Interactive Exploration of Heterogeneous Cultural Heritage Collections

Michiel Hildebrand

CWI, Amsterdam, The Netherlands
michiel.hildebrand@cwi.nl

1 Research Problem

In this research we investigate to what extent explicit semantics can be used to support end users with the exploration of a large heterogeneous collection. In particular we consider cultural heritage, a knowledge-rich domain in which collections are typically described by multiple thesauri. Many institutions have made or are making (parts of) their collections available online. The cultural heritage community has the ambition to make these isolated collections and thesauri interoperable and allow users to explore cultural heritage in a richer environment.

The MultimediaN E-Culture project [1] examines the usability of Semantic Web technology to integrate museum data and to provide effective user interfaces to access this heterogenous data. The project has collected data from multiple museum collections annotated with multiple thesauri. Based on the procedure described in [2] this data is converted to RDFS/OWL and enriched with links across collections and thesauri. The result is represented in a single repository in the form of a large RDF graph. While some of these links have formal semantics as defined by RDFS/OWL, the majority only has "weak" semantics as defined by SKOS[1] and domain specific schemas.

Within the context of the E-Culture project, our research aims at better interfaces and search functionality to support end users with the exploration of large heterogeneous RDF graphs. Here we face two general problems. First, in a heterogeneous graph we have no fixed schema on which we can base the interface design and application functionality. Second, the semantics of our domain are too weak to depend on formal reasoning alone and thus we require alternative strategies that benefit from "weak" semantic structures to provide the required search functionality.

We focus on three types of end user functionality. First, searching for terms within multiple thesauri to support manual annotation. Second, keyword search, as it has become the de-facto standard to access data on the web. Third, faceted browsing as it has become a popular method to interactively explore (image) collections. We investigate the use of explicit semantics to improve support to the user in these three tasks. We propose the following research questions:

[1] http://www.w3.org/2004/02/skos/

A. Sheth et al. (Eds.): ISWC 2008, LNCS 5318, pp. 914–919, 2008.

- How can explicit semantics be used in search algorithms to support the user with finding results in a heterogeneous graph?
- How do we organize and visualize the results found in a heterogenous graph to support exploration of this graph?
- How can we evaluate the added value of using explicit semantics in search, result organization and visualization?

2 Approach

We investigate interactive exploration in heterogeneous collections by the implementation and evaluation of three prototype systems on top of large and real world data collections. First, an annotation interface that uses autocompletion to support users with finding terms from multiple thesauri. Second, a search interface that supports the user in exploring museum objects that are semantically related to a keyword query. Third, a facet browser to support the user with the interactive formulation of queries. For all three we investigate how to improve interaction by using explicit semantics in the search algorithm, the result organization and visualization. In the following subsections we describe the details of our approach for each prototype. Due to limited space we do not elaborate on the related work for each. Note, we performed an extensive survey of semantic search applications (see work plan).

For each prototype system we propose an evaluation method. Although several applications use Semantic Web technology to support some form of exploration there are, as yet, no standard metrics to evaluate the solutions. This is probably because the aims, user tasks they intent to support and the use of semantics vary greatly among different applications. Determining appropriate evaluation methods for different forms of semantic search is an intrinsic part of this research.

2.1 Finding Terms within Multiple Thesauri

In an annotation task the user describes an object using terms from domain-specific thesauri. An annotation interface may contain annotation fields for the different properties that should be described. Typically, some form of keyword search makes the thesauri terms accessible. As thesauri become available in an interoperable format, annotators can access terms from multiple sources, including sources provided by other institutions.

At the moment there are several Semantic Web search engines that give access to terms from RDF/OWL documents. Examples of semantic search engines are Swoogle [3] and Sindice [4]. An elaborate analysis of 35 semantic search applications is available in [5]. Generic semantic search engines are not yet suited to support annotators in finding domain specific terms. A search query may return results irrelevant for a particular type of annotation. Furthermore, terms in the result set can be ambiguous in the sense that a naive visualization of these terms would not allow a user to distinguish them from each other.

Could an interface provide large coverage by using multiple thesauri while providing effective organization and visualization of the search results? The user

can be supported in the annotation process by presenting only the terms that are appropriate for an annotation field and these terms should be organized and visualized so that they are unambiguous and self explanatory. For example, we can constrain the results of the creator field to persons, and augment the visualization of person names with their birth and death date; an annotation field for the creation site can be constrained to geographical locations, and these locations can be organized in a grouping by the country.

To experiment with different configurations of search, result organization and visualization we implement a configurable term search component. The component uses autocompletion to suggest terms while the user is typing. Based on this autocompletion search component a prototype annotation interface is constructed to support the subject annotation of the Rijksmuseum Amsterdam print collection. The prototype annotation interface will give access to terms from multiple thesauri including thesauri from outside the Rijksmuseum.

Evaluation. The prototype annotation interface is evaluated qualitative with experts at the Rijksmuseum. First, we gather information about the current annotation practices at the Museum. Second, we iteratively design a prototype interface considering the feedback of the experts in each cycle. Third, the professional annotators use the prototype interface in an experiment situated in their own environment. The results of the evaluation consist of observations that impact the practical use of Semantic Web technologies in the search algorithm and result visualization and organization.

2.2 Exploring Heterogeneous Collections through Keyword Search

A common way to start the exploration of the objects in museum collections is through keyword search. The simple "Google like" interface, with a single text-entry box, has become the standard for keyword search interfaces. Using Semantic Web technology we can use keyword search functionality to find museum objects that are semantically related to a query.

Within the E-Culture project we explore graph search algorithms to efficiently perform semantic on a large collection [1]. The results of a graph search are related to the keyword query by a path in the graph, that reflects a possible semantic interpretation of the query. Hollink et. al. showed through statistical analysis that some patterns of graph paths, containing relations from WordNet, performed better than others [6]. In an heterogeneous collection it is, however, difficult to determine in advance all paths that will lead to relevant results.

An interactive interface would let the user choose which interpretation she is interested in. Explicit semantics can be used to organize and visualize the search results to support the user with disambiguation of the search results as well as with the exploration of semantically related objects. We develop a prototype interface that allows interactive exploration of semantic keyword search results. The interface will use the graph-based search algorithms as developed within the E-Culture project. We investigate how explicit semantics can be used to organize the search results. In addition, we explore different visualizations for the presentation of the search results, such as geographical maps and timelines.

Evaluation. Different types of semantic organization techniques will be evaluated in a user study using an interactive exploratory search task. Designing the details of the experimental setup is part of the future work. We strive to use objective measurements, such as click stream data, as well as subjective opinions to determine the satisfaction of discovering (new) museum objects and relations.

2.3 Exploring Heterogeneous Collections with Faceted Browsing

Faceted browsing has become popular as an interface to interactively formulate queries [7,8,9,10]. A single facet highlights a dimension of the underlying data. By visualizing the values of the facets in the user interface, the user can construct multi-faceted queries by navigating through the interface.

Marti Hearst et al. showed that faceted browsing is very well suited to explore a collection of visual resources [7]. They assume a homogeneous collection with a fixed data schema, which allows manual configuration of the facets. Using a prototype implementation we investigate the requirements to apply faceted browsing to a large heterogenous collection. We explore the use of explicit semantics to organize and visualize the many facets of a heterogeneous collection.

Evaluation. A problem with faceted browsing is that large join queries can not always be computed sufficiently fast to support reasonable response times. We evaluate the scalability of faceted browsing and investigate how caching mechanisms can be used to improve response times. We define the theoretical and practical scalability limits, and experimentally, we show the performance statistics of different types of realistic queries on a large real world data set.

3 Contributions

1. Requirements analysis on the semantic data, search algorithms and user interface needed to support annotation using multiple thesauri. Concrete, implementation of an interface along with the underlying algorithms that support efficient term search from multiple thesauri by professional annotators. The Web-based implementation is based on a more generic interface model for term lookup in heterogeneous thesauri. The algorithms provide term search, result organization and visualization and can be configured with domain-specific semantics.
2. Design of an interface and algorithms to support end users with the exploration of a heterogeneous collection. The interface provides keyword search, semantic-based organization of the search results as well as different visualizations. The algorithms provide graph search to efficiently find objects semantically related to a keyword query and result organizing techniques that can be configured with domain specific semantics.
3. Design of a scalable interface and efficient query engine to apply faceted browsing to heterogeneous RDF graphs. The facets can be automatically or manually configured using domain specific semantics. The query engine provides caching mechanism to efficiently support large join queries.

4 Work Plan

Bellow we briefly describe the work that we have achieved so far, that we are currently working on and that is planned for future work. Between brackets we mention to which contribution the work is related.

Results achieved so far
(all) Studied related work in semantic search. In 35 existing systems we analyzed how explicit semantics are used in query construction, the core search process, the presentation of the search results and user feedback on query and results. (To appear as a chapter in the book for the Network of Excellence, K-Space)
(1) Interface design for a configurable autocompletion component. We implemented a client-side autocompletion widget in JavaScript on top of the YAHOO User Interface (YUI) library. We also implemented a server-side algorithm for term search and result organization that can be configured with domain specific semantics. (CWI technical report INS-E0708)
(1) User study on result organization techniques for autocompletion suggestions. In cooperation with Alia Amin we conducted two user studies using web-based interactive surveys to test different methods of grouping term suggestions. (to be submitted)
(2) Design of a search algorithm for semantic search. In cooperation with Jan Wielemaker we implemented a best-first weighted graph search algorithm in Prolog. (Accepted for ISWC 2008)
(2) Initial interface design for organization and visualization of search results. We implemented a client side widget in JavaScript that can visualize a set of results as groups of thumbnails, on a geographical map, timeline or a graph. We also implemented server side algorithms for result organization and visualization of RDF data that can be configured to use domain specific semantics. (Intermediate results are part of the ClioPatria open source toolkit[2])
(3) Interface design for faceted browsing on a heterogeneous RDF graph. The prototype system, /facet, is, for example, used within the E-Culture Demonstrator[3], the K-Space news demonstrator[4]. (Published at ISWC 2006)

Current work
(1) Design and configuration of the interface to support subject annotation of the print collection of the Rijksmuseum Amsterdam.
(1) User study of the prototype annotation interface with professional annotators of the Rijksmuseum Amsterdam.
(3) Evaluation of scalability in /facet. Test caching solutions to improve computation of large join queries.

Planned work
(2) Continuation of interface design to support interactive exploration of search results with semantic clustering.
(2) Experimental design and user study on result clustering for semantic search.

[2] `http://e-culture.multimedian.nl/software/ClioPatria.shtml`
[3] `http://e-culture.multimedian.nl/demo/search`
[4] `http://newsml.cwi.nl/explore/facet`

Acknowledgments

I would like to thank Alia Amin, CWI Amsterdam and Jan Wielemaker, UvA Amsterdam for their close cooperation, my supervisors Jacco van Ossenbruggen, CWI Amsterdam and Lynda Hardman, CWI Amsterdam and TU Eindhoven for guiding me in this research trajectory and Guus Schreiber, VU University Amsterdam for his support of my research.

This research was supported by the MultimediaN project funded through the BSIK programme of the Dutch Government and by the European Commission under contract FP6-027026, Knowledge Space of semantic inference for automatic annotation and retrieval of multimedia content — K-Space.

References

1. Schreiber, G., Amin, A., van Assem, M., de Boer, V., Hardman, L., Hildebrand, M., Hollink, L., Huang, Z., van Kersen, J., de Niet, M., Omelayenjko, B., van Ossenbruggen, J., Siebes, R., Taekema, J., Wielemaker, J., Wielinga, B.: MultimediaN E-Culture Demonstrator. In: Cruz, I., Decker, S., Allemang, D., Preist, C., Schwabe, D., Mika, P., Uschold, M., Aroyo, L.M. (eds.) ISWC 2006. LNCS, vol. 4273, pp. 951–958. Springer, Heidelberg (2006)
2. Tordai, A., Omelayenko, B., Schreiber, G.: Thesaurus and metadata alignment for a semantic e-culture application. In: K-CAP 2007. Proceedings of the 4th international conference on Knowledge capture, pp. 199–200. ACM, New York (2007)
3. Ding, L., Pan, R., Finin, T., Joshi, A., Peng, Y., Kolari, P.: Finding and Ranking Knowledge on the Semantic Web. In: Gil, Y., Motta, E., Benjamins, V.R., Musen, M.A. (eds.) ISWC 2005. LNCS, vol. 3729, pp. 156–170. Springer, Heidelberg (2005)
4. Tummarello, G., Oren, E., Delbru, R.: Sindice.com: Weaving the open linked data. In: Aberer, K., Choi, K.-S., Noy, N., Allemang, D., Lee, K.-I., Nixon, L., Golbeck, J., Mika, P., Maynard, D., Mizoguchi, R., Schreiber, G., Cudré-Mauroux, P. (eds.) ASWC 2007 and ISWC 2007. LNCS, vol. 4825, pp. 547–560. Springer, Heidelberg (2007)
5. Hildebrand, M., van Ossenbruggen, J., Hardman, L.: An analysis of search-based user interaction on the Semantic Web. Technical Report INS-E0706, CWI (2007)
6. Hollink, L., Schreiber, G., Wielinga, B.: Patterns of semantic relations to improve image content search. Web Semant 5(3), 195–203 (2007)
7. Yee, K.P., Swearingen, K., Li, K., Hearst, M.: Faceted Metadata for Image Search and Browsing. In: CHI 2003. Proceedings of the SIGCHI conference on Human factors in computing systems, Ft. Lauderdale, Florida, USA, pp. 401–408. ACM Press, New York (2003)
8. Hyvonen, E., Makela, E., Salminen, M., Valo, A., Viljanen, K., Saarela, S., Junnila, M., Kettula, S.: Museumfinland – finnish museums on the semantic web. Journal of Web Semantics 3(2), 25 (2005)
9. Schraefel, M.C., Smith, D.A., Owens, A., Russell, A., Harris, C., Wilson, M.L.: The evolving mSpace platform: leveraging the Semantic Web on the Trail of the Memex. In: Proceedings of Hypertext 2005, Salzburg, pp. 174–183 (2005)
10. Huynh, D., Karger, D., Miller, R.: Exhibit: Lightweight structured data publishing. In: 16th International World Wide Web Conference, Banff, Alberta, Canada. ACM, New York (2007)