SWSE: Objects before documents!

Andreas Harth, Aidan Hogan, Jürgen Umbrich, and Stefan Decker

National University of Ireland, Galway
Digital Enterprise Research Institute
firstname.lastname@deri.org

Abstract. Web search engines are immensely useful for locating documents online. However, with more and more structured data being published online, the restriction to the hyperdocument model impairs the usefulness for searching and browsing. In contrast, an object-orientated model provides means to firstly integrate data about the same object from multiple sources, and secondly enable expressive queries over the integrated information space. We present SWSE, a search engine over 1.1 billion statements published on the Semantic Web. The system provides an easy-to-use end-user interface through which users can find and navigate an object-orientated information space. In addtion, the system exposes the data via a full SPARQL REST service which is open for application developers to query and integrate data in own applications.

1 Introduction

While the major part of the current web consists of hyperdocuments, there has been the (implicit) realisation of the web community at large that a more formal knowledge representation model might make sense. An indicator for the desire towards a higher abstraction level is the emergence of social media and social networking sites concerned with people, databases about companies and organisations such as CrunchBase[1], product[2] comparison sites such as Ciao [3], location-aware sites such as Rummble [4], or events databases such as Upcoming[5].

We see the potential of Semantic Web technologies helping to achieve a networked and integrated information space which operates on the abstraction level of objects (people, organisations, products, locations, events) rather than documents. The benefit of using the object abstraction is two-fold: first, objects can be described in multiple documents, and an object-orientated view allows to aggregate and integrate information about the same real-world object from multiple sources. Second, given more structured data, complex queries are possible,

[1] For a Semantic Web-compatible version of TechCrunch's company database visit http://cb.semsol.org/

[2] An ontology covering products can be found at http://www.heppnetz.de/projects/goodrelations/

[3] http://www.ciao.com/

[4] http://www.rummble.com/

[5] http://upcoming.yahoo.com/

and developers can easily re-use query result in own applications. As a consequence, search engines need to offer improved capabilities for handling structured datasets.

To demonstrate the utility of an object-orientated abstraction for web search, we present an updated version of SWSE, our Semantic Web Search Engine research prototype, operating on 1.1 billion statements from over 6.5 million sources. SWSE's user interaction model is close to the model of current web search engines. To locate objects of interest, users just have to provide a few keywords and in a matter of seconds SWSE returns a list of relevant objects. From the result list, users can then teleport to relevant objects which are displayed in a detail view. From here, users again can teleport to related objects. In addition to the end user interface, SWSE provides an API that allows application developers to pose complex queries, which surpass in expressivity the type of queries that search engines and even dedicated data APIs are offering.

In the remainder of the paper, we introduce the search and navigation functionality offered to end users, describe the query endpoint, introduce the architecture and the data set derived from the billion triple challenge data set, review related systems, and conclude with an outlook to future work.

2 Searching and Navigating Objects

The user interface of SWSE is similar to the user interfaces of traditional search engines to leverage the familiarity of web users with these systems. We exted the browsing model with functionality to navigate and explore the result objects. From a high-level perspective, SWSE allows users to perform two operations:

- Locate objects of interest via keyword search
- Navigate an aggregated view of objects on-site, or navigate to external pages.

In the following we briefly describe each operation and provide an example.

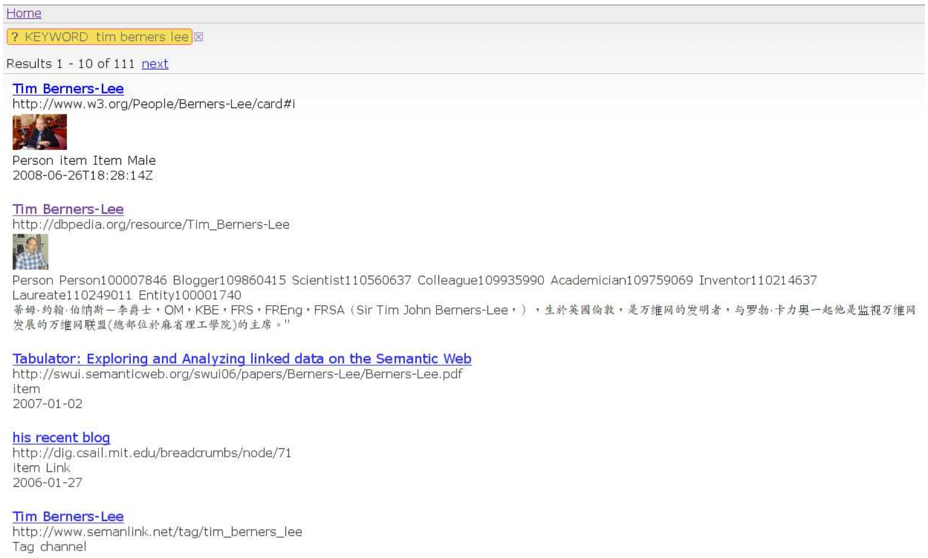
2.1 Searching Objects

The searching objects feature is similar to locating documents in hypertext search engines. Given a set of keywords the user sees first a list of the top ten matching objects. Figure 1 shows a screenshot of the results for the query "tim berners lee".

Beside normal pagination functions to navigate through all returned objects the user can access all available information for a given result object by clicking on the object.

2.2 Teleporting

The object navigation feature ("teleporting") is very similar to following a hypertext link, the only difference being that in SWSE the user is staying on-site.



Home

? KEYWORD tim berners lee ☒

Results 1 - 10 of 111 next

Tim Berners-Lee
http://www.w3.org/People/Berners-Lee/card#i
Person item Item Male
2008-06-26T18:28:14Z

Tim Berners-Lee
http://dbpedia.org/resource/Tim_Berners-Lee
Person Person100007846 Blogger109860415 Scientist110560637 Colleague109935990 Academician109759069 Inventor110214637 Laureate110249011 Entity100001740
蒂姆·约翰·伯纳斯－李爵士，OM，KBE，FRS，FREng，FRSA（Sir Tim John Berners-Lee，），生於英國倫敦，是万维网的发明者，与罗勃·卡力奧一起他是监视万维网发展的万维网联盟(總部位於麻省理工學院)的主席。"

Tabulator: Exploring and Analyzing linked data on the Semantic Web
http://swui.semanticweb.org/swui06/papers/Berners-Lee/Berners-Lee.pdf
item
2007-01-02

his recent blog
http://dig.csail.mit.edu/breadcrumbs/node/71
item Link
2006-01-27

Tim Berners-Lee
http://www.semanlink.net/tag/tim_berners_lee
Tag channel

Fig. 1. Search results page

The information displayed about an object emanates from index, whereas on the web people jump from server to server. Keeping people on-site is necessary to be able to pre-process and integrate data from multiple sources, and provide reasonable response times for queries (performing the data integration during runtime is too costly and slow).

Figure 2 shows a screenshot of the detailed view for the object Galway.

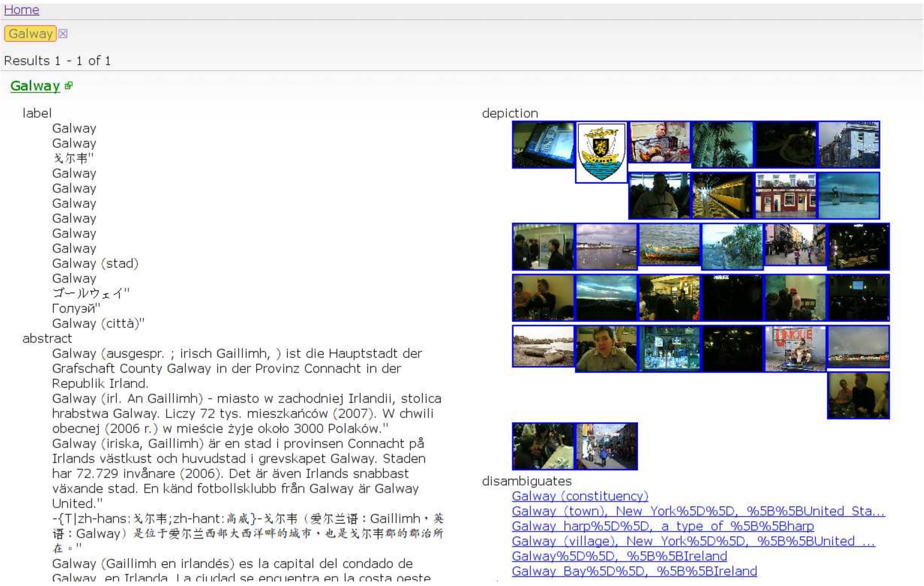
3 Query Endpoint

Besides the graphical user interface which allows users to explore the knowledge base, SWSE provides a SPARQL endpoint to execute complex queries. We provide a REST service to access the query processing functionality. To avoid overloading the servers, each query has an allotted 20 seconds processing time. The SPARQL endpoint allows application developers to process the results from the SWSE index in their own systems.

4 Dataset and System Setup

From the billion triple challenge datasets[6] we used the Falcon, Swoogle, Watson, SWSE-1 and SWSE-2 and DBpedia datasets. We excluded the other datasets

[6] http://www.cs.vu.nl/~pmika/swc/btc.html



Home

Galway ☒

Results 1 - 1 of 1

Galway

label

Galway
Galway
戈尔韦"
Galway
Galway
Galway
Galway
Galway
Galway
Galway (stad)
Galway
ゴールウェイ"
Голуэй"
Galway (città)"

abstract

Galway (ausgespr. ; irisch Gaillimh,) ist die Hauptstadt der Grafschaft County Galway in der Provinz Connacht in der Republik Irland.
Galway (irl. An Gaillimh) - miasto w zachodniej Irlandii, stolica hrabstwa Galway. Liczy 72 tys. mieszkańców (2007). W chwili obecnej (2006 r.) w mieście żyje około 3000 Polaków."
Galway (iriska, Gaillimh) är en stad i provinsen Connacht på Irlands västkust och huvudstad i grevskapet Galway. Staden har 72.729 invånare (2006). Det är även Irlands snabbast växande stad. En känd fotbollsklubb från Galway är Galway United."
-{T|zh-hans:戈尔韦;zh-hant:高威}-戈尔韦（爱尔兰语：Gaillimh，英语：Galway）是位于爱尔兰西部大西洋畔的城市，也是戈尔韦郡的郡治所在。"
Galway (Gaillimh en irlandés) es la capital del condado de Galway, en Irlanda. La ciudad se encuentra en la costa oeste

depiction

disambiguates

Galway (constituency)
Galway (town), New York%5D%5D, %5B%5BUnited Sta...
Galway harp%5D%5D, a type of %5B%5Bharp
Galway (village), New York%5D%5D, %5B%5BUnited ...
Galway%5D%5D, %5B%5BIreland
Galway Bay%5D%5D, %5B%5BIreland

Fig. 2. Details page for the object http://dbpedia.org/resource/Galway.

because of either licensing restrictions, restrictions in crawling, or the necessity of manual data conversion. From these seed data sets (around 450m triples) we extracted all unique URIs and dereferenced the URIs with the MultiCrawler framework [5] in mid-June 2008. The resulting data set contains 6.5m sources with a total of 1.1b RDF statements.

The current system for the billion triple challenge is distributed over six machines with a single Operton 2.2 GHz CPU, four GB of main memory and two 160GB SATA disks. Four machines are used to distribute the indexed RDF statements, one machine hosts the query processor component that enables efficient and distributed SPARQL queries and another machine runs the end-user interface.

5 Related Systems

Semantic web search systems can be broadly classified into two categories: systems that operate on a document abstraction (such as Swoogle [4] and Sindice [9]), which utilise algorithms and indices inspired by information retrieval research, and systems that operate on an object-orientated model (such as SWSE and Falcons [2]), which follow a more data-oriented tradition. While both approaches have their merits, we think that a distinguishing feature of search engines for the Semantic Web should be the view towards handling objects.

In contrast to KIM[7] which provides an infrastructure for information extraction, we operate on information that has been published in RDF. KIM has a end user interface that allows to pose queries beyond keyword search. Similarly, TcruziKB[8] provides complex query construction faciltites (inspired by GRQL[1]) over manually integrated and converted data, and is restricted to an enterprise search scenari. Asio[3] also has an enterprise search focus. In contrast, SWSE aims at domain-independent data aggregated from the Web; we mimic a Web search engine since in our experience users have issues with more complex input forms.

6 Conclusion

SWSE is search engine for data with a search interface[7] that enables a object-orientated result navigation, and a SPARQL endpoint[8] that allows application developers to access the indexed information via complex queries. With the modular design of the SWSE components the system can be efficiently distributed across several machines and scale up to store and retrieve several billion statements [6].

Areas of future study includes reasoning over the database[9] and more powerful selection and navigation functionality to be able to not only offer objects instead of documents but, ultimately, answers to complex information needs.

References

1. N. Athanasis, V. Christophides, and D. Kotzinos. Generating on the fly queries for the semantic web: The ics-forth graphical rql interface (grql). In *International Semantic Web Conference*, pages 486–501, 2004.
2. G. Cheng, W. Ge, and Y. Qu. Falcons: searching and browsing entities on the semantic web. In *WWW '08: Proceeding of the 17th international conference on World Wide Web*, pages 1101–1102, New York, NY, USA, 2008. ACM.
3. M. Dean. Suggestions for semantic web interfaces to relational databases. In *W3C Workshop on RDF Access to Relational Databases*, October 2007.
4. L. Ding, T. Finin, A. Joshi, R. Pan, S. R. 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 conference on Information and knowledge management*, pages 652–659, New York, NY, USA, 2004. ACM Press.
5. A. Harth, J. Umbrich, and S. Decker. Multicrawler: A pipelined architecture for crawling and indexing semantic web data. In I. F. Cruz, S. Decker, D. Allemang, C. Preist, D. Schwabe, P. Mika, M. Uschold, and L. Aroyo, editors, *International Semantic Web Conference*, volume 4273 of *Lecture Notes in Computer Science*, pages 258–271. Springer, 2006.
6. A. Harth, J. Umbrich, A. Hogan, and S. Decker. Yars2: A federated repository for querying graph structured data from the web. In 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, editors, *ISWC/ASWC*, volume 4825 of *Lecture Notes in Computer Science*, pages 211–224. Springer, 2007.

[7] http://swse.deri.org/

[8] http://swse.deri.org/yars2/

[9] We will demonstrate a dataset with materialised inferences at ISWC

7. A. Kiryakov, B. Popov, I. Terziev, D. Manov, and D. Ognyanoff. Semantic annotation, indexing, and retrieval. *Journal of Web Semantics: Science, Services and Agents on the World Wide We*, 2(1):49–79, 2004.
8. P. N. Mendes, B. McKnight, A. P. Sheth, and J. C. Kissinger. Tcruzikb: Enabling complex queries for genomic data exploration. *International Conference on Semantic Computing*, 0:432–439, 2008.
9. E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello. Sindice.com: A document-oriented lookup index for open linked data. *International Journal of Metadata, Semantics and Ontologies*, 3(1), 2008.