAINED	0.1074$^{\ll-}$	0.1563$^{\ll\ll}$	0.1156$^{\ll\ll}$	0.1732$^{\ll}$	0.0561$^{\ll<}$	0.0976$^{\ll<}$	0.0515$^{\ll\ll}$	**0.1084**$^{\ll}$
	TRAINED(nDCG)	0.1115$^{\ll*}$	0.1695$^{\ll-}$	0.1297$^{\ll\ll}$		0.0568$^{\ll-}$	0.1040$^{\ll-}$	0.0632$^{\ll-}$	
	TRAINED(τ_{ap})	0.1092$^{\ll<}$	0.1730$^{\ll-}$	0.1337$^{\ll<}$		0.0569$^{\ll*}$	0.1052$^{\ll*}$	0.0686$^{\ll-}$	
	TRAINED(ρ)	0.1091$^{\ll-}$	**0.1761**$^{\ll*}$	0.1573$^{\ll*}$		0.0568$^{\ll-}$	0.1009$^{\ll<}$	0.0840$^{\ll*}$	
Blogs	UNTRAINED	0.0433$^{\ll-}$	0.0541$^{\ll-}$	0.0421$^{\ll<}$	**0.0554**$^{\ll}$	0.0252$^{\ll-}$	0.0403$^{\ll-}$	0.0254$^{\ll\ll}$	**0.0425**$^{\ll}$
	TRAINED(nDCG)	0.0433$^{\ll-}$	**0.0554**$^{\ll*}$	0.0439$^{\ll*}$		0.0255$^{\ll*}$	0.0403$^{\ll-}$	0.0309$^{\ll*}$	
	TRAINED(τ_{ap})	0.0448$^{\ll*}$	0.0553$^{\ll-}$	0.0428$^{\ll-}$		0.0255$^{\ll*}$	0.0412$^{\ll*}$	0.0269$^{\ll\ll}$	
	TRAINED(ρ)	0.0431$^{\ll-}$	**0.0554**$^{\ll*}$	0.0439$^{\ll*}$		0.0255$^{\ll*}$	0.0402$^{\ll-}$	0.0307$^{\ll-}$	
News	UNTRAINED	0.0179$^{\ll*}$	0.0185$^{\ll-}$	0.0167$^{\ll-}$	0.0182$^{\ll}$	0.0125$^{\ll-}$	0.0166$^{\ll-}$	0.0168$^{\ll-}$	0.0182$^{\ll}$
	TRAINED(nDCG)	0.0178$^{\ll-}$	0.0185$^{\ll-}$	0.0184$^{\ll-}$		0.0126$^{\ll*}$	0.0166$^{\ll-}$	0.0182$^{\ll*}$	
	TRAINED(τ_{ap})	0.0179$^{\ll*}$	0.0185$^{\ll-}$	0.0165$^{\ll-}$		0.0117$^{\ll-}$	**0.0189**$^{\ll*}$	0.0163$^{\ll<}$	
	TRAINED(ρ)	0.0179$^{\ll*}$	0.0189$^{\ll*}$	**0.0204**$^{\ll*}$		0.0125$^{\ll-}$	0.0163$^{\ll-}$	0.0141$^{\ll<}$	
Google (Jiang et al., 2008)		0.3173				Results not reported by the authors.			

expertise appearing on Web sites other than their own. A comparison to the performance of the candidate-centric approach of Jiang et al. (2008) using results retrieved by Google (last row in the table) further attests the effectiveness of our mimicking strategy. As for the remaining WSEs, restricting Google and Yahoo! to retrieve only PDF documents degrades the retrieval performance, while still not significantly differing from the intranet baseline. The next most effective source is Google Scholar, while Google Blogs and Google News have much lower performances,[9] which is not surprising, given that these WSEs provided the least evidence, as previously discussed in Section 4.

A similar scenario is observed for the EX08 topics, however with a lower performance on most settings, and significantly lower than that of the intranet baseline. This suggests that the expertise needs from the EX08 topic set are apparently less likely to be answered on the Web when compared to those from the EX07 topic set. Given the longer time elapsed since the development and release of the EX07 topics, it seems plausible that more evidence is available on the Web for this topic set compared to the EX08 one. Indeed, using the procedure described in Section 3 for obtaining expertise evidence from WSEs, the median WSE (among the seven considered) retrieves an average of 334.20 documents per EX07 topic while only 268.97 are retrieved for each EX08 topic. Nonetheless, the performances attained by LM + Google and LM + Yahoo! are notable exceptions, as they do not significantly differ from the intranet baseline, once again showing that useful evidence can be mined from WSEs, even when the apparently more difficult EX08 topic set is considered. Overall, these results show that we can effectively leverage the general Google and Yahoo! WSEs for expert search, hence answering our second research question.

As for the considered mimicking strategies, the (untrained) DLH13 model performs generally best for the EX07 topics, while LM is overall the best for EX08. When mimicking is refined by training the pseudo-WSEs, a corresponding improvement in retrieval performance over untrained pseudo-WSEs is observed for the majority of cases. For the EX07 topic set, out

[9] Note that this is not a reflection of these WSEs' quality as such, but of their usefulness as a source of expertise evidence.

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

of 21 possible cases, training BM25, LM, and PL2 improves in 11, 16, and 18 cases, respectively. For the EX08 topic set, improvements are attained in 15, 10, and 18 out of 21 cases for BM25, LM, and PL2, respectively. Spearman's ρ yields the best results in most cases (11) for the EX07 topics, followed by τ_{ap} (8) and nDCG (3). For the EX08 topics, τ_{ap} is the most effective training measure (eight cases), followed closely by nDCG (7) and Spearman's ρ (6). As a whole, refining the mimicking of WSEs using any of nDCG, τ_{ap}, or ρ is beneficial. Improvements are particularly marked for PL2 on both topic sets, and can also be significant, particularly for the EX08 topics. This further confirms our findings in Section 5, regarding the potential of refining our mimicking approach.

7. Combining pseudo-WSEs with intranet evidence

Compared to a typical enterprise search scenario, where only intranet evidence is used, the retrieval performance achieved by the pseudo-WSEs in Section 6 is impressive, particularly for the EX07 topics. Hence, a natural question that arises is whether this external evidence can be combined with an existing expert search engine operating with intranet data. As the two sources of expertise evidence do not overlap, they should be independent and their combination should result in an increase in retrieval performance. In this section, we address our third research question, by investigating whether the external and intranet evidence can be successfully combined. Additionally, similarly to the previous section, we assess the impact on the expert search performance of this combination after refining the mimicking of the WSEs' underlying ranking, by training the corresponding pseudo-WSEs.

Table 5 presents the results of our experiments for both EX07 and EX08 topics and additionally includes the intranet baselines from Table 4. Significant increases over the baseline and the best among the untrained and trained settings are shown using the symbols previously introduced in Section 6. As an additional baseline, a row with the performance reported by Jiang et al. (2008) using Google is included. Note, however, that they do not perform an explicit integration of internal and external evidence. Instead, they simply allow Google to also retrieve results from inside the CERC collection. Nevertheless, this is the closest attempt to combine full content evidence gathered from both inside and outside this collection that we are aware of.

With respect to our third research question, we note that combining the rankings based on mimicked WSEs with that based on existing intranet evidence can markedly improve both the intranet baseline and the corresponding pseudo-WSEs alone (see Table 4). Indeed, for both the TREC EX07 and EX08 topic sets, except for Google News, all mimicked WSEs can bring improvements to the existing intranet-based expert search engine. On the EX07 topic set, marked improvements are attained by integrating expertise evidence from Google, Yahoo!, and Yahoo!-pdf. These gains can be significant, in particular for Google using LM, and for Yahoo! using both BM25 and LM. However, for the EX08 topic set, less marked improvements are observed, which are not significant with respect to the baseline intranet performance. These results confirm our observations in Section 6, regarding the increased difficulty of the EX08 topic set compared to EX07.

As for the untrained weighting models used for mimicking the document rankings produced by the different WSEs, LM performs the best, improving the intranet baseline in 6 out of 7 possible cases, followed by BM25, DLH13, and PL2, with improvements in 4, 4, and 1 out of 7 cases, respectively. On the EX08 topics, a different trend is observed, with BM25 improving the intranet baseline in 5 out of 7 cases. LM, PL2, and DLH13 bring improvements in 3, 2, and 2 out of 7 cases, respectively. When compared to the performance reported by Jiang et al. (2008) using Google, we note that marked improvements are attained using all four untrained weighting models, again attesting the effectiveness of our approach.

By examining the effect of refining the mimicking of a pseudo-WSE on the effectiveness of the combined expert search engine, we note that the retrieval performance of the latter is not consistently enhanced compared to using an untrained pseudo-WSE. For the EX07 topics, out of 21 possible cases, training BM25, LM, and PL2 improves over their untrained version in 7, 11, and 14 cases, respectively. For the EX08 topics, improvements are observed in 8, 13, and 11 cases, respectively. In particular, the additional layer of training involved in combining internal and external expert search rankings seems to attenuate the benefits of adopting a trained mimicking strategy, in contrast to the observations in Section 6.

Besides being effective on its own, as shown in Section 6, the results in Table 5—particularly those on the EX07 topics—show that our approach for mimicking WSEs' rankings can be successfully integrated to an existing intranet-based expert search engine, hence answering our third research question. Additional improvements could be achieved if we further enhanced the mimicking and training procedures by integrating features such as candidate query term proximity or candidate homepage detection, which were shown to be effective on this task (Macdonald, Hannah, & Ounis, 2008), e.g., within a learning-to-rank setting (Joachims et al., 2007). Moreover, enhancing the mimicking of WSEs by extracting additional features from the available documentary evidence of expertise could also help overcome the sparsity of this evidence for the more recent EX08 topics.

8. Leveraging alternative expertise evidence from WSEs

In the previous sections, we showed that mimicking WSEs by trying to reproduce their rankings for evidence identification queries can be an effective approach for expert search. By doing so, we have answered the hypothetical question *what if WSEs did expert search?* Since indexing documents and scoring these documents in response to a query are an integral part of any WSE's searching mechanism, efficiency would not be a concern if our approach was to be deployed in a Web search

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

Table 5
MAP results of combining expert search engines based on intranet and pseudo-WSEs—the latter trained using different measures—for the 127 topics from the EX07 and EX08 tasks.

Source	Mimicking	EX07 topics (CE-001–CE-050)				EX08 topics (CE-051–CE-127)			
		BM25	LM	PL2	DLH13	BM25	LM	PL2	DLH13
Intranet		0.3576	0.3366	0.3582	0.3560	0.3481	0.3365	0.3543	0.3656
Google	UNTRAINED	0.4198$^{=<}$	0.4134$^{>*}$	0.3857$^{==}$	0.4137$^{=}$	0.3573$^{==}$	0.3384$^{==}$	0.3536$^{==}$	**0.3824**$^{=}$
	TRAINED(nDCG)	0.4196$^{==}$	0.4129$^{>=}$	0.4035$^{==}$		0.3582$^{=*}$	0.3409$^{==}$	0.3515$^{==}$	
	TRAINED(τ_{ap})	**0.4267**$^{=*}$	0.4053$^{==}$	0.3824$^{==}$		0.3324$^{==}$	0.3592$^{=*}$	0.3571$^{==}$	
	TRAINED(ρ)	0.4265$^{==}$	0.4057$^{==}$	0.4050$^{=*}$		0.3437$^{==}$	0.3334$^{==}$	0.3588$^{=*}$	
Yahoo!	UNTRAINED	0.4348$^{==}$	0.4371$^{\gg=}$	0.3531$^{==}$	0.3988$^{=}$	0.3566$^{=*}$	**0.3677**$^{=*}$	0.3632$^{=*}$	0.3667$^{=}$
	TRAINED(nDCG)	0.4431$^{>*}$	0.4437$^{\gg=}$	0.3521$^{==}$		0.3492$^{==}$	0.3664$^{==}$	0.3605$^{==}$	
	TRAINED(τ_{ap})	0.4338$^{=<}$	0.4437$^{\gg=}$	0.3946$^{=*}$		0.3484$^{==}$	0.3657$^{==}$	0.3608$^{==}$	
	TRAINED(ρ)	0.4328$^{==}$	**0.4439**$^{\gg*}$	0.3605$^{==}$		0.3380$^{==}$	0.3666$^{==}$	0.3476$^{==}$	
Google-pdf	UNTRAINED	0.3397$^{==}$	0.3559$^{=<}$	0.3421$^{<=}$	0.3711$^{=}$	0.3493$^{==}$	0.3281$^{==}$	0.3322$^{==}$	0.3465$^{=}$
	TRAINED(nDCG)	0.3489$^{=*}$	0.3610$^{==}$	0.3502$^{=*}$		0.3532$^{=*}$	0.3292$^{==}$	**0.3602**$^{=*}$	
	TRAINED(τ_{ap})	0.3260$^{==}$	**0.3727**$^{=*}$	0.3419$^{==}$		0.3421$^{==}$	0.3516$^{=*}$	0.3506$^{==}$	
	TRAINED(ρ)	0.3444$^{==}$	0.3593$^{==}$	0.3438$^{<=}$		0.3407$^{==}$	0.3310$^{==}$	0.3430$^{==}$	
Yahoo!-pdf	UNTRAINED	0.3808$^{=*}$	0.3702$^{==}$	0.3448$^{<=}$	**0.3979**$^{=}$	0.3540$^{=*}$	0.3262$^{==}$	0.3601$^{=*}$	**0.3648**$^{=}$
	TRAINED(nDCG)	0.3797$^{==}$	0.3892$^{==}$	0.3484$^{==}$		0.3538$^{==}$	0.3461$^{==}$	0.3524$^{==}$	
	TRAINED(τ_{ap})	0.3703$^{==}$	0.3836$^{==}$	0.3603$^{=*}$		0.3526$^{==}$	0.3348$^{==}$	0.3545$^{==}$	
	TRAINED(ρ)	0.3802$^{==}$	0.3898$^{=*}$	0.3466$^{==}$		0.3457$^{==}$	0.3466$^{=*}$	0.3589$^{==}$	
Scholar	UNTRAINED	**0.3598**$^{=*}$	0.3457$^{=*}$	0.3523$^{==}$	0.3540$^{=}$	0.3488$^{=*}$	0.3384$^{=*}$	0.3480$^{==}$	**0.3648**$^{=}$
	TRAINED(nDCG)	0.3540$^{==}$	0.3269$^{<\ll}$	0.3590$^{=*}$		0.3339$^{==}$	0.3373$^{==}$	0.3570$^{=*}$	
	TRAINED(τ_{ap})	0.3584$^{==}$	0.3450$^{==}$	0.3585$^{==}$		0.3465$^{==}$	0.3384$^{=*}$	0.3552$^{==}$	
	TRAINED(ρ)	0.3544$^{==}$	0.3359$^{<\ll}$	0.3585$^{==}$		0.3329$^{==}$	0.3307$^{==}$	0.3309$^{==}$	
Blogs	UNTRAINED	0.3547$^{==}$	0.3310$^{==}$	0.3558$^{=*}$	0.3533$^{=}$	0.3430$^{==}$	0.3298$^{==}$	0.3525$^{==}$	**0.3625**$^{=}$
	TRAINED(nDCG)	0.3415$^{\ll\ll}$	0.3206$^{\ll}$	0.3552$^{==}$		0.3482$^{=*}$	0.3319$^{==}$	0.3493$^{==}$	
	TRAINED(τ_{ap})	0.3435$^{==}$	0.3376$^{=*}$	0.3505$^{\ll\ll}$		0.3482$^{=*}$	0.3362$^{=*}$	0.3364$^{==}$	
	TRAINED(ρ)	**0.3570**$^{=*}$	0.3282$^{==}$	0.3542$^{==}$		0.3449$^{==}$	0.3223$^{==}$	0.3543$^{=*}$	
News	UNTRAINED	0.3538$^{\ll}$	0.3355$^{==}$	0.3497$^{\ll}$	0.3538$^{<}$	0.3407$^{==}$	0.3326$^{==}$	0.3314$^{==}$	**0.3618**$^{=}$
	TRAINED(nDCG)	0.3470$^{\ll<}$	0.3223$^{\ll}$	0.3542$^{=<}$		0.3471$^{==}$	0.3357$^{=*}$	0.3518$^{==}$	
	TRAINED(τ_{ap})	0.3356$^{\ll\ll}$	0.3361$^{=*}$	**0.3574**$^{=*}$		0.3473$^{=*}$	0.3342$^{==}$	0.3514$^{==}$	
	TRAINED(ρ)	0.3561$^{=*}$	0.3313$^{==}$	0.3491$^{==}$		0.3449$^{==}$	0.3339$^{==}$	0.3529$^{=*}$	
Google (Jiang et al., 2008)		0.3769				Results not reported by the authors.			

scenario. Nonetheless, in this section, we address the practical problem of deploying our approach in a real enterprise scenario. Despite the costs incurred in building an initial expertise base using evidence from WSEs (which is even more problematic for candidate-centric approaches, as discussed in Section 2), maintaining our approach in a typical expert search environment should not represent a significant burden for an enterprise. For instance, the initial base could be updated on-demand as new employees entered or leaved the enterprise, or as new topics became of interest to its users. As a fall-back plan, if no external evidence is available for a given topic or candidate, the intranet alone could still be used to effectively answer users' queries, as denoted by the strong performance of our baseline intranet expert search engine used in the previous sections.

Nevertheless, for a given unseen expert search request, the system would have to query a WSE in order to identify expertise evidence and build the profile of every candidate. Moreover, the full documents that comprise these profiles would have to be downloaded. For instance, in our current setting, as detailed in Section 4, we download up to the top 24 documents retrieved for each expertise identification query. Additionally, these profiles would have to be updated from time to time—again, a problem also faced by candidate-centric approaches. These costs, however, could be overcome by the use of an alternative form of expertise evidence extracted from the WSEs' result pages.

In this section, we address our fourth and last research question, by investigating the effectiveness of leveraging the metadata associated to the results retrieved by WSEs, such as their titles and descriptive snippets, instead of downloading the corresponding full documents. In particular, for each result for an evidence identification query, we extract its title and snippet, combining them together as a single document of expertise evidence to compose the profile of a candidate. Such evidence is then indexed and scored by our pseudo-WSEs, in the same way as done for the full documents.

In Table 6, analogously to the results presented in Table 3 using full documents, we contrast the nDCG performances attained by the pseudo-WSEs using the combined indices of titles + snippets under different training settings. In particular, we train our pseudo-WSEs using the same set of 12,068 evidence identification queries used in Section 4.4, however using the derived rankings of titles + snippets. Relative to indexing the full documents retrieved by the WSEs, the evidence retrieved in

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

Table 6
nDCG performances before and after training pseudo-WSEs to better mimic the corresponding WSEs using evidence identification queries on titles + snippets. nDCG, AP correlation (τ_{ap}), and Spearman's rank coefficient (ρ) are used for training. DLH13 has no parameters to train.

Source	Mimicking	BM25	LM	PL2	DLH13
Google	UNTRAINED	0.8807	0.8817	0.8843	0.8873
	TRAINED(nDCG)	0.8846$^{\gg}$	0.8863$^{\gg}$	**0.8903**$^{\gg}$	
	TRAINED(τ_{ap})	0.8844$^{\gg}$	0.8863$^{\gg}$	0.8900$^{\gg}$	
	TRAINED(ρ)	0.8845$^{\gg}$	0.8862$^{\gg}$	**0.8903**$^{\gg}$	
Yahoo!	UNTRAINED	0.8847	0.8858	0.8873	0.8896
	TRAINED(nDCG)	0.8871$^{\gg}$	0.8894$^{\gg}$	**0.8925**$^{\gg}$	
	TRAINED(τ_{ap})	0.8870$^{\gg}$	0.8891$^{\gg}$	0.8924$^{\gg}$	
	TRAINED(ρ)	0.8871$^{\gg}$	0.8894$^{\gg}$	**0.8925**$^{\gg}$	
Google-pdf	UNTRAINED	0.8913	0.8920	0.8948	0.8976
	TRAINED(nDCG)	0.8953$^{\gg}$	0.8960$^{\gg}$	**0.9004**$^{\gg}$	
	TRAINED(τ_{ap})	0.8953$^{\gg}$	0.8957$^{\gg}$	0.8985$^{\gg}$	
	TRAINED(ρ)	0.8951$^{\gg}$	0.8960$^{\gg}$	**0.9004**$^{\gg}$	
Yahoo!-pdf	UNTRAINED	0.8960	0.8969	0.8987	0.9011
	TRAINED(nDCG)	0.8992$^{\gg}$	0.8993$^{\gg}$	**0.9039**$^{\gg}$	
	TRAINED(τ_{ap})	0.8983$^{\gg}$	0.8992$^{\gg}$	0.9038$^{\gg}$	
	TRAINED(ρ)	0.8992$^{\gg}$	0.8993$^{\gg}$	**0.9039**$^{\gg}$	
Scholar	UNTRAINED	**0.8259**	**0.8259**	**0.8259**	**0.8259**
	TRAINED(nDCG)	**0.8259**$^{=}$	**0.8259**$^{=}$	**0.8259**$^{=}$	
	TRAINED(τ_{ap})	**0.8259**$^{=}$	0.8213$^{\ll}$	**0.8259**$^{=}$	
	TRAINED(ρ)	**0.8259**$^{=}$	0.8213$^{\ll}$	**0.8259**$^{=}$	
Blogs	UNTRAINED	0.7818	0.7817	0.7850	0.7846
	TRAINED(nDCG)	0.7818$^{=}$	0.7855$^{=}$	**0.7862**$^{=}$	
	TRAINED(τ_{ap})	0.7825$^{=}$	0.7855$^{=}$	0.7850$^{=}$	
	TRAINED(ρ)	0.7825$^{=}$	0.7855$^{=}$	**0.7862**$^{=}$	
News	UNTRAINED	0.9844	0.9909	0.9892	0.9909
	TRAINED(nDCG)	0.9892$^{=}$	0.9909$^{=}$	**0.9933**$^{=}$	
	TRAINED(τ_{ap})	0.9844$^{=}$	0.9909$^{=}$	**0.9933**$^{=}$	
	TRAINED(ρ)	0.9892$^{=}$	0.9909$^{=}$	**0.9933**$^{=}$	

the form of result titles and snippets is more sparse—for instance, titles and snippets do not always contain all the query terms used to retrieve them, as these terms are only present in the corresponding full documents. Additionally, the statistics derived from a collection of snippets are usually less refined, since the summarisation of full documents into snippets often involves a reduction of duplicated content (Carbonell & Goldstein, 1998). This makes it more difficult for the pseudo-WSEs to reproduce the original WSEs' rankings, which is reflected in the lower nDCG values reported in Table 6 when compared to those in Table 3 using the full documents.

Nevertheless, we compare the retrieval effectiveness of using titles + snippets to the counterpart pseudo-WSEs in Table 4, which use the full content of each retrieved result. Table 7 presents the results of applying the expCombMNZ voting technique on top of the pseudo-WSEs using either the combined index of titles + snippets ($T + S$) or the full retrieved documents (F) for the 127 topics from the EX07 and EX08 tasks. As we are mostly interested in comparing the effectiveness of these alternative document representations for expert search, we report the results obtained by mimicking WSEs using either representation without further training.[10] Significance is given by the Wilcoxon signed-rank matched-pairs test and denoted by the usual symbols. Additionally, the last row includes the results reported by Jiang et al. (2008)—a candidate-centric approach using Google—and Serdyukov and Hiemstra (2008a)—a document-centric approach using Yahoo!—on the EX07 topics. Results on the EX08 topics are not reported in these papers. Also note that the results reported by Serdyukov and Hiemstra (2008a) for WSEs other than Yahoo! are not restricted to external evidence only, and hence are not comparable to ours. Similarly, Balog and de Rijke (2008) and Serdyukov et al. (2008) do not report their results in isolation from the intranet evidence. Moreover, the results of the latter two approaches are based on either titles or snippets, but not both.

From Table 7, we observe that mimicking WSEs using titles + snippets performs even better than using the full content of each retrieved result. Indeed, for the EX07 topics, improvements over the full content baseline are obtained in 26 out of 28 possible cases. For the EX08 topics, the use of titles + snippets is preferable in 18 cases. Improvements are significant in several cases (10 for EX07, 13 for EX08). The highest improvements are observed for PL2, raising its observed poor performance

[10] Results obtained by their trained counterparts using the training measures described in Section 4.4 show a similar trend, however with generally higher absolute performances for both titles + snippets and full results.

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

Table 7
MAP results of applying the expCombMNZ voting technique to UNTRAINED rankings produced by pseudo-WSEs using a combined evidence of titles and snippets (T + S) and the corresponding full content baseline (F) from Table 4 for the EX07 and EX08 task topics.

Source	Index	EX07 topics (CE-001–CE-050)				EX08 topics (CE-051–CE-127)			
		BM25	LM	PL2	DLH13	BM25	LM	PL2	DLH13
Google	F	0.3805	0.3797	0.3133	0.3807	0.2293	0.3044	0.1819	0.3010
	T + S	0.3836$^{=}$	**0.3970**$^{=}$	0.3790$^{>}$	0.3811$^{=}$	**0.3148**$^{\gg}$	0.2974$^{=}$	0.3062$^{\gg}$	0.3136$^{=}$
Yahoo!	F	0.4017	0.3999	0.3210	0.4116	0.2502	0.3263	0.2057	0.3177
	T + S	0.4181$^{=}$	0.4103$^{=}$	0.4187$^{>}$	**0.4208**$^{=}$	0.3283$^{\gg}$	0.3193$^{=}$	0.3324$^{\gg}$	**0.3348**$^{=}$
Google-pdf	F	0.2392	0.3197	0.1518	**0.3250**	0.1671	**0.2571**	0.1270	0.2544
	T + S	0.3085$^{\gg}$	0.3023$^{=}$	0.3098$^{\gg}$	0.3123$^{=}$	0.2567$^{\gg}$	0.2498$^{=}$	0.2472$^{\gg}$	0.2519$^{=}$
Yahoo!-pdf	F	0.2314	0.2906	0.2324	0.3036	0.1640	0.2738	0.1447	**0.2748**
	T + S	**0.3404**$^{\gg}$	0.3172$^{=}$	0.3221$^{\gg}$	0.3380$^{=}$	0.2657$^{\gg}$	0.2624$^{=}$	0.2682$^{\gg}$	0.2694$^{=}$
Scholar	F	0.1074	0.1563	0.1156	0.1732	0.0561	0.0976	0.0515	0.1084
	T + S	0.2103$^{\gg}$	0.2115$^{>}$	0.2109$^{\gg}$	**0.2255**$^{>}$	0.1447$^{\gg}$	0.1442$^{\gg}$	**0.1494**$^{\gg}$	0.1479$^{\gg}$
Blogs	F	0.0433	0.0541	0.0421	0.0554	0.0252	0.0403	0.0254	0.0425
	T + S	0.0700$^{=}$	0.0699$^{=}$	0.0700$^{=}$	**0.0718**$^{=}$	0.0456$^{=}$	0.0422$^{=}$	**0.0460**$^{>}$	0.0437$^{=}$
News	F	0.0179	0.0185	0.0167	0.0182	0.0125	0.0166	0.0168	**0.0182**
	T + S	0.0194$^{=}$	0.0194$^{=}$	0.0211$^{=}$	**0.0261**$^{=}$	0.0122$^{=}$	0.0116$^{=}$	0.0119$^{=}$	0.0119$^{<}$
Google (Jiang et al., 2008)	T + S	0.3360				Results not reported by the authors.			
Yahoo! (Serdyukov & Hiemstra, 2008a)	T + S	0.4230							

on full documents to a comparable level to the other considered weighting models when titles + snippets are used. One reason for the overall higher performance of titles + snippets is that, besides representing high quality surrogates of the full documents, the expertise evidence from the titles + snippets is more numerous. Indeed, differently from full documents, they are extracted directly from the WSEs' result listings and hence are not prone to network errors, etc. during crawling. An interesting observation, however, is that LM shows a noticeable preference for the full content documents on the EX08 topics.

By comparing our results to those reported by Jiang et al. (2008) and Serdyukov and Hiemstra (2008a) and reproduced in Table 7, we observe a marked improvement over the former and a comparable performance to the latter. Recalling our last research question, these results attest the feasibility of using this alternative evidence from WSEs, drastically reducing the incurred costs, particularly in terms of network traffic (Serdyukov & Hiemstra, 2008a). On average, for each evidence identification query, the downloaded content was reduced from 550KB (i.e., the result pages and up to 24 full documents) to only 8KB (i.e., the result pages only)! Moreover, the results in Table 7 also show that leveraging titles + snippets can be even more effective than using the corresponding full documents. Indeed, DLH13 + Yahoo! using the titles + snippets evidence provided the best expert search retrieval performance across all individual strategies investigated in this article.

9. Conclusions

In this article, using a uniform experimental setting, we have conducted a thorough investigation of the usefulness of Web search engines (WSEs) as expert search systems. In particular, we have shown that WSEs can be effectively used for expert search by explicitly trying to mimic their underlying ranking mechanisms.

Using a standard enterprise test collection and two topic sets from the TREC 2007 and 2008 Enterprise Expert Search tasks, we investigated the effectiveness of leveraging external evidence using seven different WSEs. Our results showed that this evidence can be successfully used for finding experts within an organisation. Moreover, by combining it with the existing intranet-based evidence, we showed that the expert search effectiveness can be further improved on the 2007 topic set. On the apparently more difficult 2008 topic set, although less evidence could be mined from the Web, we showed that marked improvements could still be attained, hence demonstrating the consistency of our approach.

Finally, we investigated the effectiveness of using the titles and snippets rather than downloading the full content of the results retrieved by WSEs. Our results showed that mimicking WSEs' rankings using this alternative evidence is unexpectedly more effective for expert search than using the corresponding full documents, while providing a drastic reduction in the incurred network costs. This demonstrates that our approach can effectively and efficiently be deployed in a typical enterprise setting.

Overall, our mimicking experiments show that, should a WSE start offering expert search services for enterprise organisations like CSIRO, they would likely be effective. Moreover, they confirm that exploiting the WSEs' underlying ranking mechanism through mimicking can be an effective approach to estimate the relevance of expertise evidence gleaned from outside the enterprise sphere. As future work, we wish to investigate the impact of different voting techniques (Macdonald & Ounis, 2006) on the effectiveness of using this external evidence, as well as different training procedures for further enhancing the mimicking of WSEs' document rankings, including the simulation of such rankings (Macdonald & Ounis, 2009). Finally, another direction of investigation is the study of combining multiple sources of expertise evidence (Fang, Si, & Mathur, 2010a).

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009

References

Amati, G., 2003. *Probabilistic models for information retrieval based on divergence from randomness*. Ph.D. thesis, University of Glasgow.

Amati, G., 2006. Frequentist and bayesian approach to information retrieval. In *Proceedings of the 28th European conference on information retrieval* (pp. 13–24). London, UK.

Aslam, J. A., & Montague, M. (2001). Models for metasearch. In *Proceedings of the 24th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 276–284). New Orleans, LO, USA: ACM.

Bailey, P., Craswell, N., de Vries, A. P., Soboroff, I. (2007). Overview of the TREC-2007 enterprise track. In *Proceedings of the 16th text retrieval conference*. Gaithersburg, MD, USA.

Balog, K., Azzopardi, L., & de Rijke, M. (2006). Formal models for expert finding in enterprise corpora. In *Proceedings of 29th annual international ACM SIGIR conference on research and development in information retrieval* (pp. 43–50). Seattle, WA, USA: ACM.

Balog, K., de Rijke, M. (2008). Combining candidate and document models for expert search. In *Proceedings of the 17th text retrieval conference*. Gaithersburg, MD, USA.

Balog, K., Soboroff, I., Thomas, P., Bailey, P., Craswell, N., de Vries, A. P. (2008). Overview of the TREC-2008 enterprise track. In *Proceedings of the 17th text retrieval conference*. Gaithersburg, MD, USA.

Carbonell, J., & Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in informaion retrieval* (pp. 335–336). Melbourne, Australia: ACM.

Craswell, N., de Vries, A. P., Soboroff, I. (2005). Overview of the TREC-2005 enterprise track. In *Proceedings of the 14th text retrieval conference*.

Fagin, R., Kumar, R., McCurley, K. S., Novak, J., Sivakumar, D., Tomlin, J. A., Williamson, D. P. (2003). Searching the workplace Web. In *Proceedings of the 12th international world wide Web conference* (pp. 366–375).

Fang, H., & Zhai, C. (2007). Probabilistic models for expert finding. In *Proceedings of the 29th European conference on information retrieval* (pp. 418–430). Rome, Italy: Springer-Verlag.

Fang, Y., Si, L., & Mathur, A. (2010a). Discriminative probabilistic models for expert search in heterogeneous information sources. *Information Retrieval*.

Fang, Y., Si, L., & Mathur, A. P. (2010b). Discriminative models of integrating document evidence and document-candidate associations for expert search. In *Proceedings of the 33rd annual international ACM SIGIR conference on research and development in information retrieval* (pp. 683–690). Geneva, Switzerland: ACM.

Hawking, D. (2004). Challenges in enterprise search. In *Proceedings of the 15th Australasian database conference* (pp. 15–24).

He, B., Macdonald, C., Ounis, I., Peng, J., Santos, R. L. T. (2008). University of Glasgow at TREC-2008: experiments in blog, enterprise, and relevance feedback tracks with terrier. In *Proceedings of the 17th text retrieval conference*. Gaithersburg, MD, USA.

Hiemstra, D. (2001). *Using language models for information retrieval*. Ph.D. thesis, University of Twente.

Järvelin, K., & Kekäläinen, J. (2002). Cumulated gain-based evaluation of IR techniques. *TOIS, 20*(4), 422–446.

Jiang, J., Han, S., Lu, W. (2008). Expertise retrieval using search engine results. In *Proceedings of the SIGIR 2008 workshop on future challenges in expertise retrieval* (pp. 11–16). Singapore, Singapore.

Joachims, T., Li, H., Liu, T.-Y., & Zhai, C. (2007). Learning to rank for information retrieval. *SIGIR Forum, 41*(2), 58–62.

Kendall, M. G. (1938). A new measure of rank correlation. *Biometrika, 30*(1/2), 81–93.

Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science, 220*(4598), 671–680.

Kwok, K. L., Chan, M. (1998). Improving two-stage ad-hoc retrieval for short queries. In *Proceedings of the 21st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 250–256). Melbourne, Australia.

Macdonald, C., 2009. *The voting model for people search*. Ph.D. thesis, University of Glasgow.

Macdonald, C., Hannah, D., Ounis, I. (2008). High quality expertise evidence for expert search. In *Proceedings of the 30th European conference on information retrieval* (pp. 283–295). Glasgow, UK.

Macdonald, C., He, B., Plachouras, V., Ounis, I. (2005). University of Glasgow at TREC-2005: Experiments in terabyte and enterprise tracks with terrier. In *Proceedings of the 14th text retrieval conference*.

Macdonald, C., Ounis, I. (2006). Voting for candidates: Adapting data fusion techniques for an expert search task. In *Proceedings of the 15th international conference on information and knowledge management* (pp. 387–396).

Macdonald, C., & Ounis, I. (2008). Voting techniques for expert search. *Knowledge and Information Systems, 16*, 259–280.

Macdonald, C., & Ounis, I. (2009). The influence of the document ranking in expert search. In *Proceedings of the 18th international conference on information and knowledge management* (pp. 1983–1986). Hong Kong, China: ACM.

Mukherjee, R., & Mao, J. (2004). Enterprise search: Tough stuff. *ACM Queue, 2*(2), 36–46.

Ounis, I., Amati, G., Plachouras, V., He, B., Macdonald, C., Lioma, C. (2006). Terrier: A high performance and scalable information retrieval platform. In *Proceedings of the ACM SIGIR workshop on open source information retrieval*.

Robertson, S. E., Walker, S., Hancock-Beaulieu, M., Gatford, M., Payne, A. (1995). Okapi at TREC-4. In *Proceedings of the 4th text retrieval conference. Gaithersburg, MD, USA*.

Serdyukov, P., Aly, R., Hiemstra, D. (2008). University of Twente at the TREC-2008 enterprise track: Using the global Web as an expertise evidence source. In *Proceedings of the 17th text retrieval conference. Gaithersburg, MD, USA*.

Serdyukov, P., Hiemstra, D. (2008a). Being omnipresent to be almighty: the importance of global Web evidence for organizational expert finding. In *Proceedings of the SIGIR 2008 workshop on future challenges in expertise retrieval* (pp. 17–24). Singapore, Singapore.

Serdyukov, P., Hiemstra, D. (2008b). Modeling documents as mixtures of persons for expert finding. In *Proceedings of the 30th European conference on information retrieval* (pp. 309–320). Glasgow, UK.

Soboroff, I., de Vries, A. P., Craswell, N., 2006. Overview of the TREC-2006 enterprise track. In *Proceedings of the 15th text retrieval conference*.

Wackerly, D. D., Mendenhall, W., & Scheaffer, R. L. (2002). Mathematical statistics with applications. *Duxbury Advanced Series*.

Yilmaz, E., Aslam, J. A., Robertson, S. (2008). A new rank correlation coefficient for information retrieval. In *Proceedings of the 31st annual international ACM SIGIR conference on research and development in information retrieval* (pp. 587–594). Singapore, Singapore.

Yimam-seid, D., & Kobsa, A. (2003). Expert finding systems for organizations: Problem and domain analysis and the DEMOIR approach. *Journal of Organizational Computing and Electronic Commerce, 13*(1), 1–24.

Please cite this article in press as: Santos, R. L. T., et al. Mimicking Web search engines for expert search. *Information Processing and Management* (2011), doi:10.1016/j.ipm.2010.11.009