Determining Relevant Product Information Sources

Matthias Wauer
Chair for Computer Networks, Institute for Systems Architecture
Technische Universität Dresden, Germany
matthias.wauer@tu-dresden.de

***Abstract*—This position paper describes the challenges related to federated enterprise search over heterogeneous product information sources. It focuses on the aspect of finding relevant information sources w.r.t. the user's information need and identifies core research questions with regards to the design and implementation of a potential solution, based on Semantic Web techniques.**

***Keywords*-distributed information retrieval; federated information systems; resource selection; ontologies;**

I. Introduction

A. Motivation

Within companies, product information is usually stored in a large number of different information systems. In contrast to centralized search indexes, federated search is a reasonable approach if the individual information systems are either too large for additional indexing, the information is updated very frequently, or there are too many information systems.

One major concern of federated search is the performance aspect. Not only can slow information sources delay a query. A large amount of users will also quickly stress such a federated search system if the number of queries is multiplied with the number of information sources, as shown in figure 1.

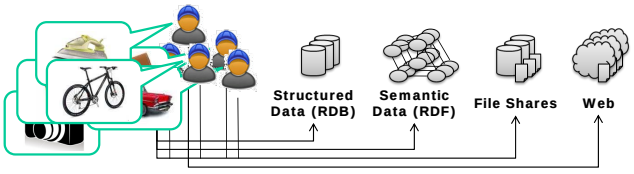

Figure 1. A federated search system directing each query to all available information sources

For environments with a restricted set of queries, the information sources to be queried may be selected using certain rules, depending on the query type. If less restricted search interfaces are required, other options for limiting the search space need to be taken into account.

This proposal describes an approach to semantic resource selection for heterogeneous product information sources. Figure 2 outlines the core components: different information sources, a *data integration broker* for accessing these sources, and a *resource selector* determining relevant sources based on a semantic representation of the query.

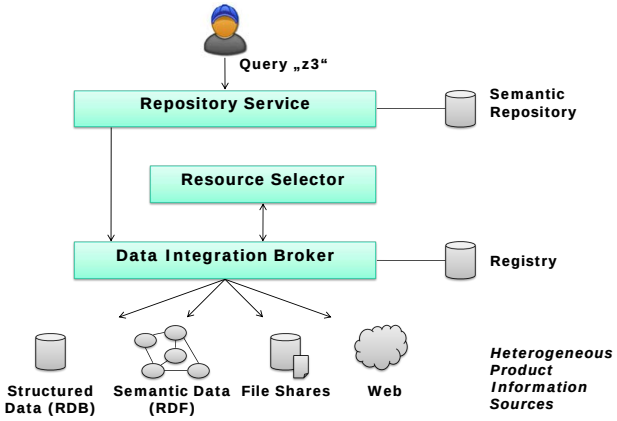

Figure 2. Simplified conceptual overview

B. Related Work

The selection of information sources has been covered independently by the information retrieval, semantic web and database communities. For text information retrieval, Schwartz et al. [1] identified and described the resource discovery problem. *GlOSS* (Glossary of Servers Server) [2] is an early work based on prior knowledge about the information sources. The relevance decision is executed by different estimators. Originally limited to boolean queries, it was later extended towards a vector space model [3]. Similar approaches, such as UUM [4] and CRCS [5] further extend the centralized index approach with e.g. learning, and SUSHI [6] directly uses the scoring of sampled documents to estimate the source relevance, without requiring training.

In Web information retrieval, more recent distributed approaches like [7] are motivated by the exponential growth of Web data and, hence, the slow update intervals of an estimated four weeks. In connection with increased Internet of Things activity and trends like corporate microblogging, a federated information integration approach appears to be similarly relevant for enterprise environments.

In the semantic web, finding relevant information is either done by creating and following links [8] [9] or indexing the information sources [10], or integrating both concepts [11]. Even though initial research [12] claims that this can be applied in an enterprise context, it is uncertain whether Semantic Web techniques can improve the retrieval of different product information that can also reside in legacy sources.

Federated database management systems, as discussed in [13], can be distinguished between tightly and loosely coupled architectures. In the context of this thesis, loosely coupled systems are assumed to be more interesting, although tightly coupled federated databases are preferred by the database community in order to ensure transactional (consistency etc.) properties. The influence of a semantic repository for schema integration (matching) of such heterogeneous databases needs to be studied further.

To the best of our knowledge, we are not aware of any research that integrates these individual aspects in order to enable resource selection based on existing semantic information.

II. Research Theses

The following theses summarize the goals of major research questions to be addressed.

Source model: The topic profile of any product information source can be represented by an appropriate source model, e.g. semantic content description or summary.

Query Model: The information need of a user can be expressed by a model that transforms queries to a representation according to the source model.

Matching: The relevance of information sources can be distinguished by means of this transformation.

Efficiency: The proposed solution leads to a significant reduction of queried information sources with only marginal loss of search results.

Extraction: The contents of the source model can be obtained semiautomatically from existing product ontologies.

III. Concept

In addition to the rather abstract theses, a more concise description of the actual intended system is given in this section. The keynote of this approach is the use of a semantic model for relevance distinction. Because a given ontology of the considered domain may have inappropriate dimensions for the selection process, this proposal introduces a smaller taxonomy that can be derived from the ontology.

An initial concept of the required components and sample data is depicted in figure 3 and explained subsequently (referring to the indicated steps):

1) Before queries are handled, the *taxonomy extractor* first connects to the *semantic repository* for generating the taxonomy index. Defined by the predicates used in the ontology, a taxonomy can be inferred directly or semi-automatically. For example, the semantic repository may contain the statement `<#BMW> skos:broader <#CarMake>`[1] that can be directly translated into a taxonomy relationship. The taxonomy index contains a mapping of all subjects found in the semantic repository to related taxonomy terms.
2) Depending on the specific query, keywords are disambiguated to URIs that identify semantic concepts or instances.
3) The *repository service* passes the returned URIs to the *query mapper*.
4) For each URI, the taxonomy index contains weighted taxonomy concepts that indicate the relevance.
5) In this case, the tuple contains the maximum weight of each taxonomy term.
6) The *resource selector* looks up the resource descriptions of available data providers from the registry and applies a ranking function.
7) The selected most relevant information sources are queried using generic stubs for the respective access type.

In order to gather resource descriptions, a *resource indexer* may either use explicitly provided descriptions from the information sources or perform sampling (a) using terms in the repository. The updated resource descriptions can then be stored in the registry (b).

IV. Evaluation

The evaluation of the proposed theses requires several foundations and a gold standard. This section summarizes the requirements, the estimated effort required to generate them, and which of the theses rely on them.

Queries

are expected to represent the information need of the user. They are required as an input for the query transformation. The query set is directly required by the thesis *query model* and can indirectly be necessary for the evaluation of thesis *efficiency* in order to measure the actual relevance of an information source regarding a certain information need.

About 50 queries should be sufficient in order to represent a reasonable variety of information needs. They can be derived from already analyzed requirements in the Aletheia [15] project with manageable effort. Additionally, the KDD Cup provides a data set[2] that may be exploited.

Information sources

exhibit heterogeneous resources that contain a variety of product information and related data. The evaluation of the theses *source model* and *efficiency* relies on these entities.

[1]The property `skos:broader` is defined in the SKOS ontology [14]

[2]http://www.sigkdd.org/kddcup/index.php?section=2005&method=data

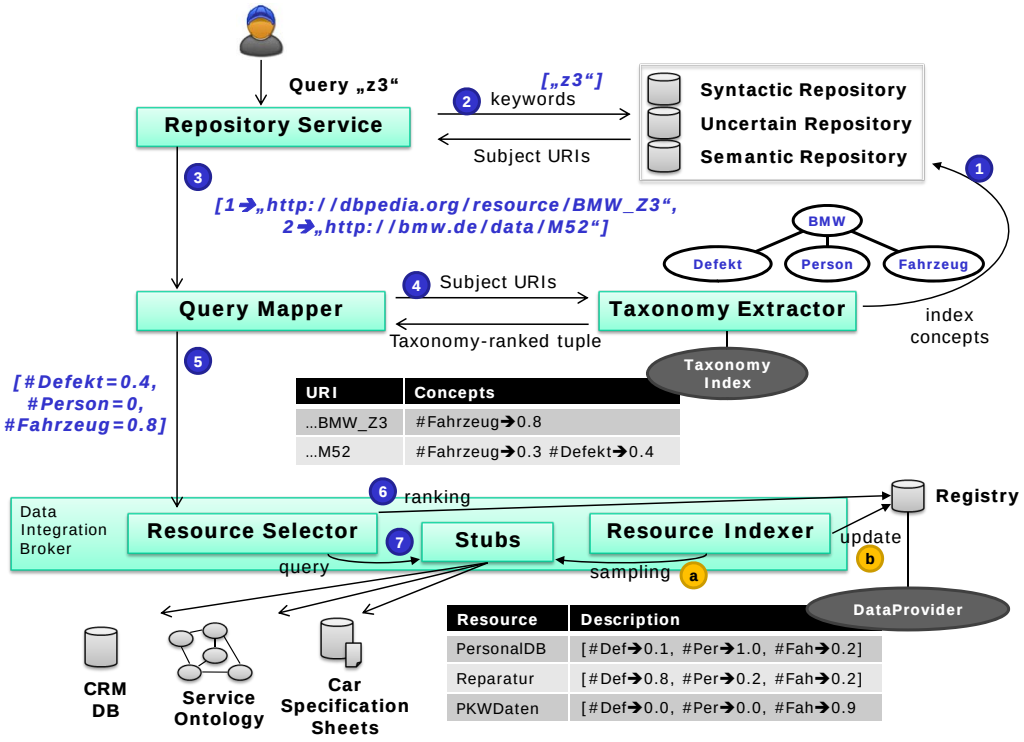

Figure 3. Detailed concept of affected components and individual data flows

Similar to the queries, a reasonable evaluation requires approximately 50 information sources of roughly 7 different access types[3]. As the designated Aletheia scenarios only provide samples of these data sets, generating such information sources of reasonable sizes requires significantly more effort. This process can be supported by utilizing publicly available data or, to some lesser degree, artificial content.

Ontology

provides the semantic foundation for extracting the taxonomy. This semantic model has to be sufficiently complete in order to provide a reasonable coverage of query terms. It is required for the *extraction* and *query model* theses and indirectly necessary for all other theses.

The Aletheia project scenarios already designed and implemented domain-specific ontologies that should only require minor adjustments.

Relevance Matching

is a gold standard defining whether each of the defined information sources is relevant for each of the queries. The *efficiency* thesis depends on it for evaluating the recall and precision of the selected information sources.

Taken into account the number of queries and information sources, the relevance matching gold standard requires $50 * 50 = 2500$ ratings.

V. Issues

This proposal presents an approach that suggests an improved solution to the resource selection problem for federated product information systems. This section summarizes issues that need to be addressed in the scope of this work.

Extraction of resource descriptions is a non-trivial task. Initially, manually defined descriptions have been used. A more sophisticated solution can use sampling of e.g. literals from the ontology in order to estimate the relevance of information sources with regards to a concept of the taxonomy.

Balancing the required repository size while still preserving the federated characteristic is difficult:

Minimum:: the repository only consists of labels directly related to the taxonomy terms. For example, the term “car” is stored, but “z3” as a certain model is too specific. A transformation can only be found for a few very abstract queries. Hence, for most actual queries the system has to

[3]These categories should include file shares, text databases, relational databases, RDF/SPARQL-based data sets, Web pages in general, Wiki pages as a specific Web page type, and Internet of Things data.

perform a fall back mechanism of querying all information sources.

Maximum: the repository consists of an entire index of all information sources. Even very specific query terms can be found, but all the information of the federated information sources would be stored centrally, effectively rendering a central index.

Further Aspects

More complex queries pose additional difficulties and require specific solutions. For example, the query "DD-XY1" searching for information on the specific car with the given license plate can only be related to certain information sources if either this precise value is stored in the repository, which is unlikely, or the term exhibits a known pattern that can be detected as a kind of identifier.

Additional enhancements can be achieved by addressing relevance feedback, i.e. an implicit user rating for an information source by using its information, or query logs for contiuous updates on the information source contents.

VI. Outline Draft

The main sections of the thesis will be arranged as follows:

1) introduction and motivation
2) distributed information systems *focusing on federation and semantic approaches*
3) source model *refers to thesis 1*
4) query model *refers to thesis 2*
5) selection model extraction *refers to thesis 5*
6) matching of product information sources to queries *refers to thesis 3*
7) evaluation *refers to thesis 4*
8) summary and conclusion

Acknowledgment

This work was partly funded by the German Ministry of Education and Research under the research grant number 01IA08001F.

References

[1] M. F. Schwartz, A. Emtage, B. Kahle, M. A. Sheldon, A. Duda, R. Weiss, J. W. O'toole, E. M. Voorhees, and N. K. Gupta, "A comparison of internet resource discovery approaches," *Computing Systems*, vol. 5, pp. 461–493, 1992.

[2] L. Gravano, H. Garcia-Molina, and A. Tomasic, "The effectiveness of GlOSS for the text-database discovery problem." in *ACM International Conference on Management of Data (SIGMOD 1994)*, 1994. [Online]. Available: http://ilpubs.stanford.edu:8090/68/

[3] ——, "GlOSS: text-source discovery over the internet," *ACM Transactions on Database Systems*, vol. 24, pp. 229–264, 1999.

[4] L. Si and J. Callan, "Unified utility maximization framework for resource selection," in *In Proc. ACM CIKM Conf.* ACM Press, 2004, pp. 32–41.

[5] M. Shokouhi, "Central-rank-based collection selection in uncooperative distributed information retrieval," in *Proceedings of ECIR Conference*, 2007, pp. 160–172.

[6] P. Thomas and M. Shokouhi, "Sushi: scoring scaled samples for server selection," in *SIGIR '09: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*. New York, NY, USA: ACM, 2009, pp. 419–426.

[7] S. Bockting and D. Hiemstra, "Collection selection with highly discriminative keys," in *Proceedings of the 7th workshop on large-scale distributed systems for information retrieval (LSDS-IR'09), SIGIR 2009*, 2009. [Online]. Available: http://citeseerx.ist.psu.edu/viewdoc/summary?doi=10.1.1.147.6264

[8] T. Berners-Lee, "Linked data," http://www.w3.org/DesignIssues/LinkedData.html, July 2006, last accessed February 22nd, 2010.

[9] C. Bizer, J. Volz, G. Kobilarov, and M. Gaedke, "Silk - a link discovery framework for the web of data," in *18th International World Wide Web Conference*, April 2009. [Online]. Available: http://www2009.eprints.org/227/

[10] L. Ding, T. Finin, A. Joshi, R. Pan, R. S. Cost, Y. Peng, P. Reddivari, V. Doshi, and J. Sachs, "Swoogle: a search and metadata engine for the semantic web," in *CIKM '04: Proceedings of the thirteenth ACM international conference on Information and knowledge management*. New York, NY, USA: ACM, 2004, pp. 652–659.

[11] G. Tummarello, R. Delbru, and E. Oren, "Sindice.com: Weaving the open linked data." in *ISWC/ASWC*, ser. Lecture Notes in Computer Science, K. Aberer, K.-S. Choi, N. F. Noy, D. Allemang, K.-I. Lee, L. J. B. Nixon, J. Golbeck, P. Mika, D. Maynard, R. Mizoguchi, G. Schreiber, and P. Cudr-Mauroux, Eds., vol. 4825. Springer, 2007, pp. 552–565. [Online]. Available: http://dblp.uni-trier.de/db/conf/semweb/iswc2007.html#TummarelloDO07

[12] F.-P. Servant, "Linking enterprise data," in *Linked Data on the Web (LDOW2008)*, 2008. [Online]. Available: http://data.semanticweb.org/workshop/LDOW/2008/paper/17

[13] A. P. Sheth and J. A. Larson, "Federated database systems for managing distributed, heterogeneous, and autonomous databases," *ACM Comput. Surv.*, vol. 22, no. 3, pp. 183–236, 1990.

[14] W3C Semantic Web Activity, "SKOS simple knowledge organization system," http://www.w3.org/2004/02/skos/, August 2009. [Online]. Available: http://www.w3.org/2004/02/skos/

[15] M. Wauer, D. Schuster, J. Meinecke, T. Janke, and A. Schill, "Aletheia - towards a distributed architecture for semantic federation of comprehensive product information," in *Proceedings of IADIS International Conference WWW/Internet*, Rome, Italy, 2009.