Integrating Explicit Semantic Analysis for Ontology-based Resource Selection

Matthias Wauer
TU Dresden
Computer Networks Group
Dresden, Germany
matthias.wauer@tu-dresden.de

Daniel Schuster
TU Dresden
Computer Networks Group
Dresden, Germany
daniel.schuster@tu-dresden.de

Alexander Schill
TU Dresden
Computer Networks Group
Dresden, Germany
alexander.schill@tu-dresden.de

ABSTRACT

In federated information systems, deciding whether an information source is relevant for a given query is crucial for its overall performance. Focusing on uncooperative unstructured information sources, we analyze several drawbacks of the popular CORI resource selection algorithm by evaluating it in a federated product information scenario. Based on these results, we propose and describe a novel approach using an ontology-based sampling method, which is used to initialize an Explicit Semantic Analysis index.

Categories and Subject Descriptors

H.3.4 [**Information Storage and Retrieval**]: Systems and Software; H.4.0 [**Information Systems Applications**]: General

General Terms

Design, Management, Experimentation

Keywords

Federated Information System, Distributed Information Retrieval, Resource Selection, Explicit Semantic Analysis, Product Information Management, Ontology

1. INTRODUCTION

Product-related information is generated, accessed and manipulated along the product lifecycle in heterogeneous formats. Only part of this information can be accessed using state-of-the-art product information systems, because most of this information is only available in unstructured sources. A comprehensive Federated Product Information System (FPIS) therefore has to integrate and harmonize data from all phases of the product lifecycle, and different source formats like unstructured documents, sensor information or product databases for the design, production, delivery, and service of a product [11].

Most of these sources are either legacy sources, or they belong to different administrative domains. The sheer number of such resources, such as department-specific document management systems with their own search Web services, often prevents physical integration approaches and comprehensive semantic analysis of the documents. Virtual integration, like federated search, provides means to access such distributed information sources in an ad-hoc fashion.

We will focus on such approaches for unstructured information sources. In distributed information retrieval, three processing steps can be distinguished [2]:

Resource description describes the process and result of analyzing the content of information sources,

Resource selection denotes the method for distinguishing between relevant and non-relevant sources for a given query,

Result merging specifies how individual result lists from selected sources can be merged and re-ranked according to the user's information need.

Existing approaches to these problems are based on vector space or language models. They provide computationally efficient means to describe and select different resources, but they have several drawbacks, as discussed in section 2. According to Gray [5], network costs have the highest influence on distributed computing economics, which is another reason to improve resource selection performance. The FPIS context provides additional information that can potentially improve the performance of such methods. For example, the Aletheia reference architecture [12] defines a semantic product information repository that can be extended with information from a company's database using ontology mapping. In this paper, we discuss an approach on how such an ontology can be utilized, focusing on the resource description and resource selection problems.

The contributions of this paper consist of

1. An evaluation of the performance and shortcomings of existing resource selection algorithms in the context of a federated product information system scenario in Section 3,

2. A novel approach to the resource description and resource selection problems using the Explicit Semantic Analysis method for enabling ontology-based resource selection in Section 4.

iiWAS2011, 5-7 December, 2011, Ho Chi Minh City, Vietnam.

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.

The actual resource selection estimates the relevance probability using two components for each query term. A term-based measure $T_{i,t}$ 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. Terms in a query $Q = \{t_1, t_2, \ldots, t_n\}$ are simple strings and weighted equally.

Some drawbacks of CORI have already been identified and addressed by other approaches. ReDDE [8] 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. Thomas and Shokouhi [9] 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. All of these approaches have a relatively similar performance, with minor differences depending on individual test sets.

Collections are typically assumed to be independent, so relationships between them are typically not taken into account by these algorithms. Hong and Arguello et al. [6, 1] proposed classification-based and feature-combining methods that perform marginally better, with performance increasing significantly for high precision and small sample size scenarios.

Neither of the proposed resource selection methods make use of explicitly semantic information, such as an ontology and documents labelled accordingly. However, there are information retrieval models that attempt to utilize it. Paralic and Kostial [7] use semantic concepts similar to a boolean model with manually tagged documents. Vallet et al. [10] add heuristics-based semantic tagging of documents and support for complex semantic queries on a knowledge base. However, these approaches require a large and detailed ontology in order to provide any advantage.

To avoid this ontology coverage problem, concept-based information retrieval such as Explicit Semantic Analysis may be applied. Egozi et al. [4] propose the creation of a high-dimensional concept space by, e.g., extracting Wikipedia articles. The title of the article denotes the concept, whereas the terms in the article content are used to build an inverted index which links each term to a weighted concept, thus indicating the concept's relevance for each term. Then, documents and queries can be represented by concept vectors, which can be compared using common similarity metrics. The obvious drawback of this approach in the context of this paper is the need for a Wikipedia-like description of the domain terminology which should be covered by the FPIS.

3. EVALUATING SYNTACTIC METHODS

Since none of the existing test sets accurately captures the specific requirements of a FPIS scenario, we decided to define an appropriate test set and respective gold standard. We assume that specific products are designed, manufactured, and managed by certain departments of a company specialized on this product range. The test sets are therefore based on a set of about 3600 documents (2.89GB in total) provided by industrial device and service company ABB, which have been divided into individual collections using the respective product type reference known for each document. Details on the test set generation process can be found in [13]. The characteristics of the test sets used in this evaluation are shown in Table 1.

Table 1: Test sets created from the document base

Test set	#Collections	Size range
TS_{small}	$\approx$ 230	$\approx$ 20–50 documents
TS_{large}	9	$\approx$ 250–500 documents

Similarly, the gold standard defines a query set that has partially been captured at ABB, with the user base consisting mainly of service engineers. Each query has a number of relevant documents assigned, i.e., the gold standard contains no detailed weighting of document relevances. It consists of 16 queries. These queries can be roughly divided into 3 types:

Factual queries contain only direct references to a certain product.
Example: "FXE4000" (a specific ABB product)

Extended factual queries contain a direct reference to a product, augmented by additional constraints such as properties or aspects of the product or the context it is used in.
Example: "FXE4000 configuration"

Non-factual queries do not contain a specific product reference. They are typically posed by customers with a certain need, but no detailed knowledge of the specific products or parts that are related to it.
Example: "chemical flow precise density measurement" (indicates a need for precisely measuring the density of a flow of matter in a chemical branch application)

Using these test sets, we evaluated the widely used CORI resource selection algorithm. We implemented it as a plugin for an extensible federation framework [13], which can access remote information sources via Web service connections and connect to legacy providers by custom wrappers.

3.1 Precision and Recall

When evaluating the relevance of resources for a specific query, CORI assigns a score for each resource, but does not define a relevance decision, i.e., which of the resources should be selected for a federated query. Hence, the initial evaluation focused on the accuracy of the CORI ranking. Ideally, the algorithm should return all relevant resources (high *recall*) and return no irrelevant resources (high *precision*) at the same time.

Figures 1 and 2 shows that CORI is very precise when the query contains very discriminative terms, such as product names. In the case of non-factual queries, it is much more difficult to disseminate the relevant resources by syntactic comparisons alone.

The above results are idealized, because the CORI resource descriptions have been generated from all documents of each collection. In reality, only a small subset of documents can and should be retrieved for this purpose. Figure 3

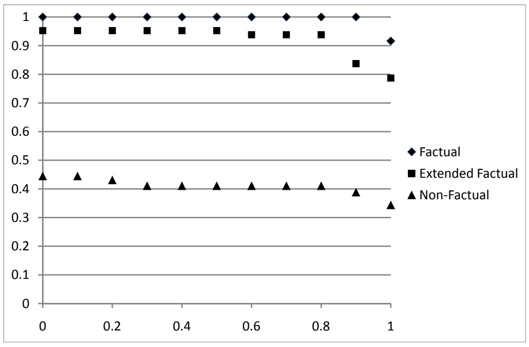

Figure 1: Precision and recall for TS_{small}

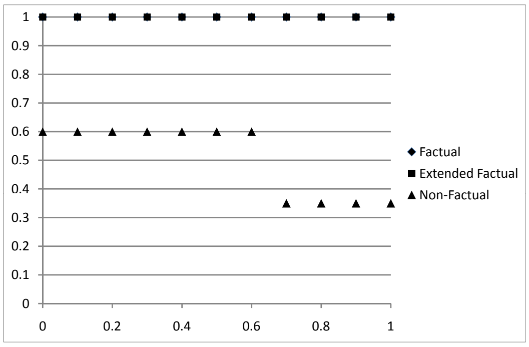

Figure 2: Precision and recall for TS_{large}

shows that CORI performs much worse when a subset of 5 documents or 10% of each collection is retrieved for creating the resource description, in particular when considering the case when all relevant sources should be selected (recall=1.0).

3.2 Relevance Assessment

As CORI does not distinguish relevant from non-relevant collections, usually a fixed number (cutoff) of sources is selected. This is not a sufficient behaviour, because queries vary from highly selective to broad scope w.r.t. the relevant collections. A fixed cutoff therefore leads to either low precision or low recall if the cutoff is higher or lower than the number of relevant collections, respectively.

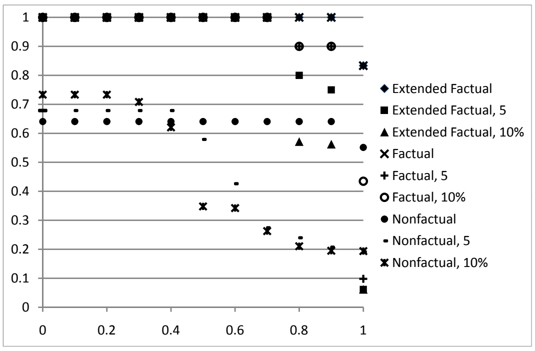

Figure 3: Precision and recall for samples of 5 documents and 10% of a collection for TS_{small}

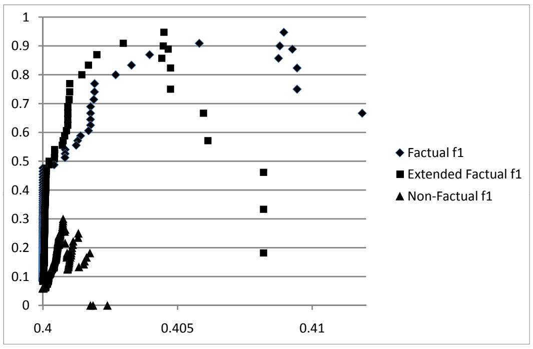

Figure 4: F-measure vs. CORI scores for TS_{small}

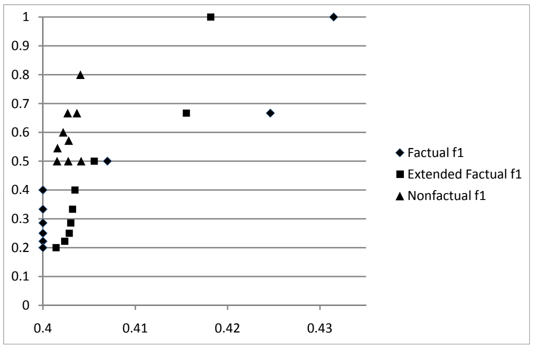

Figure 5: F-measure vs. CORI scores for TS_{large}

Thus, we evaluated whether the score of the CORI ranking can be utilized for this purpose. Figures 4 and 5 show the F-measure, which is the harmonic mean of recall and precision, when the cutoff is set to the CORI score of each ranked resource. The most appropriate score-based cutoff (threshold t) would be around the highest F-measure when precision and recall are weighted equally.

The results indicate that the score is not a good indicator for relevance classification. TS_{small} particularly shows very different appropriate threshold values. Whereas for factual queries the threshold t should be between 0.404 and 0.408, non-factual queries instead have a maximum score of 0.4025. As a result, a good t for factual queries would not classify any resource relevant for a non-factual query. The maximum F-measure for non-factual queries is again at a much lower theshold of 0.4008.

4. ONTOLOGY-BASED EXPLICIT SEMANTIC ANALYSIS

The low precision of CORI-based resource selection of non-factual queries indicates that the syntactic approach is not able to distinguish the meaning, and therefore the target of such a query. We therefore propose to apply ontology-based semantics available in a FPIS to extend the resource selection capabilities. In order to circumvent the low ontology coverage issue, we suggest to use the ontology for bootstrapping the Explicit Semantic Analysis (ESA) approach as follows:

1. *Ontology extraction* returns the concepts C stored in the ontology, along with additional information such as related instances and their labels L.

2. For each concept in C, all collections are queried for the label l (query-based sampling). The results from each resource are merged and the highest-ranked document(s) are given as input to the ESA analyzer for concept c. Thus, this *ontology-based sample* document content is similar to the Wikipedia document described in Section 2, with c being the Wikipedia article's title.

3. When sample documents for all concepts have been indexed using ESA, the individual collections are sampled and get an ESA vector assigned. This vector v_{c_i} is a weighted set of all concepts a collection is relevant for.

After this initial learning phase, the following steps are executed during resource selection for a query q:

4. ESA analyzes q, thus transforming it to a concept vector v_q.

5. This query representation v_q is compared to all collection vectors v_c using a similarity metric, such as cosine similarity.

6. The similarity score distinguishes whether a source is relevent for this query.

There are several possible adjustments in this process. In step 1, the extraction process may be restricted to a subset of the concepts, e.g., only the topmost concepts. The number of documents retrieved in step 2 is relevant for the coverage of ESA terms. Initial tests have shown that pre-processing of the documents can be necessary to improve the quality of the ESA index. Finally, the similarity metric in step 5 can be adapted so the score allows for a more precise differentiation of relevant and non-relevant sources.

5. CONCLUSIONS

The CORI resource selection algorithm has limitations w.r.t. the selection accuracy and its relevance assessment capabilities. For the envisioned federated product information system, we propose an advanced resource selection mechanism that utilizes domain knowledge from an ontology.

An initial prototype of this concept is implemented using an F-Logic ontology and a Java-based ESA implementation[1]. We plan to further enhance this prototype and evaluate against more comprehensive gold standards. In general, we believe this component is a crucial aspect of a federated product information system that is capable of efficiently integrating distributed Web information sources, performing adequate to a centralized solution.

Compared to a previous publication [13] focusing on the framework for resource selection mechanisms in FPIS architectures, this paper proposes a method using the semantic knowledge base for applying an explicit semantic analysis resource selection algorithm.

6. ACKNOWLEDGMENTS

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

[1] http://code.google.com/p/research-esa/

7. REFERENCES

[1] J. Arguello, J. Callan, and F. Diaz. Classification-based resource selection. In *Proceeding of the 18th ACM Conference on Information and Knowledge Management*, CIKM '09, pages 1277–1286, New York, NY, USA, 2009. ACM.

[2] J. Callan. Distributed information retrieval. In *Advances in Information Retrieval*, pages 127–150. Kluwer Academic Publishers, 2000.

[3] J. P. Callan, Z. Lu, and W. B. Croft. 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, pages 21–28, New York, NY, USA, 1995. ACM.

[4] O. Egozi, S. Markovitch, and E. Gabrilovich. Concept-based information retrieval using explicit semantic analysis. *ACM Trans. Inf. Syst.*, 29:8:1–8:34, April 2011.

[5] J. Gray. Distributed computing economics. *Queue*, 6(3):63–68, 2008.

[6] D. Hong, L. Si, P. Bracke, M. Witt, and T. Juchcinski. 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, pages 98–105, New York, NY, USA, 2010. ACM.

[7] J. Paralic and I. Kostial. Ontology-based information retrieval. *Information and Intelligent Systems, Croatia*, pages 23–28, 2003.

[8] L. Si and J. Callan. 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, pages 298–305, New York, NY, USA, 2003. ACM.

[9] 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*, pages 419–426, New York, NY, USA, 2009. ACM.

[10] D. Vallet, M. FernÃąndez, and P. Castells. An Ontology-Based Information Retrieval Model. In *The Semantic Web: Research and Applications*, Lecture Notes in Computer Science, pages 455–470, 2005.

[11] M. Wauer, D. Schuster, and J. Meinecke. 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, pages 325–332, New York, NY, USA, 2010. ACM.

[12] 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.

[13] M. Wauer, D. Schuster, and A. Schill. Advanced resource selection for federated enterprise search. In *Proceedings of BIS Workshops 2011*. Springer, 2011.