Chapter 9

Semantic Search on Unstructured Data: Explicit Knowledge through Data Recycling

Alex Kohn
Roche Diagnostics GmbH, Germany

François Bry
University of Munich, Germany

Alexander Manta
Roche Diagnostics GmbH, Germany

ABSTRACT

Studies agree that searchers are often not satisfied with the performance of current enterprise search engines. As a consequence, more scientists worldwide are actively investigating new avenues for searching to improve retrieval performance. This paper contributes to YASA (Your Adaptive Search Agent), a fully implemented and thoroughly evaluated ontology-based information retrieval system for the enterprise. A salient particularity of YASA is that large parts of the ontology are automatically filled with facts by recycling and transforming existing data. YASA offers context-based personalization, faceted navigation, as well as semantic search capabilities. YASA has been deployed and evaluated in the pharmaceutical research department of Roche, Penzberg, and results show that already semantically simple ontologies suffice to considerably improve search performance.

INTRODUCTION

Nowadays most data produced in business is captured electronically and stored in computer systems. Search engines are of key importance in making this “hidden” information visible to the employees. Spoiled by the improvements in Web search, experts expect now a similar search performance in their intranet environment. However, current state-of-the-art enterprise search engines underperform (Feldman & Sherman, 2004). In effect, search for information becomes a central problem in companies.

DOI: 10.4018/978-1-4666-0185-7.ch009
A particularity of enterprise search is the lack of scientific publications. In case of commercial products, the information provided in booklets or white papers give only a vague picture of the applied algorithms. An aggravating factor is that the methods’ effectiveness in improving information retrieval in enterprise search is barely empirically investigated. Indeed, published methods often restrict to synthetic evaluations. Further, scientific publications often describe methods which are optimized for the Web but not for intranet environments. Lastly, papers addressing intranet search are often focused on the intranet web, ignoring the fact that file shares, e-mails, databases, applications, etc. are also part of an intranet which needs to be searched.

We conclude that search for information in intranet environments is theoretically and practically disappointing. The rising question is: why is search for information in the enterprise such a challenge?

Many reasons can be given (Fagin et al., 2003; Hawking, 2004): Heterogeneous data sources and formats, complex security permissions, less user observations, few or missing metadata, growing amounts of data, etc.

The World Wide Web is dominated by the hypertext protocol. This is in contrast to intranets, where only a small portion of the data is in a Web accessible format. This heterogeneity makes data integration a difficult task, as large portions of the intranet are not search engine friendly. Further, ranking of search results is made more difficult due to a different or missing linkage structure (Xue et al., 2003).

The complex security permissions present in companies are a mixed blessing. On the one hand side the information landscape is fragmented into many silos, i.e. any employee can only see a small subset of all data. On the other hand, ranking of search results is eased as only a subset of all data needs to be sorted by relevance. The degree of fragmentation depends of course on the company’s philosophy of information sharing across departments.

Observing a user’s search behavior enables search engines to detect the context of a user, which ultimately leads to personalization services (Micarelli et al., 2007). Such services are already part of the leading Web search engines. Offering personalization services in the enterprise however, is a difficult task due to the lack of feedback data: a few users are facing a lot of data.

Being confronted with barely explicit metadata at hand and mostly unstructured free-text documents represents another challenge. Therefore, it is difficult to offer semantic search capabilities—a problem, well known from the Internet.

Considering the mentioned challenges, the problem is how to improve search for information in the enterprise. Could integration, i.e. federated search, make the information landscape accessible? Could the ranking of search results be improved by applying facetted navigation or personalized search? Could high-quality metadata be obtained by applying automatic information extraction? Could domain knowledge (e.g., organizational charts, project databases, etc.) be used to set the searcher as well as the results in context?

Technically, we contribute by compiling and developing several approaches for facing the listed challenges, namely role-based adaptation, guided navigation, and incorporation of domain knowledge. The approaches are implemented into YASA (Your Adaptive Search Agent). YASA is deployed in the pharmaceutical research department of Roche in Penzberg.

Scientifically, we contribute by conducting an empirical evaluation of the applied methods in a real world setting. In particular, we analyze whether the ranking of search results can be improved by our role-based adaptation method. We investigate how useful facetted navigation is for finding information. Further, we analyze whether automatic metadata extraction can help improve information acquisition. Last, we determine whether search for information can be improved by including domain knowledge into the search process. These aspects have previously only insufficiently been investigated or not all.
Related Content

Named Data Networking: A Promising Architecture for the Internet of Things (IoT)
[www.igi-global.com/article/named-data-networking/203694?camid=4v1a](www.igi-global.com/article/named-data-networking/203694?camid=4v1a)

Technologies to Support the Market of Resources as an Infrastructure for Agile/Virtual Enterprise Integration
[www.igi-global.com/chapter/technologies-support-market-resources-infrastructure/35761?camid=4v1a](www.igi-global.com/chapter/technologies-support-market-resources-infrastructure/35761?camid=4v1a)

Enhancing the Access to Public Procurement Notices by Promoting Product Scheme Classifications to the Linked Open Data Initiative
Jose María Alvarez Rodríguez, Luis Polo Paredes, Emilio Rubiera Azcona, Alejandro Rodríguez González, José Emilio Labra Gayo and Patricia Ordoñez de Pablos (2013). *Cases on Open-Linked Data and Semantic Web Applications* (pp. 1-27).
[www.igi-global.com/chapter/enhancing-access-public-procurement-notices/77197?camid=4v1a](www.igi-global.com/chapter/enhancing-access-public-procurement-notices/77197?camid=4v1a)

Pillars of Ontology Treatment in the Medical Domain
[www.igi-global.com/chapter/pillars-ontology-treatment-medical-domain/38043?camid=4v1a](www.igi-global.com/chapter/pillars-ontology-treatment-medical-domain/38043?camid=4v1a)