Advanced Resource Selection for Federated Enterprise Search

Matthias Wauer, Daniel Schuster, and Alexander Schill

TU Dresden, Faculty of Computer Science, Chair for Computer Networks,
Helmholtzstraße 10, 01062 Dresden, Germany
{matthias.wauer,daniel.schuster,alexander.schill}@tu-dresden.de

Abstract. Distributed information retrieval is a well-known approach for accessing heterogeneous, highly autonomous sources of unstructured information. Selecting and querying only a number of relevant sources can help improve its performance, but most resource selection algorithms are limited to syntactic comparisons.
We present a framework for applying resource selection in the context of a semantic federated product information system, and evaluate the performance of the well-known CORI resource selection algorithm in this context.

Key words: resource selection, distributed information retrieval, federated search, enterprise search, enterprise information systems

1 Introduction

Product information of companies are stored in many different sources, typically due to organizational and technical requirements. This diversification prevents a comprehensive view and seamless access to this information. Furthermore, most of the relevant product information is only available in unstructured form.

We are therefore developing a federated product information system (FPIS) [1] with regards to innovative application scenarios of five collaborating industry partners (ABB, BMW, Deutsche Post DHL, Otto, SAP). It connects federated heterogeneous information providers using semantic middleware. Structured information can already be integrated by semantically mapping their respective schemas to an ontology using existing tools. One of our goals is to extract information from documents as needed in order to associate them with an ontology.

The sheer number of documents produced in companies today can barely be managed in a central document database. Distributed management of documents allows for greater autonomy of the individual *collections* but requires a more sophisticated approach to search. Centrally indexing these documents can only be done if all collections are collaborative, i.e., if they provide immediate access to all of their documents. An alternative is distributed information retrieval [2], which issues a query to the search engines of each collection and merges all results. It can be applied to all collections that provide a search interface.

Drawbacks of distributed information retrieval are the potential processing and communication overhead for a large number of collections and increased response times if the response of a collection is delayed. Hence, a resource selection algorithm is supposed to reduce the number of queried collections, typically to those that are estimated to be relevant with regards to the query.

We evaluate the applicability of an existing well-known resource selection algorithm for an FPIS on a corpus of industrial service documents, and propose a framework which utilizes the available semantic information for improving resource selection performance.

2 Related Work

With regards to resource selection, CORI [3] is one of the most popular algorithms. It uses *per collection* statistical features to estimate the relevance of collections, based on inference network document ranking. Queries are expected to be a simple set of terms.

The actual computation estimates two components for each term: a term-based measure $T_{i,t}$ which uprates a term that occurs frequently in collection i w.r.t. average and collection-specific number of different terms, and a collection-evaluative measure which increases the impact of highly distinctive terms, e.g., terms that only occur in few collections. Each term in query $Q = \{t_1, t_2, \ldots, t_n\}$ is weighted equally.

Some drawbacks of CORI have been identified [4] and addressed by other approaches. ReDDE [5] is less prone to disregard large collections if the collections are skewed, i.e., the collections vary considerably in size. For similarly sized collections, results improve marginally. CRCS [6] and SUSHI [7] find that these algorithms barely use the document samples of each collection and their scores for each query, although they are valuable for assessing a collection's relevance. They also determine how many collections should be selected, whereas CORI usually selects a fixed amount of them.

Collections are typically assumed to be independent, so relationships between them are typically not taken into account by these algorithms. Hong et al. [8] present a model that classifies resources not only on singular features for each resource, but also on joint similarity between resources as an additional feature. They estimate the importance of detecting the similarity by applying different algorithms, and conclude that a similarity metric based on relevance for each query performs better than a language model based Kullback-Leibler metric, which performs worse than the common independent approach. The differences are fairly small with TREC testbeds. However, the performance increases significantly for a real-world testbed, in particular for high precision values, i.e., the topmost source ranking results.

Arguello and colleagues [9] extend the document-based selection with both an estimated query topic and query click-through data. These three evidences, namely corpus-based, query-categorical, and click-through features, are combined using a machine learning algorithm, which is initialized with automatically generated training data. Evaluation of this approach shows that the categorization of a query can improve the accuracy significantly if the collection sample is small.

3 Resource Selection Concept

The resource selection is part of our current FPIS, the Aletheia prototype [1]. Similar to Arguello et al. [9], the proposed solution should be able to combine several features for the final collection relevance assessment, but in a much more extensible way as shown in Figure 1. The processing of such features is wrapped as *plugin* components that can be applied flexibly depending on the actual scenario. Connector components are the actual mediators communicating with the federated information providers.

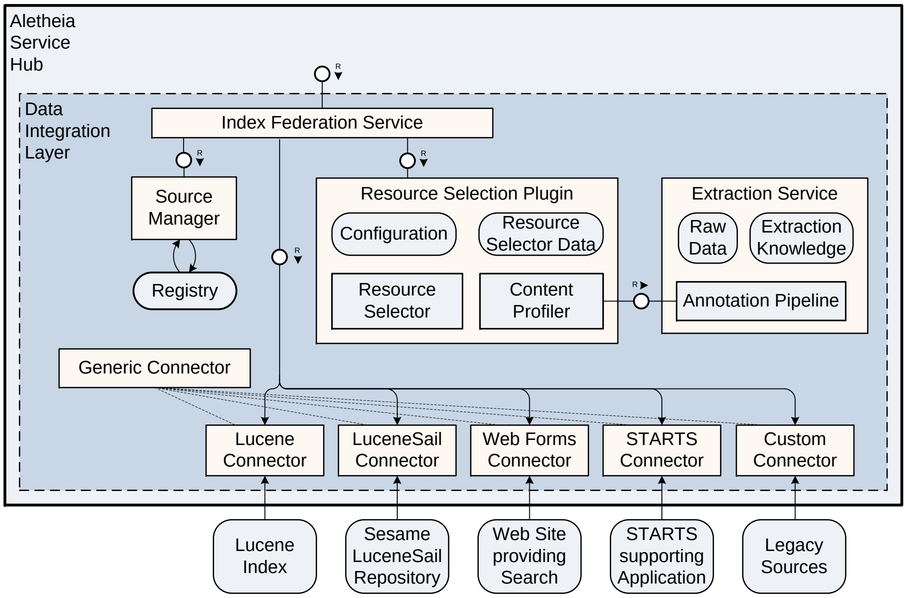

Fig. 1. Architecture of a resource selection framework, in FMC notation

Knowledge based resource selection can be applied by using the Extraction Service provided by the FPIS, which annotates the sampled documents semantically. A semantic resource selection plugin may adapt this component by applying custom UIMA [10] annotators.

The integration of this federated query processing with other components of the Aletheia Service Hub (not shown here) enables many other features, e.g., adding and modifying semantic tags for documents by the user.

4 Preliminary Evaluation

In order to find out how existing algorithms perform, a CORI resource selection plugin indexed different test sets before evaluating a range of queries.

4.1 Test Sets

The evaluation was executed on multiple test sets in order to find out how the framework and algorithms perform. All test sets were derived from a collection of real industrial documents, a subset of a project partner's digital library compiled for offline use. This library consists of

- A topical structure T (tree.xml) containing links to
- Node files (*nodeId*.xml) describing a set of documents D related to the node's topic, and
- 3.624 folders, each containing one of the documents d in D, with a total of 2.89GB of files.

This library is analysed and split into appropriately sized collections, assuming that the content of a sub-tree's referenced documents are related to a limited set of topics as in, e.g., files of a certain workgroup. can As a first attempt, the sum s_{node} of the number of documents in its own node file and all sub-node's files is appended to each node in T. Then, an XPath [11] expression can be applied to find all nodes having a defined minimum and maximum collection size. This approach, however, does not result in the expected collections because of similar product's node files often reference the same documents. Hence, the sum of *unique* documents $s_{u_{node}}$ is usually much less than expected.

A second algorithm therefore not only counts the number of documents, but traverses through T computing the list of unique documents for each node, not including the documents of sub-nodes that form a collection themselves.

For some collections, the sum $s_{u_{node}}$ can still exceed the expected maximum collection size. If they are composed of documents from multiple nodes in T they may be split, but for the evaluated collections it is not reasonable to do so due to the topical clustering. Using this algorithm, three test sets have been generated as shown in Table 1.

Table 1. Test sets generated from the document samples

Test set	#Collections	Expected size range	Overflow of $s_{u_{node}}$
TS_{small}	≈ 230	20–50 documents	35 collections (4 > 100 documents)
TS_{large}	9	250–500 documents	1 collection (557 documents)
TS_{skew}	≈ 50	manually compiled from TS_{small} and TS_{large} aiming for low overlap, to analyse shortcomings w.r.t. collection skew	

The queries have partly been taken from a developing gold standard of the FPIS. As an exceptionality of the FPIS, they are typically *hybrid*, i.e., they consist of semantic elements identifying concepts or instances of the ontology and literals which resemble keywords.

The query intents classified by Broder [12] for Web search (informational, navigational, transactional) can not be applied directly, but the queries can roughly be distinguished between:

- *immediate informational:* the query should return one document and ideally answer the information need in the first document snippet
- *composed informational:* the information need can't be answered by a single document, but several relevant documents need to be studied for an answer

Navigational queries can be considered similar to immediate informational queries in that they focus on a single document (instead of a certain Web site), whereas transactional queries are inapplicable here.

4.2 Results

The CORI algorithm has been modified to select a variable set of collections, based on a fixed threshold. It produced mixed but consistent results for a set of 12 queries. With short queries identifying a certain product, CORI typically selected very few selections and ranked the most relevant with a high accuracy.

For *immediate informational* queries, such as "[product] error 3", performance dropped significantly, apparently because the discriminating first query part was suppressed. The algorithm failed to rank the most relevant resource topmost for about half the queries, but it always remained above the threshold.

Composed informational queries showed a worse performance, with a distinct uncertainty in the selection results indicated by a low precision. For example, the query "'sensor drive' fitting procedure" yields some documents explaining how to install such a product option, but CORI fails to accurately distinguish collections using these barely specific terms. Furthermore, applying clustering to the distribution of CORI scores would not clearly discern relevant sources.

5 Conclusions

The preliminary evaluation shows that the performance of existing syntactic algorithms varies considerably regarding the kind of query. For more ambiguous queries, the syntactic approach is blatantly limited. Resource selection performance will probably benefit from a more thorough knowledge based analysis. The envisioned federated product information system supports this extension by providing semantic annotation services and an integrated hybrid query processing. Thus, users are encouraged to explicitly define the intended query terms in order to improve precision.

Future research will evaluate the existing algorithm quantitatively, based on an extended set of queries, and propose an index structure for efficient matchmaking of semantic query terms. We expect that an independent feature model and algorithms like Naive Bayes can be applied to combine the individual plugins' results.

6 Acknowledgement

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

References

1. Wauer, M., Schuster, D., Meinecke, J.: Aletheia: an architecture for semantic federation of product information from structured and unstructured sources. In: Proceedings of the 12th International Conference on Information Integration and Web-based Applications & Services. iiWAS '10, New York, NY, USA, ACM (2010) 325–332
2. Callan, J.: Distributed information retrieval. In: In: Advances in Information Retrieval, Kluwer Academic Publishers (2000) 127–150
3. Callan, J.P., Lu, Z., Croft, W.B.: Searching distributed collections with inference networks. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '95, New York, NY, USA, ACM (1995) 21–28
4. Si, L., Lu, J., Callan, J.: Distributed information retrieval with skewed database size distributions. In: Proceedings of the 2003 annual national conference on Digital government research. dg.o '03, Digital Government Society of North America (2003) 1–6
5. Si, L., Callan, J.: Relevant document distribution estimation method for resource selection. In: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in information retrieval. SIGIR '03, New York, NY, USA, ACM (2003) 298–305
6. Shokouhi, M.: Central-rank-based collection selection in uncooperative distributed information retrieval. In: Proceedings of ECIR Conference. (2007) 160–172
7. Thomas, P., Shokouhi, M.: 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) 419–426
8. Hong, D., Si, L., Bracke, P., Witt, M., Juchcinski, T.: A joint probabilistic classification model for resource selection. In: Proceeding of the 33rd international ACM SIGIR conference on Research and development in information retrieval. SIGIR '10, New York, NY, USA, ACM (2010) 98–105
9. Arguello, J., Callan, J., Diaz, F.: Classification-based resource selection. In: Proceeding of the 18th ACM Conference on Information and Knowledge Management. CIKM '09, New York, NY, USA, ACM (2009) 1277–1286
10. Ferrucci, D., Lally, A.: UIMA: an architectural approach to unstructured information processing in the corporate research environment. Nat. Lang. Eng. **10**(3-4) (2004) 327–348
11. Clark, J., DeRose, S.: XML Path Language (XPath) version 1.0. Recommendation, World Wide Web Consortium (November 1999) See `http://www.w3.org/TR/xpath.html`.
12. Broder, A.: A taxonomy of web search. SIGIR Forum **36** (September 2002) 3–10