Exploiting a Company's Knowledge: Your Adaptive Search Agent

Alex Kohn[12], François Bry[1], Alexander Manta[2]
[1]Institute for Informatics, University of Munich, Germany; [2]Roche Diagnostics GmbH, Penzberg, Germany

1 Introduction

Popular tools for searching the Internet like *Google* and *Yahoo!* give access to billions of web pages and to deep web content like *PubMed* (medical abstracts DB), *US Patents or Wikipedia*. Having experienced the performance of Internet search tools, company employees have similar expectations of the intranet search: large data coverage and excellent ranking of results. To their disappointment, they are confronted with a heterogeneous landscape and a multitude of search functions spread across different search engines, groupware and databases. The search process becomes fragmented, tedious and often inefficient. Enterprise search engines like *FAST, Verity* or *Google Search Appliance* are designed to fit the variety of technical environments typical for companies. Despite their technical flexibility and adaptability, they remain generic on the semantic level. Being agnostic with respect to roles and information needs of the employees, lacking even basic domain knowledge, they are not able to return differentiated and specific information, as expected by the intranet user when performing specific tasks.

YASA (Your Adaptive Search Agent) is a domain-aware complement to standard search engines. It builds on top of them and incorporates domain specific knowledge for improving search and navigation in specific usage contexts.

2 Hypothesis

One single entry point: By simplifying the user interface, one single entry point to the information (internal and external) would dramatically increase the use of specialized but valuable data repositories of the company.

Role-specific ranking: A search agent should have a ranking reflecting as good as possible the specific interests of the user's role within a company.

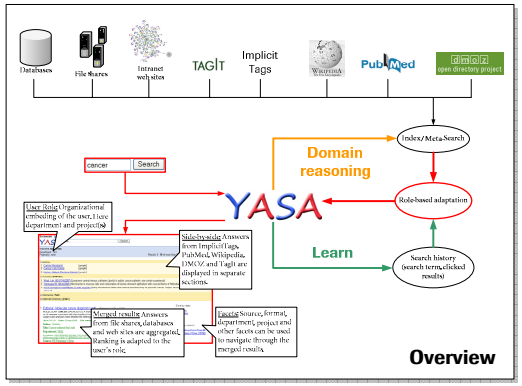
A company's knowledge is instrumental for search: The search precision and ranking of results can be dramatically improved by exploiting domain knowledge.

3 Sources of Metadata

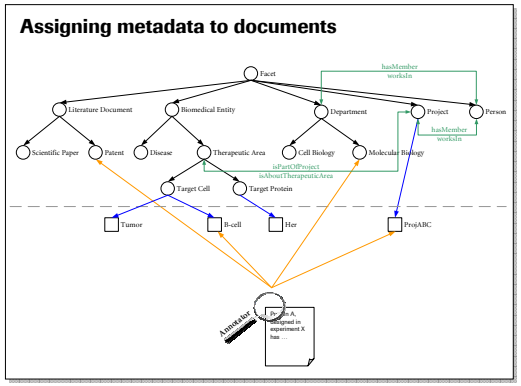
File system	Document attributes and structure	Business context of the user	Domain related knowledge
Size, path, timestamps (creation, last modification, last access), security (read & write permissions, owner)	Author, title, subject, company, manager, width, height, resolution, text content, links, comments, ...	Contact details (name, e-mail, phone, office), department, involved projects and groups, skills, position, …	Organisms, compounds, genes, targets, projects, applications, ...

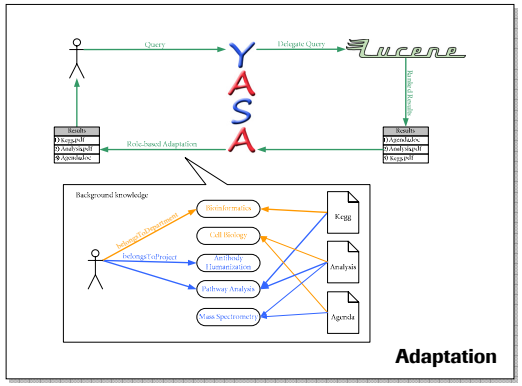
4 Architecture & Design

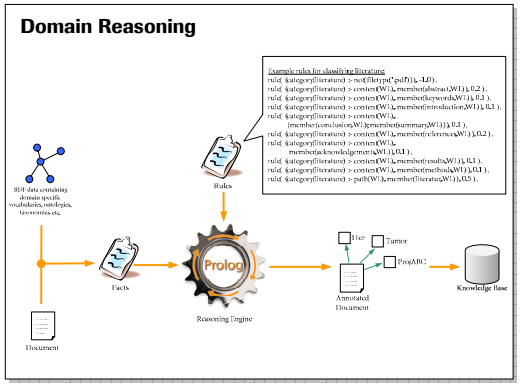
YASA takes the output of standard search engines and returns user and domain specific information due to features like:
1) Domain reasoning using domain ontologies and knowledge bases
2) Role-based adaptation
3) Learning from log data (log-based re-ranking, ImplicitTags)



The ability to offer faceted navigation and role-based adaptation requires to know the concepts a document is associated with. These associations can be detected by means of annotators which scan the document's known (meta-)data and decide based on configurable rules to which concept a given document belongs. Association between instances like "person worksIn project" can be used by the faceted navigation to give more options for browsing.





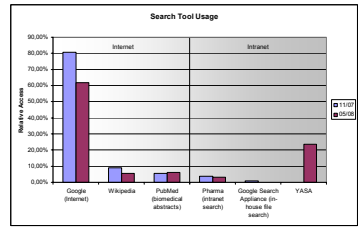


A query is delegated to the underlying search engine (here: *Lucene*) and the answer set is re-ranked considering in particular the user's role. It is known to which department a worker belongs and the projects he is involved in. Domain reasoning allows the mapping of different document categories to user roles. Our re-ranking algorithm uses background knowledge to rank items lying close to a user's role and interests higher than others.

Data extracted from the document index like content, title, subject, path, etc. are supplied to a rule engine (here: Prolog) together with domain specific information derived from ontologies and taxonomies. Based on rules with certainty factors, the reasoning engine classifies a document into categories.

5 Preliminary Results

To measure the effectiveness of *YASA* we have conducted two log file analyses. Each over a period of one month. The first was made in 2007 when *YASA* did not yet exist and the second after its pilot release in 2008. The results are comparable because both, *YASA* and the commercial search engines, were accessed from the same web page and up to the users' choice.



The statistics shows that *YASA* is the preferred tool on the intranet. However, further analyses are necessary to determine the reasons for the users' choice.

6 Conclusion

We introduced *YASA*, a domain-aware complement to standard search engines. *YASA* gives access to multiple resources – a first step towards a "one single entry point". Further, it applies "role-specific ranking" of results by means of adaptation. Finally, it exploits a company's knowledge, that is existing (meta-)data and log data to improve several aspects of search. The preliminary analysis doesn't point out which principles are of value and which not. However, domain specific knowledge, applied in almost all parts of *YASA*, is without a doubt a key to the success of *YASA* in its first prototype.

References

[Baeza-Yates, 99] R. Baeza-Yates, B. Ribeiro-Neto, Modern Information Retrieval, ACM Press, Addison Wesley, 1999
[Bry, 03] F. Bry, P. Kröger. A Computational Biology Database Digest: Data, Data Analysis, and Data Management. Distributed and Parallel Databases, 2003, Vol. 13, No. 1, pp. 7-42
[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, Doctoral Thesis, 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