Reinventing the Web Browser for the Semantic Web

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*—The paper extends the traditional browser concept with a Semantic Web tailored faceted browser thus providing integrated end-user grade support for both legacy Web and Semantic Web content. The new browser provides interactive exploratory search and navigation capabilities as well as user adaptation and personalization. We describe the operation, usage scenarios and dependencies of the browser.**

***Keywords*-personalization; faceted exploration; semantic web**

I. Introduction

The idea of a powerful Semantic Web browser is probably as old as the idea of the Semantic Web itself. Still, no generic "product grade" Semantic Web browser has been created as of today, which partly reflects the chicken-and-egg problem it poses: will there first be abundant Semantic Web content or applications that consume it?

The creators of Piggy Bank—a Firefox extension that perhaps came the closest to being a somewhat practical semantic browser for web content—tried to break this chicken-and-egg problem by allowing users to extract data from different web sites and browse them via a simple faceted browser [1]. In practice, Piggy Bank downloaded RDF content associated with web pages or employed custom built screen scrapers to automatically extract RDF data from the pages directly. In addition to browsing the acquired data, it could also be stored for future reference or optionally published and shared with other users via a Semantic Bank server. Since Piggy Bank required custom built screen scrapers for every site, its use on the open Web was limited. Furthermore, it only worked on web pages directly visited by the user and thus could not be used for generic web search.

A lot of research from several fields has shown that faceted browsers are a suitable platform for exploratory search integrating both search and navigation. The information retrieval branch of research focuses on extensions of the faceted querying model [2] and automatic extraction of facet hierarchies from textual data [3]. The human-computer interaction perspective was explored in the pioneering Flamenco browser, which focuses on information exploration and usability evaluation in the digital image domain [4]. Generic construction of faceted browser interfaces proved to be a complex task and clustering was proposed to support automatic facet (i.e., hierarchy/classification) generation [5].

Semantic Web approaches such as represented by Sindice.com address data discovery, caching, indexing, inference and querying for documents containing semantic information about resources via keyword-, URI- and triple-based queries [6]. The mSpace explorer [7] leverages RDF data to provide a flexible column-based browsing interface similar to a faceted browser, while the /facet browser also automatically extracts and ranks facets from raw RDF data, primarily exploiting rdfs:subClassOf hierarchies [8]. Similarly, the BrowseRDF faceted browser generates facets from RDF data and defines several metrics used to identify useful facets [9].

Despite these results and the fact that several generic domain independent RDF browsers for SPARQL enabled repositories or linked data already exist (e.g., Disco Hyperdata browser, Tabulator, Zitgist Dataviewer), no *end-user grade* integrated semantic search and browsing solutions for the generic (Semantic) Web have been widely deployed.

II. Unified "NextGen" Web Browser

We propose a novel unified web browser concept which augments *end-user experience* by integrating access to and interaction with legacy Web and Semantic Web content via a generated faceted browser interface. We first describe our browser in terms of user experience, i.e. how a user—Alice—would employ its capabilities for an exploratory search task, and next elaborate on its design.

Alice needs to find papers relevant to her research so she starts her session using the general keyword-based search of our browser. Somewhat expectedly, most of the top results appear to be her own papers. Although normally Alice would try to *guess* better keywords which others might use to describe relevant results, she instead takes advantage of the search by example capability of the browser to see similar/related results and rank them via a positive example selecting one of her better papers. The browser returns a (large) mixed set of her papers, other papers and also various somewhat related results as returned by a back-end search engine. In order to filter out her own papers, she places a negative faceted restriction saying *not my papers*.

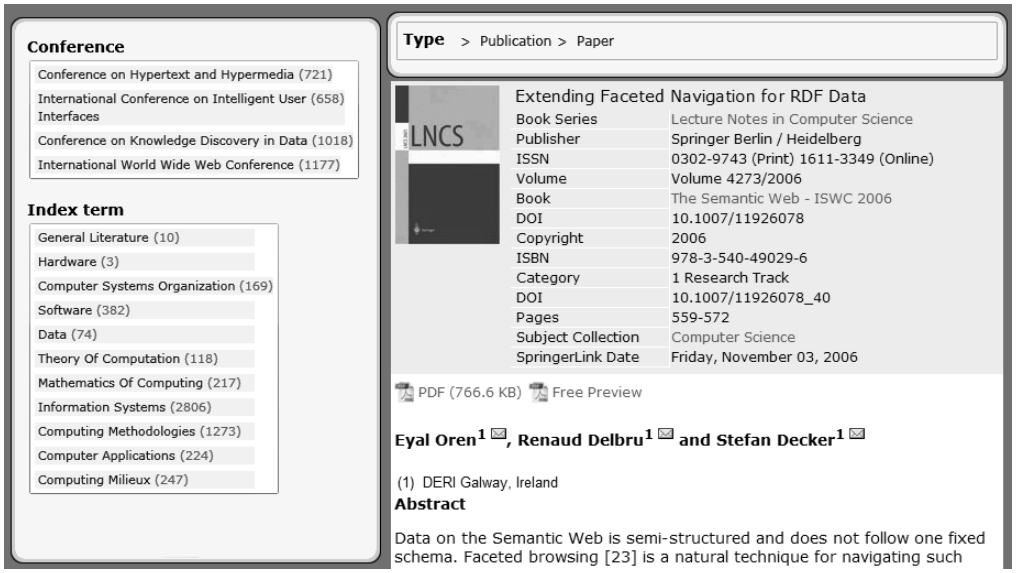

Figure 1. Example of our browser in the scientific publications domain showing legacy web content (right) with additional navigational options extracted as hierarchical facets from ACM index term annotations (left).

Looking at the results, Alice sees papers she had already read, digital library pages of new papers, bookstore sites, conference programs and some broken links. To reduce the number of irrelevant results, she employs negative search by example saying *no shops and no programs* (i.e., ranking those results low) while also restricting the results set to *not older than 4 years*. The browser returns an interesting looking paper on a digital library page warning Alice that she does not have an account to access the full paper. Since the paper is effectively unavailable, Alice rather explores another paper (described only by a bibliographic reference), which was recommended by the browser based on her social network data – Alice knows the authors personally.

This gave Alice an idea, which she decides to explore – how could she select all papers by all authors she knows to work in her field and the papers they reference? She starts by using nested facets to restrict the results to papers (see Fig. 1), then to papers that are authored by people she knows. Alice also adds papers authored by people whose papers are referenced by the people she knows. Lastly, since the browser tracks her profile, it recommends her to hide all the resources she had already seen in the past leaving her with a good set of results from relevant authors.

A. Semantic Browser Interface Generation

Our browser acts as an integrated tool for search when it acts like a search engine front-end, and for navigation when it supports navigation across a collection of "pure" information artifacts. In the Semantic Web these correspond to individual resources (or sets of resources) while in the legacy Web, they correspond to HTML pages stripped of non-essential parts such as hard-coded site navigation menus, banners, language selectors or external links. Since non-content parts of web pages often correspond to (navigational) metadata created by site authors to aid users in navigation we use these to acquire annotations describing the corresponding information artifacts. For example a hierarchical site navigation menu is considered to be a facet, while the links in non-content parts are used as annotations for the content present on the pages they link to (see Fig. 1).

We generate the browser interface (i.e., facets corresponding to patterns in metadata) based on estimated user characteristics and the user's current position in the information space taking advantage of both ontologies describing resources and the navigational metadata extracted from web pages. Note that we use the term *facet* somewhat loosely compared to its strict definition in library science. Still, this is why in our example Alice could use faceted search restrictions on arbitrary content while also having the appropriate restrictions available in the user interface in the first place.

Moreover, our ontological (metadata) description of the generated user interface allows us to exploit existing methods of personalization and social aspects to adapt the users' navigational experience in the open Web which was previously impossible due to the hard-coded nature of links. Also since many sites employ content management systems which already internally process navigational metadata these could be extracted or attached to pages automatically. Alternatively, screen scrapers or text mining approaches for automated faceted hierarchy construction can be employed [3].

B. Multi-Paradigm Exploration

While we primarily employ faceted navigation, Alice also used several other approaches in our scenario. Our browser

supports users during query construction and refinement until a suitable result set is found. When Alice was satisfied with the results she switched from the search result overview towards a detailed view of single search results, at which point the original facets served as means for horizontal browsing (i.e., show more of this kind of content) or as containers for acquired navigational metadata.

To further improve query construction capabilities, we employ a multi-paradigm search that integrates basic faceted search with keyword-based search in facets, restrictions or general information artifacts along with content-based search using positive and negative examples [10]. This allowed Alice to explore the information space using alternative means, when her current options failed or were exhausted (e.g., when she could not guess any more keywords).

In this respect, our browser functions as a powerful query construction front-end for other search engines similarly to current browsers, which offer a search box yet delegate the search itself to a search engine of the user's choice. The browser might also behave as a meta-search tool by aggregating or personalizing the results of several search engines (e.g., Google or Sindice) and/or content providers.

C. Adaptation and Social Aspects

We use the semantic metadata about information artifacts (or even the artifact themselves) to tailor the user experience to the specific needs of individual users. This functionality is provided by our built-in personalization and user modeling engine which continuously track user behavior and adjust the user interface to meet the estimated needs of users [11]. This is why Alice had facets related to scientific publications in her user interface at the time and it is also the reason why only attributes she prefers were shown for individual search results instead of showing everything that was available.

Our browser renders several adaptive views of the information artifacts. For example, quantitative data are visualized in an attribute table of search results, while photos would be shown in a matrix of image thumbnails. Furthermore, the current user task also effects the available interaction tools. If Alice decided to browse photos of her colleagues exploiting the social dimension (i.e., *photos* instead of papers of people who she *works with*) she would get a matrix of image thumbnails, while for her own photos the browser would also provide interactive editing tools for (metadata) content creation/annotation (see Fig. 2).

Once users explore a single search result the views switch from *query refinement and search result overviews* to *information artifact exploration views* which focus on the attributes of the result and change the facet contents into horizontal navigation menus. As before different views are available, which include tabular views of (selected) attributes, large image (multimedia content) views or graph views for Linked data exploration [10].

III. Browser realization

We designed our browser as a client-side Silverlight application running in an existing web browser simplifying deployment (see Fig. 3). The client-side browser renders the graphical user interface and handles user interaction, which is then transformed into server-side requests as necessary. The server back-end consists of several web (WCF) services – the Factic faceted search engine, the Steltecia service for ontological repository access, and the optional SemanticLog event logging service for global statistics tracking.

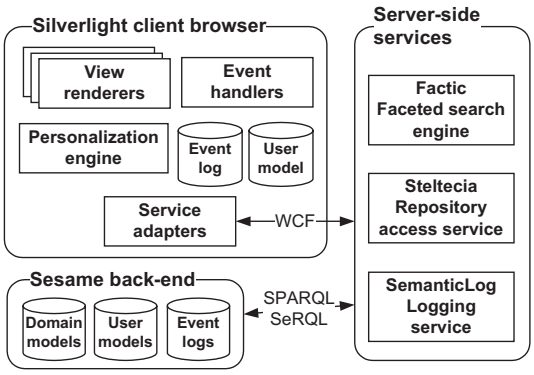

Figure 3. Architecture of our browser prototype.

Our early browser prototype works with ontological representation of information artifacts, facets, restrictions and user preferences in RDF/OWL while also consuming non-ontological data such as publicly available images. Our primary data set contains roughly 8 000 manually and semi-automatically annotated images, while our secondary data sets contain up to hundreds of thousands of publications.

We focused also on prototype functionality and performance testing, which showed improvement (response times in seconds) from our previous implementation also due to asynchronous request processing and a platform switch. Although the bottleneck still appears to be the Sesame ontological repository, we see (significant) room for improvement by caching results, reducing the number of required queries via personalization and using up-to-date hardware.

IV. Conclusions

We described the novel concept of a unified (Semantic) Web browser providing users with exploratory search and navigation capabilities in a single *end-user grade* user interface. We address these issues that decrease user experience:

- The navigational problem – we provide a single interface users can familiarize with instead of the one-size-fits-all-users GUIs which are different for each web site.
- Static/complex navigation on web sites – we provide automated personalization of the GUI corresponding to the navigational structure of web sites which is now dynamic based on usage statistics and information space evolution (not hard-coded into HTML pages).

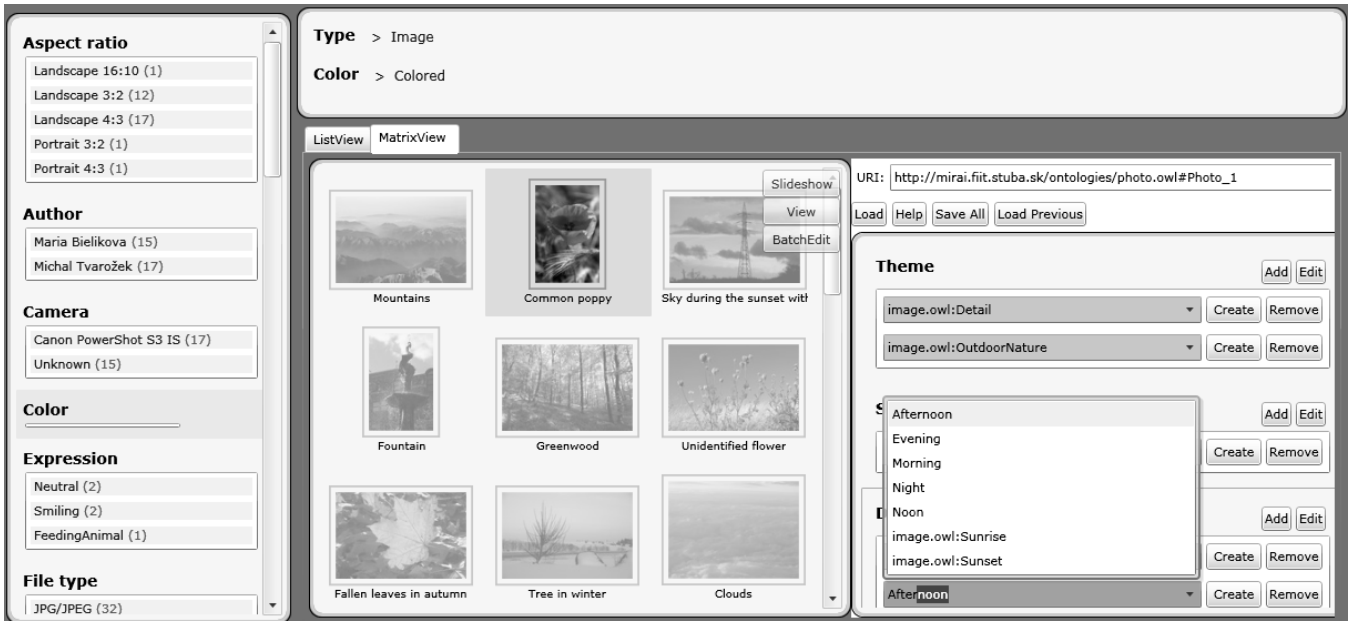

Figure 2. Our browser prototype in the image domain showing generated facets (left), a matrix of image thumbnails (center) and the metadata viewer/editor generated for the currently selected image (right).

- Personalized browsing privacy – user modeling can be performed entirely on the client side with only optional publishing of aggregate anonymous usage statistics.

Furthermore, our approach has potential to speed up Semantic Web adoption by streamlining end-user experience and legacy content annotation, thus providing viable applications that would convince content providers to supply true semantic content and cooperate via existing ontologies.

ACKNOWLEDGMENT

This work was partially supported by the Scientific Grant Agency of the Slovak Republic, grant No. VG1/0508/09 and by the Cultural and Educational Grant Agency of the Slovak Republic, grant No. KEGA 3/5187/07.

REFERENCES

[1] D. Huynh, S. Mazzocchi, and D. Karger, "Piggy bank: Experience the semantic web inside your web browser," *Web Semant.*, vol. 5, no. 1, pp. 16–27, 2007.

[2] Ben-Yitzhak, et al., "Beyond basic faceted search," in *WSDM '08: Proc. of the Int. Conf. on Web search and Web Data Mining*. New York, NY, USA: ACM, 2008, pp. 33–44.

[3] W. Dakka and P. G. Ipeirotis, "Automatic extraction of useful facet hierarchies from text databases," in *ICDE*. IEEE, 2008, pp. 466–475.

[4] K.-P. Yee, K. Swearingen, K. Li, and M. Hearst, "Faceted metadata for image search and browsing," in *Proc. of the SIGCHI Conf. on Human factors in Computing Systems*. ACM Press, 2003, pp. 401–408.

[5] M. A. Hearst, "Clustering versus faceted categories for information exploration," *Commun. ACM*, vol. 49, no. 4, pp. 59–61, 2006.

[6] E. Oren, R. Delbru, M. Catasta, R. Cyganiak, H. Stenzhorn, and G. Tummarello, "Sindice.com: a document-oriented lookup index for open linked data," *Int. J. Metadata Semant. Ontologies*, vol. 3, no. 1, pp. 37–52, 2008.

[7] schraefel,, m. c. et al., "The evolving mspace platform: leveraging the semantic web on the trail of the memex," in *Proc. of Conf. on Hypertext and Hypermedia*. New York, NY, USA: ACM, 2005, pp. 174–183.

[8] M. Hildebrand, J. van Ossenbruggen, and L. Hardman, "/facet: A browser for heterogeneous semantic web repositories," in *ISWC 2006*, ser. LNCS, I. Cruz et al., Ed., vol. 4273, 2006, pp. 272–285.

[9] E. Oren, R. Delbru, and S. Decker, "Extending faceted navigation for rdf data," in *ISWC 2006*, ser. LNCS, I. Cruz et al., Ed., vol. 4273, 2006, pp. 559–572.

[10] M. Tvarožek and M. Bieliková, "Collaborative multi-paradigm exploratory search," in *WebScience '08: Proc. of the Hypertext 2008 Workshop on Collaboration and collective intelligence*. New York, NY, USA: ACM, 2008, pp. 29–33.

[11] ——, "Personalized faceted navigation in semantically enriched information spaces," in *Advances in Semantic Media Adaption and Personalization*, Marios C. Angelides et al., Ed., vol. 2. CRC Press, 2009, pp. 181–201.