

Enhanced Federated Search for Digital Object Repositories

Christian Kohlschütter¹, Dierk Höppner², and Maria Nejd1¹

¹ iSearch IT Solutions GmbH

{kohlschuetter, nejd1}@isearch-it-solutions.de

² German National Library of Science and Technology (TIB)

Dierk.Hoepfner@tib.uni-hannover.de

1 Introduction

Federated Search provides interoperability between distributed Digital Object Repositories and Digital Libraries for cooperatively retrieving information elements by keywords and parametrized search via a common user interface, regardless of the underlying search systems. This task inevitably poses several challenges, first of all the harmonization of all information and parameters (object schemata, ranking statistics, query operators, protocols) required for conducting the search across all participating libraries.

In the context of the Federated Search project at the German National Library of Science and Technology (TIB), started in 2006, we developed a novel Federated Search infrastructure which addresses these points, introducing an extension of the well-established SRU/SRW standard [2]. The system connects TIB bibliographic collections to the repositories of its customers and partners, e.g. FIZ Technik. Unlike traditional systems, our approach provides an efficient, homogeneous relevance ranking across the participating (and potentially incompatible) repository systems, faceted search for result drill-down as well as connectivity to other federations such as the German *vascoda* federation.

2 The SRX/FS Protocol

The SRU/SRW standard has proven to work well for basic Federated Search. However, to support relevance ranking and drill-down, we had to extend the standard in several ways; we call this extension SRX/FS.

For a consistent relevance ranking, the underlying algorithms need to work on federation-wide statistics (such as the Term-DF distribution), which we can retrieve via SRU's *scan* operation. SRX/FS adds an *override-scan* operation which transmits term statistics across the federation's participants for later aggregation. This way, every search engine knows exactly how to independently rank its part of the search results. Search requests can be processed in parallel by each federation member's search engine, without transmitting ranking parameters along with the query (this would simply be too expensive for wildcard queries, for instance). Faceted Search (drill-down) can be provided without federation-wide information or interaction between search engines, which allows for very fast execution times. SRX/FS provides a corresponding SRU extension

encapsulated into *extra-request-data*. Drill-down can be activated on a per-field basis, with refinement options sorted alphabetically or ranked by the number of hits, with an optional maximum number of terms to be returned. We chose SRU's XML-based XCQL query language, which is easily parseable and better extensible than the string-based CQL. SRX/FS extends XCQL by providing additional structured and parametrized query types, including fuzzy search and query term boosting.

3 Implementation and Evaluation

The reference implementation of SRX/FS is based on Lucene [1]. The system can be queried using SRX/FS (XCQL) as well as natively using Lucene `Query` objects. It transparently integrates into existing Lucene contexts (providing a custom `Searcher` class) and can query any locally connected Lucene index as well as SRX/FS-based remote collections. The integration of virtually any other type of repository is possible through indexer plug-ins [4], which provide the data to the federation via a compatible Lucene index. We have already written one such plug-in for FAST Data Search; plug-ins for SQL databases and Verity/Autonomy are currently being developed. For systems which already utilize Lucene (such as SOLR, Dspace and Fedora), re-indexing is not necessary if the used object schema is compatible. The system has also been connected to the *vascoda* federation [3] (the *vascoda* portal does not rank the results but merges them in a round-robin manner, though). The TIB GetInfo portal incorporating our system will go live in Q3/2008.

We have evaluated the system in a remote setup between TIB and FIZ Technik as well as in the *vascoda* federation and compared distributed search against a local setup. The test collection consisted of about 3 million bibliographical objects. In all cases, using SRX/FS-based Federated Search we were able to achieve the same relevance ranking as for the local Lucene system, and comparable search times (50-100ms per query including network latency) and drill-down operation characteristics (usually 100ms to 3s, depending on the query).

4 Discussion

We have three main objectives up for discussion:

1. Integration of existing repositories
2. Support of existing DOR system
3. Protocol Enhancements (e.g. for access control).

References

1. Lucene. <http://lucene.apache.org/>.
2. SRU/SRW: Search/Retrieval via URL. <http://www.loc.gov/standards/sru/>.
3. Vascoda. <http://www.vascoda.de/>.
4. Sergey Chernov, Christian Kohlschütter, and Wolfgang Nejd. A Plugin Architecture Enabling Federated Search for Digital Libraries. In *ICADL*, pages 202–211, 2006.