Bridging Semantic and Legacy Web Exploration: Orientation, Revisitation and Result Exploration Support

Michal Tvarožek
Institute of Informatics and Software Engineering
Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
Email: tvarozek@fiit.stuba.sk

Mária Bieliková
Institute of Informatics and Software Engineering
Slovak University of Technology
Ilkovičova 3, 842 16 Bratislava, Slovakia
Email: bielik@fiit.stuba.sk

***Abstract*—In order to address issues such as information overload and the navigation problem, which plague users on the Web, we need to improve user support for query construction, modification, result browsing and information exploration. The Semantic Web aimed to address many of these issues by providing machine processable information and application interoperability, but as of today failed to reach widespread acceptance. We build upon our previous work with Semantic Web exploration and propose a browsing solution which allows users to browse both legacy and semantic web information transparently while taking advantage of advanced exploration features provided by our personalized faceted semantic browser. In this paper, we describe our approach to integrating legacy web content into our semantic browser via web crawling and page annotation, and orientation and revisitation support including tree-based history visualization and incremental graph-based result exploration.**

***Keywords*-Exploratory search, Web, Semantic Web, interface generation, revisitation support, graph visualization**

I. Introduction

Authors often cite information overload, the infamous navigation problem or query complexity as main issues affecting web users. In practice however, these are "just" the consequences of the lack of support for the three primary actions users perform during typical web search sessions:

- *query construction and refinement*,
- *result browsing and selection*,
- *result exploration and understanding*.

Based on the specific search intent—informational, navigational, transactional—as defined by Broder [1], system support might focus on different actions. Informational queries might stress query construction to get the best results, while navigational searches would quickly find good starting points but also provide result exploration support, e.g. in terms of showing user trails on target sites.

These issues are even more pronounced in the Semantic Web environment, as semantic query construction is a highly complex task requiring not only the knowledge of semantic query languages (e.g., SPARQL) but also knowledge about domain concepts to use in the query (e.g., URIs of classes). Similarly, the browsing of results or their exploration must address the fact that Semantic Web information has no default visualization as opposed to typical HTML pages.

II. Related Work

Current approaches for the legacy Web range from simple query auto-complete functionality (e.g., in Google) to sophisticated query expansion and disambiguation solutions performing clickstream analysis and data mining [2], [3], while support for result browsing and selection is much more limited to snippets or more scarcely simple ratings of results. Both these steps are usually performed or supported by a web search engine and thus their support is fairly widespread. Result exploration support however is virtually non-existent as it would require individual web sites to have been designed and developed to provide user support.

Mayer provides a broad survey of existing history and revisitation approaches, along with open problems including acquisition, search and visualization of history entries and metadata [4]. While current browser and search engine extensions support features such as full-text search in history (e.g., the Firefox plugin WebMynd) or tree-based history visualization more suited to the recursive nature of web navigation (e.g., the Firefox plugin HistoryTree, Pad Tree or WebView [4]), users still encounter issues with keyword guessing, disorientation and dead links.

Exploratory search approaches [5] aim to provide support in all steps of the search process, but usually focus on exploration of closed information spaces, thus making support during wild web exploration scarce. Kules et al. examined how users used faceted browsers and found that facets were an integral part of the exploration experience accounting for about one half of the time spent on actual search results [6].

VisGets is an advanced visualization and querying solution for legacy web data [7]. It crawls the web and gathers news articles, and in turn enables users to explore the data based on three dimensions – time, location and topic. It does not however provide any kind of social recommendation support nor supports navigation after selecting a search result (i.e., once the user leaves the original search engine).

Similarly in the Semantic Web context, Tabulator enables users to browse Linked Data [8]. While Tabulator enables users to take advantage of different visualizations (e.g., map, calendar), it offers only very limited search support. Other Semantic Web browsers / query builders such as Disco Hyperdata browser or Zitgist Dataviewer offer even less user support and are thus useful only to experts.

III. Unified "NextGen" Browser Enhancements

We previously devised a unified "NextGen" web browser concept, which focused on end-user experience by integrating access to and interaction with legacy Web and Semantic Web content via a generated faceted browser interface [9].

In this paper, we extend our original approach with additional user support for *revisitation tasks* and *result exploration*. While our original approach allowed users to browse regular web pages in the same browser as true semantic content, it was unable to search those pages via the faceted semantic browser. We now improve upon our original concept by providing a *lightweight semantics extraction* approach for legacy web content that crawls web sites, which can then be searched as if they were any other semantic content in our faceted semantic browser.

A. Legacy web content integration

In order to enable true semantic exploration of legacy web content we devised a lightweight semantics acquisition approach on the page-level (i.e., we do not try to extract and link individual objects within a page). We gather:

- *content-related metadata*, which is derived from actual page content using term extraction algorithms,
- *usage-related metadata*, which is based on how users browse the specific site.

To acquire content metadata, we crawl web sites, identify page content stripped of banners, navigational menus and other "irrelevant" items. Next we index the pages via Lucene, and apply several metadata extraction approaches:

- *Metadata extraction* from page content using an external term extraction library, which also queries public bookmarking systems to identify existing tags.
- *Hierarchical classification extraction* from local navigation menus interlinking web pages within a site.
- *Annotation extraction* from incoming contextual links in page content.

We acquire usage data from an external proxy server, which improves web search via social-context driven query expansion based on user action tracking and evaluation [10].

Consequently, each page is indexed for fulltext search via Lucene, and has additional metadata describing its size, document type, recency, links to other pages, associated topics (also classified via external taxonomies, e.g., from Delicious), association to the local site hiearchy extracted from menus, annotations from incoming contextual links, and usage data (e.g., how many users visited the site, (anonymous) social relations to other users), which can be used for exploration via our faceted semantic browser.

B. Revisitation and orientation support

Our tree-based visualization of search and browsing history improves user orientation within complex navigation sessions and provides revisitation support between sessions. We continually record user actions performed within our browser (e.g., facet selections, result exploration) and construct a tree of query modifications and result visits (see Fig. 1). The tree is shown to users while they are browsing and also stored for future reference and processing.

We identify user agendas (i.e., goals users aimed to achieve) defined as a set of weighted terms related to individual sessions. We extract terms from queries and from visited results using term extraction approaches, and modify weights of extracted terms by the factor of user interest in the result, computed based on time spent on a result or after explicit bookmarking. We employ cosine similarity, with vectors consisting of weighted terms, multiplied by the factor of time elapsed between the last two actions to measure the distance between the actual agenda and a new query in order to distinguish different user agendas.

Next, we combine individual history trees into a single history map by merging common history tree nodes (e.g., result visits, queries). The history map covers a user's entire browsing history, with support for keyword search and personalized presentation (e.g., hiding less visited subgraphs).

C. Result exploration support

We provide result exploration support via a graph-based visualization of resource properties (see Fig. 2). The graph view is generated directly from a domain ontology showing individual resources and their relations, also taking advantage of relevance evaluation from the personalization engine.

Graph exploration starts with a single central node selected by the user (e.g. via facets or its URI). The view displays the selected resource and its properties (i.e., relations to other resources). We visualize both resources and properties as nodes to reduce information overload and to improve graph layout as a single property can connect multiple resources at the same time. We employ a force-based layout algorithm, but also allow the user to lock and manually move nodes in the resulting graph. Our graph view supports incremental horizontal exploration of resources, as users can move the view's focus to different nodes or further expand nodes to show their properties. To further improve user orientation, we use personalization to adapt the displayed properties and/or attributes, while also allowing users to manually customize the visible properties of resources.

IV. Validation

Our browser is a Silverlight application running inside existing web browsers, where it provides content rendering,

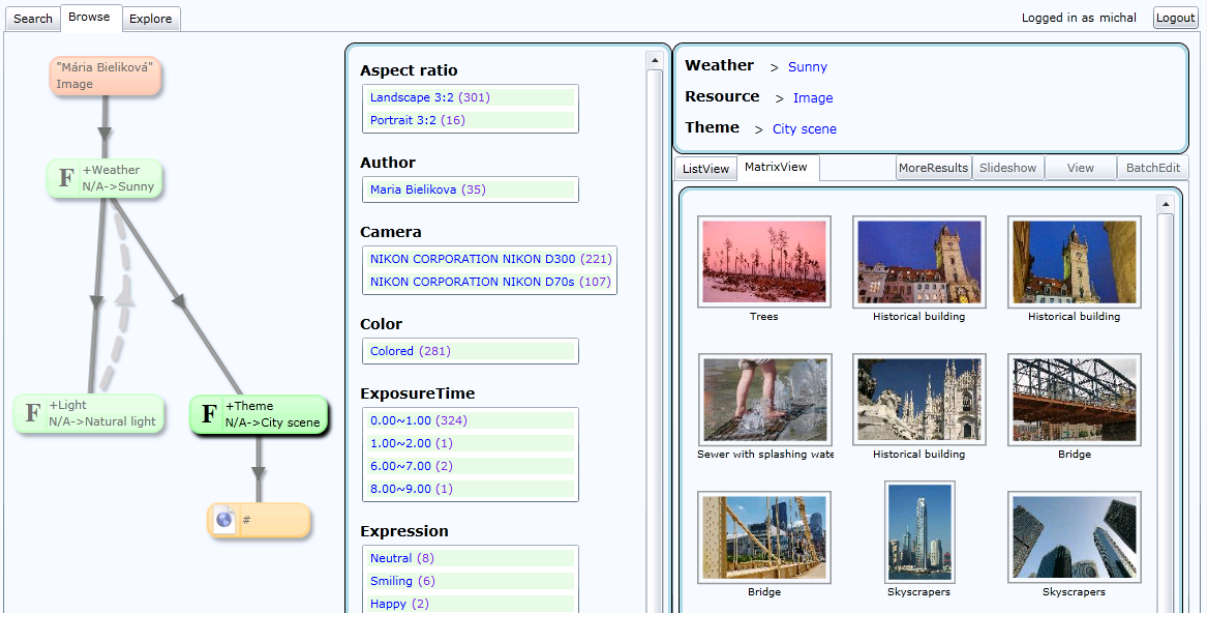

Figure 1. Example of our tree-based history visualization showing an initial keyword query (top left) and the successive faceted query refinements (left). The rest of the interface shows the list of available facets (center) and the list of search results (right).

personalization and plugin support (see Fig. 3). The browser handles all user interaction and acts as a front-end to server-side web (WCF) services, which serve as search providers, content providers or as support services. These include the Factic faceted search engine, the Steltecia service for ontological repository access, and the optional SemanticLog event logging service for global statistics tracking.

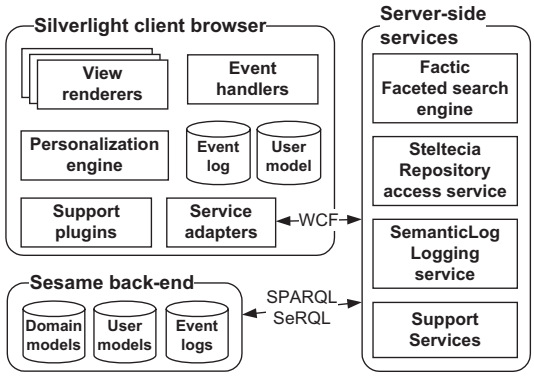

Figure 3. Architecture of our browser prototype.

We validate our approach via experiments with our faceted semantic browser prototype, which takes advantage of ontological representation of information artifacts, facets, restrictions, user preferences, and metadata describing legacy web pages in OWL. We work primarily with an image dataset containing about 8 000 manually and semi-automatically annotated images. Our web-based data set contains roughly 700 automatically annotated web resources crawled from our faculty web site. Since exact analytical validation of user-centered approaches is difficult, also considering the novelty of the exploratory search field and immaturity of methodologies for task design and browser evaluation [6], we focus on user studies and proof of concept validation of our individual approaches.

Our preliminary experiments showed that we can get sensible metadata by extracting web site navigational links, while our user study with graph-based visualization proved its viability with users who had no previous experience with graph-based exploration. Still, further evaluation is needed to confirm these results with a larger user group over a longer period of time in order for revisitation support to make sense.

We also evaluated the performance of our prototype, which despite some optimizations, shows a bottleneck in the Sesame ontological repository. Nevertheless, we still see room for improvement by caching results, reducing the number of requests per user click, e.g. by personalization.

V. Conclusions and Future Work

We extended our original faceted semantic browser with novel features aimed to support *end-user experience* with specific focus on *revisitation and orientation support*, *result exploration*, *enabling semantic legacy web site search and exploration*. To achieve these goals we devised methods for:

- *Tree-based visualization of search and browsing history*
- *Incremental graph-based search result exploration*
- *Lightweight semantic metadata acquisition from legacy web sites including (external) usage statistics tracking*

Our initial experiments have shown the viability of the proposed approaches for their intended purposes in terms of their practicality (i.e., it can be done) and improved user

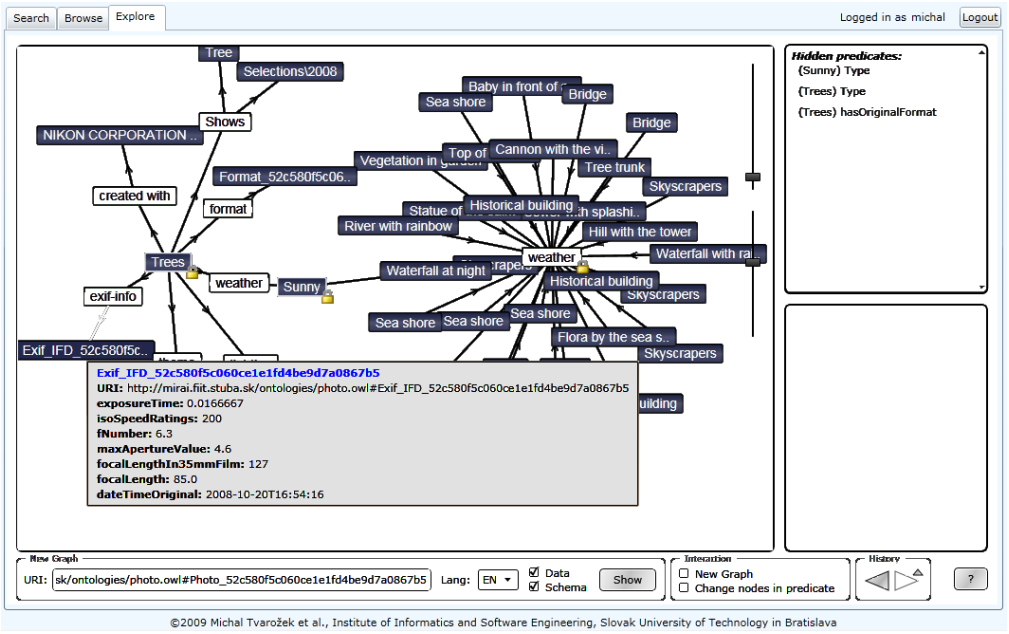

Figure 2. Example of our generated graph exploration interface. Dark nodes represent individual resources, white nodes correspond to relations (top). Hovering over nodes shows the attributes of a node (center); additional tools include zooming, spatial expansion, node hiding and history (right), with additional filtering options for languages and data/schema only visualization (bottom).

experience (i.e., improved task times, better understanding of the information space, efficient resource revisitation). Based on our experiments, we believe that we successfully addressed the original goal of our work:

To empower end-users with seamless access to both legacy web content and semantic information spaces by providing an end-user grade exploratory browser with support for effective query formulation, result overview browsing and individual result exploration.

Still, future work will include refining the proposed interface and more comprehensive experiments to validate the overall benefit of the proposed combination of approaches.

Acknowledgment

This work was supported by the Scientific Grant Agency of SR, grant No. VG1/0508/09, the Cultural and Educational Grant Agency of SR, grant No. 028-025STU-4/2010, and it is a partial result of the Research & Development Operational Program for the project Support of Center of Excellence for Smart Technologies, Systems and Services II, ITMS 25240120029, co-funded by ERDF.

References

[1] A. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

[2] G. Bordogna, A. Campi, S. Ronchi, and G. Psaila, "Query disambiguation based on novelty and similarity user's feedback," in *WI-IAT '09: Proc. of the 2009 IEEE/WIC/ACM Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology*. Washington, DC, USA: IEEE CS, 2009, pp. 125–128.

[3] P. t. Braak, N. Abdullah, and Y. Xu, "Improving the performance of collaborative filtering recommender systems through user profile clustering," in *WI-IAT'09: Proc. of Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology*. Washington, USA: IEEE CS, 2009, pp. 147–150.

[4] M. Mayer, "Web history tools and revisitation support: A survey of existing approaches and directions," *Foundations and Trends in HCI*, vol. 2, no. 3, pp. 173–278, 2009.

[5] G. Marchionini, "Exploratory search: from finding to understanding," *Com. of the ACM*, vol. 49, no. 4, pp. 41–46, 2006.

[6] B. Kules, R. Capra, M. Banta, and T. Sierra, "What do exploratory searchers look at in a faceted search interface?" in *JCDL '09: Proc. of the Joint Conf. on Digital libraries*. New York, NY, USA: ACM, 2009, pp. 313–322.

[7] M. Dörk, S. Carpendale, C. Collins, and C. Williamson, "Visgets: Coordinated visualizations for web-based information exploration and discovery," *IEEE Trans. on Visualization and Computer Graphics*, vol. 14, no. 6, pp. 1205–1212, 2008.

[8] T. Berners-lee, Y. Chen, L. Chilton, D. Connolly, R. Dhanaraj, J. Hollenbach, A. Lerer, and D. Sheets, "Tabulator: Exploring and analyzing linked data on the semantic web," in *In Proc. of the 3rd Int. Semantic Web User Interaction Workshop*, 2006.

[9] M. Tvarožek and M. Bieliková, "Reinventing the web browser for the semantic web," in *WI-IAT '09: Proc. of the Int. Joint Conf. on Web Intelligence and Intelligent Agent Technology*. Washington, DC, USA: IEEE CS, 2009, pp. 113–116.

[10] T. Kramár, M. Barla, and M. Bieliková, "Disambiguating search by leveraging a social context based on the stream of users activity," in *UMAP 2010*, ser. LNCS, 2010, accepted.