Towards a Scalable Search and Query Engine for the Web*

Aidan Hogan, Andreas Harth, Jürgen Umbrich and Stefan Decker
National University of Ireland, Galway
Digital Enterprise Research Institute
Galway, Ireland
firstname.lastname@deri.org

ABSTRACT

Current search engines do not fully leverage semantically rich datasets, or specialise in indexing just one domain-specific dataset. We present a search engine that uses the RDF data model to enable interactive query answering over richly structured and interlinked data collected from many disparate sources on the Web.

Categories and Subject Descriptors

H.3.0 [**Information Storage and Retrieval**]: General

General Terms

Algorithms, Design

Keywords

entity-centric, semantic web, web search

1. INTRODUCTION

Google and Yahoo! index vast amounts of documents on many topics, and rely on keyword user input to offer links to relevant documents.

In recent years, there has been a trend towards domain specific, database-backed sites such as CiteSeer, DBLP, IMDB and Wikipedia, which support browsing of information within the interface and offer structured search functionality. Instead of only links to relevant documents, these domain specialised sites return answers to user queries. However, the sites are isolated and have a finite coverage of data.

The search engine presented draws upon technologies from the research area of the Semantic Web to provide search and query functionality over data integrated from a large number of sources. We call our system Semantic Web Search Engine (SWSE). SWSE aims to present the user with the information they require within the interface.

SWSE is an entity-centric application; results are structured descriptions of entities, including people, conferences, papers and documents, as opposed to just HTML pages. SWSE integrates data on entities from multiple sources, hence attention is given to tracking the provenance of data. Complementary data on entities is consolidated for more succinct results.

The user interface is intended to remain close to current search engine interfaces; casual users pose queries without a priori schema knowledge and explore datasets of any schema using uniform processes. Results navigation and browsing is offered through relationships from the current result entities to related entities. Presentation of results is ordered according to relevancy and importance.

Our initial use-case for the system is to provide search, query and browsing functionality over a dataset relevant to computer science research.

2. SYSTEM OVERVIEW

In this section we will briefly describe our data model and present the SWSE architecture shown in Figure 1.

We use a graph-based data model established in the area of databases for the descriptions of entities. One commonly used framework for semantic graphs is RDF, which maps to triples (subject, predicate, object). We deem entities to be resources characterised by URIs and their description to consist of the triples containing that URI in the subject or object position.

Given that we deal with web data which may originate from a plethora of sources, we need to keep track of the provenance of information. Thus, we extend the classic RDF data model with the notion of context [1]; sometimes called named graphs. The context part of the quadruple (subject, predicate, object, context) encodes the URL from hence the triple originated.

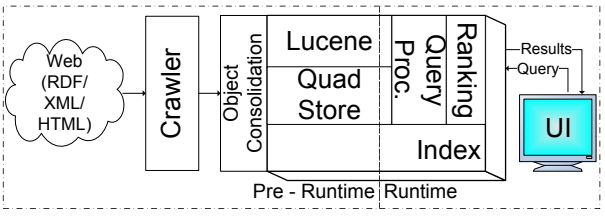

Figure 1: High-level architecture

For our initial test-case of computer science research data, we require extraction of relevant data from the Web (e.g., Friend Of A Friend data) for which we use the crawling architecture MultiCrawler [2]. The crawler is able to syntactically transform data from a variety of sources (e.g., HTML,

*This work has been supported by Science Foundation Ireland (SFI/02/CE1/I131), EU projects DIP (FP6-507483) and TripCom (IST-4-0027324-STP).

WWW 2007, May 8–12, 2007, Banff, Alberta, Canada.
ACM 978-1-59593-654-7/07/0005.

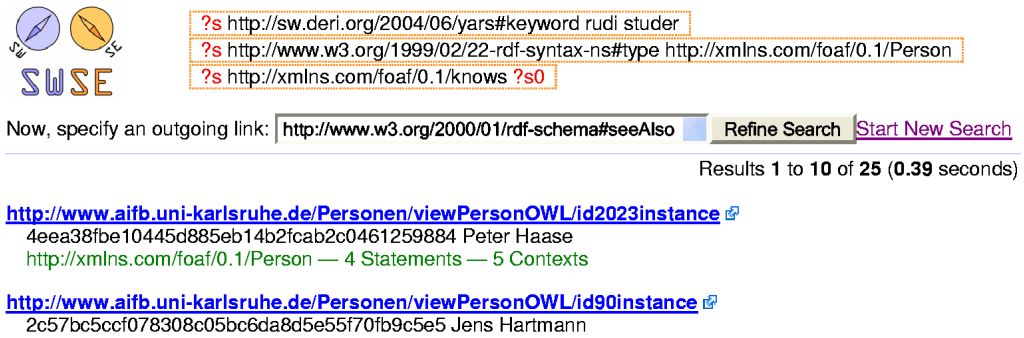

Figure 2: SWSE prototype screenshot

XML) into RDF, to allow for further integration. As the web of crawled data expands, the collected graph becomes more densely interlinked as URIs match up, e.g., the crawler fetches and extracts the title of a HTML page which is a `homepage` property of a person in the graph. Thus, we observe better connectivity in our data graph versus crawling of the RDF web alone.

Our computer science research dataset also includes the DBLP and CiteSeer knowledge bases. These required manual conversion to RDF via XSLT. We also created a research publications ontology which we used as the common target specification, along with current specifications such as FOAF descriptions for authors.

The RDF data model allows for good data integration both on a schema and instance level. However in practice, on the instance level, there exists poor agreement on common URIs for the same entities across data sources. Using object consolidation, we merge equivalent instances under one identifier. We detect equivalence by analysing shared values of inverse functional properties for multiple instances. For example, we can merge equivalent `Person` instances from the web-crawl data and the DBLP author data through having the same value for the `homepage` property. In merging the instances, we improve in-degree and out-degree in the graph and move towards resolving the fragmentation of metadata on a particular entity over multiple instances.

For indexing the data we use JARS2; a hybrid indexing structure to provide both information retrieval and database query functionality necessary for SWSE. The information retrieval functionality stems from Apache Lucene, an inverted on-disk keyword index which returns relevant entities given a keyword query. For graph based lookups, we employ an optimised custom built quad index structure. We are currently finalising a distributed architecture which uses distributed hash tables and multiple Lucene indexes; distribution and parallelisation provide scalability and speed.

To score importance of results, we use ReConRank[3]. ReConRank is a links based analysis for deriving ranks of resources and contexts which incorporates tf-idf scores from Lucene with link weighting mechanisms. Thus, for example, papers with a high keyword query tf-idf match that are well-linked (e.g., cited) will appear higher in the results listings.

The web user interface offers an interaction model within the comfort zone of the current average user. A user begins by specifying a keyword query to narrow in on a spectrum of entities relevant to a topic. A list of summaries of the relevant entities are then displayed. The user can click on an entity in the list view and get all of the information available about that entity.

Additionally, the user is offered guided exploration. Users may restrict the type of entities in the results from a list of possibles, e.g., `Person, Publication`. They may then optionally traverse the data-graph through the relationships between entities, i.e., the inlinks or outlinks of the current entity(s) such as `knows` inlinks, `authorOf` outlinks; selection is from a list of possibles. Such traversal is offered from one or many current results. The possible restrictions and navigations offered comprise the guided incremental complex query creation system which we coin the 'nodebrowsing compass'.

Figure 2 shows a screenshot of the prototype system. At the top, the query trail is displayed and denotes that the user has issued the keyword query "rudy studer", restricted results to type `Person` and from these results has followed the `knows` relation outwards. Also shown in Figure 2 is the offer of possible outlinks to traverse through next. The user can click on any entity in the results pane, getting the entity's information and allowing the user to browse from that entity.

3. CONCLUSION

We have given the overview of a system to acquire, integrate, index and provide advanced search over information from a large number of data sources on the Web.

The user interface, which is supported by a scalable and distributable index structure, provides novel ways of exploring a graph-based dataset, including incremental query formulation. An early prototype of the system is available online at `http://swse.deri.org/`.

4. REFERENCES

[1] A. Harth and S. Decker. Optimized Index Structures for Querying RDF from the Web. In *Procs of 3rd Latin American Web Congress*, 2005.

[2] A. Harth, J. Umbrich, and S. Decker. MultiCrawler: A Pipelined Architecture for Crawling and Indexing Semantic Web Data. In *5th International Semantic Web Conference*, 2006.

[3] A. Hogan, A. Harth, and S. Decker. ReConRank: A Scalable Ranking Method for Semantic Web Data with Context. In *2nd Workshop on Scalable Semantic Web Knowledge Base Systems*, 2006.