FINDING EXPERTS ON THE WEB

Fabian Kaiser, Holger Schwarz and Mihály Jakob
Institute of Parallel and Distributed Systems, University of Stuttgart
Universitaetsstrasse 38, 70569 Stuttgart, Germany

Keywords: Expert Finder, Search Engine, Information Retrieval, Knowledge Management.

Abstract: In this paper, we present an integrated approach on finding experts for arbitrary user defined topics on the World Wide Web. We discuss the special challenges that come along with this issue and why solely applying standard techniques and standard tools like Web search engines is not suitable. We point out the necessity for a dedicated expert search engine, based on a Focused Crawler. The main contribution of our work is an approach to integrate standard Web search engines into the process of searching for experts to utilize the search engines' knowledge about content and structure of the Web.

1 INTRODUCTION

Knowledge is becoming more and more a crucial factor in today's economical environment. The knowledge on how technologies and markets will evolve or how the political and social environment will change, holds enormous potential for economic success. On the other hand, the lack of such knowledge bears uncertainties and expert knowledge is required to clarify these. While in larger companies such knowledge is often found in-house, smaller companies regularly lack this knowledge due to missing human resources. Therefore, they need to rely on external experts. The challenge on hiring such external experts is to find persons with adequate skills and knowledge. Various sources like articles, books or personal contacts can be utilized for this purpose. However, in this paper we focus on finding experts on the WWW as it forms a huge and steadily growing information base. While existing approaches are either limited to enterprise-internal data or require advanced search skills from the user, our solution is more general. We propose the Expert Search Engine *EXPOSE* which guides the user through the whole search process. An example-based search field specification, autonomous Web crawling and content analysis as well as the integration of standard search engines are the key concepts on which our approach builds.

The paper is organized as follows: In Section 2 we point out the main problems people have to face when searching for experts on the Web. Section 3 outlines existing approaches that deal with either expert search or websearch in general and shows their shortcomings when being applied to a webwide expert search. In Section 4 we describe a set of techniques that are valuable for finding experts on the Web. We show how we developed and integrated these techniques and what makes our approach superior to others. Finally, Section 5 gives a short conclusion.

2 PROBLEMS AND MOTIVATION

In our experiments, we found out that searching for experts on the Web cannot be efficiently accomplished by solely using standard Web search engines and that there is a need for support by dedicated methods and tools. We identified several problems that can be summarized as the lack of support for an automated search process.

First, there is no integrated tool that efficiently allows to search for experts on the Web. Thus, any such search will result in lots of data that has to be handled manually by the user. Secondly, due to the lack of obliging standards, most platforms and services on the Web follow their own rules on processing and publishing content, optimized for the interaction with human users. Only in a few niche applications like in the context of B2B, approaches propagated by the Semantic-Web-Community (Koivunen

Kaiser F., Schwarz H. and Jakob M. (2006).
FINDING EXPERTS ON THE WEB.
In *Proceedings of WEBIST 2006 - Second International Conference on Web Information Systems and Technologies - Internet Technology / Web Interface and Applications*, pages 363-368
DOI: 10.5220/0001252703630368

& Miller 2001) are applied to provide better support for automatic processing of content. A broad use of metadata as postulated by the Semantic Web Community would result in better results in classifying and rating the relevance of arbitrary text resources. In particular, questions like *"find authors of publications about fuel cells"* could then be answered by relative simple search engines. Valuable results would be authors of scientific publications, authors of articles that are available online, writers of posts to topic-related mailinglists or newsgroups, authors in bulletin boards and the like. Due to the heterogeneity of these diverse sources and a lack of available metadata, each of them requires a different approach for analyzing the relevance and extracting author information. However, this is not covered by standard search tools.

While the Web and thus the amount of available information steadily grows, a similar increase in search skills of standard search engine users cannot be observed. Jansen, Spink & Pedersen (2005) found, that most queries to search engines consist of maximum three keywords, only the top ten results are used in most cases and only few users refine a query. This often results in an inadequate query specification. Even if search engines could potentially provide the desired information, the user only gets what he asks for - and with an insufficient query, this might differ very much from what he wants to get. As users of standard search engines will probably not train their search skills to a required amount and hiring a dedicated searcher is most often not an option as well, this lack of skills needs to be compensated by software support. Apparently, solely with an exact problem specification, no experts can be found, but as the whole search process builds on it, such a detailed specification is a fundamental requirement. As this pre-condition is not met for most users, search engines are not of much help for them in the complex process of searching for experts.

3 RELATED WORK

Finding experts is by no means a new research issue. A valuable overview about various research work in this area is given by McDonald & Ackerman (2000) and Yimam & Kobsa (2002). Most of the work is done in the context of Knowledge Management and focuses on enterprise-internal search for experts like in Yellow Pages systems. In the following, we depict several existing approaches and tools that deal with expert identification or websearch in general and show their shortcomings in the context of a webwide expert search.

The *Xpertfinder* (Heeren & Sihn 2002) is a tool that monitors email traffic as well as several kinds of files like PDF- or Word-documents located on servers within the enterprise. It analyzes authorship, content and communication structures from these sources and derives a mapping from persons to predefined expertise fields. Thus it simplifies the creating of Yellow Pages Systems and makes their maintenance less extensive. The main reason why it can not be applied to a webwide expert search is that it compiles the search results from a relatively small set of resources that need to be indexed and analyzed previously. This, in turn, is almost impossible for a private search engine in the scope of the whole WWW, as the number of resources to analyze would be ways too big, resulting in too much effort. Furthermore, the main focus of *Xpertfinder* is analyzing email exchange. While this is suitable for an enterprise-internal network, it can not be achieved for email in general, as only such communication can be monitored that involves email servers under the control of the search tool. Thus, the results are limited to enterprise-internal experts which is not suitable, as pointed out in Section 1.

A different aspect of expert identification is targeted by research in the area of *virtual communities*. Here, techniques and tools are being developed for measuring the degree of interaction between participants in a discussion. This is not primarily used to identify experts, but to monitor discussions and to gain information about the evolution of such discussion groups. Still, such techniques can compile valuable information about who belongs to the most active persons in a discussion and which knowledge networks are formed by the interconnection of users. The *Management Cockpit* (Trier 2005) is such a tool for the visualization of knowledge networks. Based on wrappers for several bulletin boards, the communication structure of such boards can be analyzed and visualized. For our goal of finding domain specific experts on the Web, this approach is too isolated because the user has to manually identify a board prior to the analysis and potentially implement a wrapper to make the data readable. Furthermore, tools like the *Management Cockpit* just analyze and visualize the communication structure of a whole board but do not differentiate on the content of every single thread and post. Their level of detail is too low and thus results are not of much use for expert identification.

Several commercial platforms for finding experts on the Web are available as well. Like on *http://www.bizwiz.com/* users can register as experts in several domains. These platforms, however, suffer from the fact that people have to manually register and describe their skills themselves. This brings the disadvantage that mainly people with commercial interests will register and that they will tend to overrate their skills and expertise to get into contact with more potential clients. Thus, the scope of such platforms is rather limited.

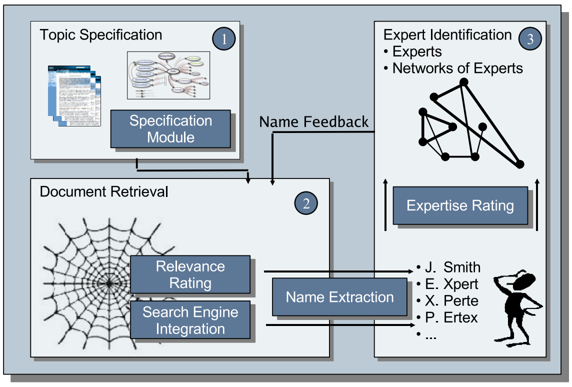

Figure 1: EXPOSE- Expert Search Engine.

4 EXPOSE: AN INTEGRATED APPROACH TO FINDING EXPERTS

As shown in Section 3, several problems occur when using existing tools in the Web context or even make it impossible to use them. To our knowledge, there is neither a tool nor a technique that meets our requirements in identifying arbitrary experts for arbitrary fields on the Web. Apparently, several techniques that are used by the systems described above are also valuable in our scenario. We therefore present an approach that integrates several known techniques from the information retrieval domain. Especially we discuss the role standard search engines can play and how they can be used to increase quantity and quality of search results. The conclusion of these considerations is our Expert Search Engine *EXPOSE*.

We identified three steps in the process of searching experts, as indicated in Figure 1:

1. Specify the information need.
2. Search for related documents.
3. Identify the experts mentioned in these documents.

Our approach follows these three steps and will be explained in the next sections.

4.1 Topic Specification

The first step towards finding experts is specifying the topic, experts are searched for. This is a common problem that most Web users face day by day when running keyword queries on Web search engines. The problem consists of mainly two aspects: (a) Different resource authors use different vocabularies to phrase content. (b) As pointed out by Jansen, Spink & Pedersen (2005), most users only specify their information need by providing just very few keywords. This is not sufficient for searching information on highly sophisticated topics, as the search results are often simply too numerous or too general. We chose a different approach for specifying an information need which is more intuitive for humans: instead of compiling a set of keywords that describe a topic, it is easier to characterize a topic using examples, which in this case can be any topic-related documents. The search process then focuses on finding further documents that are similar to the ones provided as samples. However, these samples also need to be provided by the user which results in a bootstrapping problem. Yet, we found it less extensive for the user to search for only some two or three topic-related documents using standard search engines and then let an automated system find similar resources. However, two further questions arise from this approach:

1. How to define the similarity of documents?
2. The input to standard search engines are mainly keywords instead of sample documents. Thus support for the latter technique is very rudimental if existing at all. How can this problem be solved?

Question (1) is a well known problem in the field of information retrieval. Like in many other approaches we transform documents into the vector space model and calculate their similarity based on the cosine similarity of the representing vectors. Question (2) on the other hand is more complicated. As standard search engines in general cannot be used with this example based technique, all their knowledge about content and structure of the Web cannot be utilized either. However, running a private fully-featured search engine is not an option for standard users. Thus a compromise between minimum effort and utilizing maximum knowledge must be found. In Section 4.2 and 4.3 we present our approach to tackle this problem.

4.2 Document Retrieval

As pointed out in Section 4.1, standard search engines are not suitable for searching and retrieving documents based on specification by examples. Still, for searching documents, Web resources need to be analyzed for relevance. To gather such resources, our expert search engine also contains a crawler that downloads such resources and then feeds it into a classifier (Figure 1). While for search engine crawlers, the document sequence in a crawl is not that important, as for most documents it does not make a difference if it is crawled now or any later, the crawler to be used here indeed has to focus as much as possible on topic-related documents. Therefore, it should crawl documents that are likely to be topic-related before those that are less likely to be relevant.

To solve this problem, Chakrabarti, van der Berg & Dom (1999) proposed an approach they called *Fo-*

cused Crawling. A Focused Crawler has functionality similar to that of standard search engine crawlers. Starting from a set of initial URLs, it downloads the specified resources, extracts the URLs these resources contain and recursively crawls them. The main difference between a Focused Crawler and crawlers of standard search engines is, that it aims to only download resources that are related to a specific topic, that is, the crawler focuses on this topic. Following the idea of specifying the search field with sample documents, the Focused Crawler starts from this sample set. Each document it crawls is then compared to the set of sample documents and if it is similar enough, the links found in this document are followed as well. See e.g. Diligenti at al. (2000) for some heuristics to improve the recall when using this technique.

The appliance of a Focused Crawler significantly differs from using standard search engines. While a search engine delivers the answer to keyword queries within seconds, a Focused Crawler may run for minutes to hours or even days. While this is a shortcoming in terms of getting immediate results, it can offer the user a way to influence the search progress by redirecting the crawler or refining the specification. Furthermore, the knowledge of search engines about the Web can be utilized to support a Focused Crawler. This will be discussed in the following.

4.3 Integrating Web Search Engines

As depicted in the preceding section, a Focused Crawler can reduce the number of unnecessary downloads. Still, a major challenge is that many topic-related resources are not well connected, neither directly nor by reasonably short click-paths (Diligenti et al. 2000). Thus, many relevant resources can not be crawled because a Focused Crawler will not find them. Especially for the context of finding experts, we developed a technique to tackle this problem of Focused Crawlers and find more and better suited experts on the topic in question. Our approach bases on the integration of the above mentioned techniques and the use of a Focused Crawler in combination with standard search engines. In the following, we will describe this integration.

4.3.1 Backlinks

Crawling websites starts from a set of documents, extracting the links they contain and following these links recursively. Thereby the Web can be seen as a graph where documents are represented by nodes and links are directed edges between two such nodes. The edge property *directed* implies, that even if from document A (see Figure 2) another document B can be reached following edge E_{ab}, there is not necessarily a way back from B to A, that is, E_{ba} does not necessarily exist. However, for finding topic-related documents, getting document A when only knowing document B would be a great benefit as then the Web could be seen as a non-directed graph which would make crawling much easier.

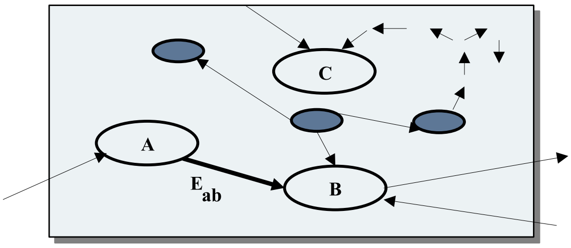

Figure 2: Interconnection of Web resources.

We tackle this problem by utilizing standard search engines (Diligenti et al. 2000). Besides the document content, such search engines also store information about the link structure, as this structure is used for ranking the results of a user's query. Some search engines make this information available to the user by providing special keywords in the query string. While without this technique we are only guaranteed to find documents that are linked from the source document, we now can also find arbitrary pages that link to the document in question. This is extremely helpful especially for popular sites, that are referred to by many other topic-related sites.

4.3.2 Search for Similar Pages

To find documents that are not connected via short link paths (document *C* in Figure 2) or not connected at all, more information about structure and content of those unreachable resources is required. To some extent, this can be achieved by utilizing standard search engines, as some of them offer keywords to indicate a search for documents similar to a given one. Such a query returns the resources that show a minimum difference to the source document, independent of their interconnection with this source.

While we can not influence how these search engines calculate the similarity of documents and which documents they return, there is another way of finding resources similar to a given one, again utilizing search engines. Standard search engines are normally interfaced using keyword queries while our Focused Crawler does not use keywords but relies on measuring the similarity of two or more documents. Thus, to utilize search engines, we need to extract keywords from documents that are rated as highly relevant, to feed these keywords into the search engine.

However, the difficulty with this approach is the extraction of keywords from arbitrary text. A closer look to the words of a document shows that each word contributes differently to the semantics of the whole text. Some words, the keywords, more or less specify the topic whereas others are just fillers, like the stopwords “as”, “so”, “for” etc. They can be filtered as they do not contribute to the semantics of the text. To find out, which of the remaining words are keywords, we calculate the $TF * IDF$ vector (Salton & McGill 1983) for a document. Our calculation is based on a text corpus consisting of several million international websites. The terms with the highest $TF * IDF$ values are then fed into a standard search engine and the resulting resources are analyzed for relevance.

Apparently, the resulting documents are not necessarily relevant, as this approach suffers from the very same problems like users of a search engine do: there is an increasing percentage of Web-spam that is not related to the topic in question but just contains the keywords (Henzinger, Motwani & Silverstein 2002). Furthermore, we can not be sure having extracted the right keywords, that is, any automated technique for extracting keywords from a text will be error-prone to some extent. Thus, in the worst case, the search engine returns a list of resources that might not be relevant. This is the same for any link the Focused Crawler follows and thus, there is no disadvantage in applying this technique besides a slightly increased number of downloads, as already the Focused Crawler can discard irrelevant documents. But the benefit is that if there are relevant documents among the search engine results, we are able to find them even if they are not well connected to already crawled resources.

4.4 Expert Identification

Once relevant documents have been identified, the third phase (Figure 1) can be entered: extracting expert information from these documents.

4.4.1 Name Extraction from Text

To identify all names in a text, advanced techniques would be needed because the text to be analyzed usually is natural language. As natural language follows complex grammars and is highly ambiguous, full understanding of such a text cannot be achieved in general. A full understanding, on the other hand, would be required to reliably identify persons in a text. However, Palmer & Day (1997) as well as Mikheev, Moens & Grover (1999) showed, that two simplifications reduce the effort in both implementation and runtime while still producing good results:

- Searching for known names, based on a name database. Lots of names can be identified that way if some simple points are considered: often there are abbreviations like “J. Smith” or the order of surname and forename is inverted (“Smith, John”).
- Searching for phrases indicating that a name is mentioned close to this phrase. Some examples are “according to J. Smith”, “Mr. Smith”, “Smith says” etc. This way, also a single forename or a single surname can be identified that would otherwise be ignored because many names are also used as terms in different contexts. For example “April” is a common forename as well as the name of a month.

Using these techniques in EXPOSE already led to quite good results. As we wanted to show the feasibility of our approach, we did not yet focus on optimizing the name recognition. However, in future work, we will use more advanced techniques, e.g. from the *Named Entity Recognition* domain.

All names identified that way form the input for the next step. We identified a set of four roles in which a person is named in a text: (1) the person is the author of the text, (2) there is a discussion the person is (actively or passively) involved in, (3) the person is referred to, (4) the person is mentioned although he/she is not related to the topic at all. In case 1-3, the assumption that the named person is an expert is at least potentially right. In the latter case however, this assumption is likely to be wrong. This shows, simply from the occurrence of names in a single topic-related text, no experts can be identified. In this next step we therefore have to find out, which of the named persons are really experts on the domain in question. Two problems have to be tackled then: (1) The occurrence of a name is not recognized although a person has been named in the text. (2) One or more terms in the text are assumed to name a person while in fact they do not (e.g. “*. . . for fuel cell manufacturers in the U.S. Smith denotes that . . .*” may produce *U.S. Smith* while *U.S.* is the last term in sentence *A* and refers to the *United States of America* whereas sentence *B* starts with some text referring to a person *Smith*, but not *U.S. Smith*). While (1) can be attended by increasing the name database or the associative rules, (2) is more complex and will be discussed in the following.

4.4.2 Expertise Rating

As the extraction of expertise from a single text requires a good understanding of the text, which in general we do not have, our approach bases mostly on statistical properties that are extracted from a set of relevant documents. We identified four criteria from which we derive the expertise of a person. Therefore, we evaluate each of these criteria and compile an overall rating from the singular results by normalizing and summing up the results from each rating.

The first quantity is simply how often a person is named in any relevant document. The more often a person occurs in texts related to the topic, the more likely this person has expertise on this topic.

Only relying on the sole presence of names in a text is error-prone: Imagine a conference website that lists the authors and abstracts of all accepted papers. If some of the papers where relevant to the topic in question, then not only the authors of the relevant paper, but also the authors of all the other papers would be rated experts. Therefore, also the position of the naming and the structure of the text (Song et al. 2004) has to be taken into account. Simple heuristics for HTML sites are (1) apply better ratings to persons that are named close to relevant keywords. (2) apply better ratings to persons that are named in the center of a text instead of near the margins. In contrast, names often occur within navigation bars, as a reference to the webmaster or in similar places. They name persons who are likely to be of no interest for the expert search. Thus, persons found in a context like "In *J. Smith's* talk about the growing demand for *high-energy fuel cells…*" would be rated higher than persons not related to the topic but named somewhere else in the document.

A third approach to rate expertise is based on analyzing communication structures. The idea is that a person is likely to have expertise when taking part in a discussion with other persons. Such discussions can be monitored in bulletin boards or in the archives of mailing lists. The more a person is in contact with other persons, especially with other persons rated as experts, the better the rating will be for this person. This is similar to the PageRank approach (Brin & Page 1998) for rating the relevance of websites.

Lastly, the interconnection of resources that name potential experts can also be taken into account. Resources that are highly connected with each other by hyperlinks are likely to form some kind of knowledge cluster. Thus, persons that are named frequently on such highly connected resources are likely to be involved into the topic and thus to have some expertise. That is, from the structure of a resource-graph, we derive hints for the expertise of persons that are named on these resources.

5 SUMMARY

In this paper, we proposed a three-step-approach to finding experts on the Web: First specifying the information need, secondly searching for related Web resources and finally identifying experts on these resources. This process can be efficiently supported by our expert search engine *EXPOSE*. We pointed out how to integrate techniques from the information retrieval domain with methods that originate from social network analysis and especially how to utilize Web search engines' knowledge about content and structure of the Web.

ACKNOWLEDGMENTS

This work was conducted in the context of the project *nova-net*, which is funded by the German Federal Ministry of Education and Research (BMBF).

REFERENCES

Brin, S. & Page, L. (1998). The Anatomy of a Large-Scale Hypertextual Web Search Engine. In *Proc. of the 7th International World Wide Web Conference*.

Chakrabarti, S., van der Berg, M. & Dom, B. (1999). Focused crawling: a new approach to topic-specific web resource discovery. In *Proc. of the 8th International World Wide Web Conference*.

Diligenti, M., Coetzee, F., Lawrence, S., Giles, C. & Gori, M. (2000). Focused Crawling using Context Graphs. In *Proc. of the 26th VLDB*.

Heeren, F. & Sihn, W. (2002). Xpertfinder - message analysis for the recommendation of contact persons within defined topics. In *Proc. of the 6th IEEE AFRICON*.

Henzinger, M., Motwani, R. & Silverstein, C. (2002). Challenges in Web Search Engines. In *SIGIR Forum Vol. 36, No 2*.

Jansen, B., Spink, A. & Pedersen, J. (2005) A temporal comparison of AltaVista Web searching. In *Journal of the American Society for Information Science and Technology, Vol. 56, Issue 6*.

Koivunen, M. &Miller, E. (2001). W3C Semantic Web Activity. In *Proc. of the Semantic Web Kick-off Seminar in Finland*.

McDonald, D. & Ackerman, M. (2000). Expertise Recommender: A Flexible Recommendation System and Architecture. In *Proc. of the 2000 ACM Conference on Computer Supported Cooperative Work*.

Mikheev, A., Moens, M. & Grover, C. (1999). Named entity recognition without gazetteers. In *Proc. of EACL 1999*.

Palmer, D. & Day, D. (1997). A Statistical Profile of the Named Entity Task. In *Proc. of the 5th Conference on Applied Natural Language Processing*.

Salton, G. & McGill, M. (1983). *Introduction To Modern Information Retrieval*. McGraw-Hill, Singapore.

Song, R., Liu, H., Wen, J. & Ma, W. (2004). Learning Block Importance Models for Web Pages. In *Proc. of the 13th World Wide Web Conference*.

Trier, M. (2005). A Tool for IT-supported Visualization and Analysis of Virtual Communication Networks in Knowledge Communities. In *Proc. of the Wirtschaftsinformatik 2005*.

Yimam, D. & Kobsa, A. (2002). Expert Finding Systems for Organizations: Problem and Domain Analysis and the DEMOIR Approach. In *Beyond Knowledge Management: Sharing Expertise*. M. Ackerman, A. Cohen, V. Pipek and V. Wulf.