Exploiting a Company's Knowledge: The Adaptive Search Agent YASE

Alex Kohn, François Bry
(Institute of Informatics, University of Munich, Germany)

Alexander Manta
(Roche Diagnostics GmbH, Penzberg, Germany)

Abstract: This paper introduces YASE, a domain-aware search agent with learning capabilities. Initially built for the research community of Roche Penzberg, YASE proved to be superior to standard search engines in the company environment due to the introduction of some simple principles: personalized ranking based on a user's role and organizational embedding, automatic classification of documents by using domain knowledge and learning from search history. While the benefits of the learning feature need more time to be fully realized, the other two principles have proved to be surprisingly powerful.

Keywords: search agent / engine, metadata, personalization, information retrieval, YASE
Categories: H.3.3, H.4.0, L.2.2

1 Introduction

All of us, regardless of our domain, field or specialty, face the same problem of information overload. In this paper we describe some promising approaches to find the relevant information in steadily growing information flows. The concepts are examined in the context of a research department of Roche Diagnostics GmbH.

In the following sub-sections we describe the hypothesis drawn from two complementary analyses performed last year of the way scientists access information. Section two introduces the search agent[1] YASE which incorporates original ideas of how to improve and personalize the ranking of results. In the last section we conclude the paper and show some perspectives.

1.1 1st hypothesis: One single entry point is what scientists prefer

In order to understand how Roche scientists retrieve information, two complementary in-house studies have been conducted: a survey and a log file analysis. The survey [Mühlbacher, 08] was based on personal questionnaires, targeting approx. 90 scientists from R&D. The second study was a log file analysis based on the monitoring of the usage of a subset of the information sources. During a period of one month we monitored 5 different search engines targeting approx. 400 employees from research and measured their relative usage (cp. Table 1).

[1] The term search *agent* is used in this context to distinguish our approach from standard search engines.

Search Engine	Relative access
Google (Internet)	80,8 %
Wikipedia	8,9 %
PubMed (biomedical abstracts DB)	5,6 %
FAST (intranet search engine)	3,8 %
Google Search Appliance (in-house file search)	0,9 %

Table 1: Usage of search engines linked from a Pharma Research homepage.

Both analyses show that a small minority of resources are heavily used by almost all scientists, while the majority of resources are barely accessed. Interestingly, because of the familiarity with the interface and due to its search performance even in specialized data sources like PubMed, patents and Wikipedia, scientists use Google more and more as the main entry point, even for scientific information. With Google there is no need to start an extra search at e.g. Wikipedia or PubMed. This suggests that – similarly to Google for external information - one single entry-point to internal resources would dramatically increase the use of the specialized but valuable data repositories of the company.

1.2 2nd hypothesis: Standard search engines less used because of poor ranking

A closer look at the usage analyses shown in the Table 1 suggests that valuable sources of in-house information and knowledge (those covered by Fast 1 and Google Search Appliance) are not accessed via search but rather by navigating the folder tree.

Enterprise search engines usually use the vector-space-model for results ranking. Algorithms successful in the Internet like PageRank show bad performance because the linkage structure of the Intranet is either poor [Fagin, 03], [Xue, 03] or completely missing as is the case with most document repositories. Besides, high redundancy (many versions of the same document) and notational heterogeneity (synonyms) distort the search results. Complex queries which go beyond the simple full text search can't be carried out with standard search engines. While cross products or joins are almost impossible to compute on the Internet, this would be possible in an intranet environment as this is comparatively much smaller.

1.3 3rd hypothesis: Dynamically built navigational structures can compensate for the missing linkage structure

The wealth of meta data available in the company (distribution lists, departmental and project membership lists, domain related thesauri and ontologies, access lists and other meta data extracted from the file system) can be used to assess the relevance of the documents to a certain user, to cluster and classify the documents, to improve the ranking and to create ad-hoc navigational structures. The search history (tuples of search terms and clicked documents), combined with functionality for manual document annotation adds a learning dimension to the repository of metadata, with potential of continuous self-improvement. By adding adequate reasoning features an adaptive, context-aware search agent can be built, as demonstrated by the prototype YASE.

2 The adaptive search agent YASE

Some of the metadata existing in a company and exploited by YASE are given in the table below:

File system	**Document attributes and structure**	**Business context of the user**	**Domain related knowledge**
Size, path, time (creation, last modification, last access), security (read & write permissions, owner)	Author, title, subject, company, manager, width, height, resolution, text content, links, comments, ...	Contact details (name, e-mail, phone, office), department, involved projects and groups	Controlled vocabularies (gene names, protein names, project names), databases, applications, ...

Table 2: Sources of metadata.

These metadata have no well-defined semantics in any RDF formalism. Some sources can be even messy, e.g. the “title” attribute (file format metadata) which can contain values like “untitled” or “slide 1”. Sources like the controlled vocabularies on the other hand can be considered clean and curated. Regardless of the source, YASE will treat all accessible data as metadata, whether it is correct or not.

Using metadata annotators of different types (statistical, machine learning or knowledge based) these sources can be used to associate appropriate attributes to documents. As an example consider the annotation of project relevance to a certain document. Project names from a controlled vocabulary are matched against the lexical substrings of the path and against the document vocabulary using a fuzzy string matching approach. In just the same manner we assign departmental relevance to documents. By using the annotator we basically put a file in several categories. At query time, facets according to the annotated categories are automatically displayed, by which a user can further browse through the data. In this way the navigational freedom to browse by project categories which otherwise are spread over several folders is enabled.

An even more powerful join is the association of document metadata with administrative user data, i.e. the user’s working context. We know the documents belonging to a project and we also know in which project a scientist is working. Hence, by joining both we know which project-related documents are relevant to a scientist. The true potential of this join is exploited when personalizing the ranking of results. After a user enters a query in YASE his administrative metadata is automatically retrieved and a temporary user profile reflecting his role in the company is created. A first ranked results list is obtained by the vector space model. In the next step the hit list is re-ranked according to the user’s profile. Documents lying closer to the user’s assumed interests are ranked higher than others. This is a key difference between YASE and search engines. Our ranking idea assumes that a user’s interests are reflected by his role and context embedding. However, this assumption does not always hold. Therefore we plan to allow a user to slip into different roles during a search session.

After having released YASE as a prototype for a significant part of the research community, we did a usage analysis based on log files over a period of three months. The results show a 39% usage of YASE (much more than Fast or Google Search Appliance from Table 1). This is an indication that YASE is accepted by the users and that it has a higher value compared to the other two internal search engines. It also suggests that at least some of the applied hypotheses are valid.

3 Conclusion and Perspectives

We have successfully used existing metadata which isn't exploited by standard search engines. Faceted navigation over the documents has been enabled and in addition the ranking of results was improved by applying a role-based adaptation. Exploiting existing metadata was the key to the success of YASE in its first prototype version.

Even though YASE is tailored to the specific environment of a research department, we argue that the concepts behind YASE allow its use in other intranet environments as well with only minor adjustments. First of all, the described shortcomings of standard search engines prevail in many other intranet environments as well. Further, domain metadata or administrative data, such as those described earlier, are available in every company or institution.

The learning features based on the search history and the inference capabilities using domain thesauri and ontologies, though partially implemented, are still to be investigated in depth. These anonymized data can be used for various purposes: recommendations of alternative queries or additional documents (URLs), improving ranking of results, etc. The integration of "deep knowledge" extracted from company databases with published documents will reveal further potentials of the adaptive search agent.

References

[Baeza-Yates, 99] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, Addison Wesley, 1999

[Bry, 03] F. Bry, P. Kröger, Bioinformatics Databases: State of the Art and Research Perspectives, In Proc. 7th ADBIS, 2003

[Fagin, 03] R. Fagin, R. Kumar, K. McCurley, J. Novak, D. Sivakumar, J. Tomlin, D. Williamson, Searching the Workplace Web, In Proc. 12th Int. WWW Conference, 2003

[Mühlbacher, 08] S. Mühlbacher, University of Regensburg / Roche Diagnostics GmbH Penzberg, Dissertation, Q4 2008, forthcoming

[Xue, 03] G. Xue, H. Zeng, Z. Chen, W. Ma, H. Zhang, C. Lu, Implicit Link Analysis for Small Web Search, SIGIR'03, 2003