User-Centric Mapping for RDFa Web Mining

Tewson Seeoun
seeount@scss.tcd.ie

Rob Brennan
rob.brennan@scss.tcd.ie

Declan O'Sullivan
declan.osullivan@scss.tcd.ie

Knowledge and data Engineering Group
School of Computer Science and Statistics, Trinity College Dublin, Ireland

ABSTRACT

There has been a substantial rise in semantic data publishing in the form of RDFa. This raises new opportunities for mining RDFa-enabled web pages and merging the data with local knowledge repositories. An important issue in merging or mapping differently-structured knowledge schemas is the limits of automatic mapping which leads to a need for human intervention. In this paper we propose requirements for and the design of a tool that supports end users in RDFa mining and mapping tasks. Mining applications have the additional constraint that they must often support enterprise users who, unlike traditional mapping users, are not knowledge engineers or programmers.

Categories and Subject Descriptors

H.3.4 [**System and Software**]: World Wide Web (WWW)

General Terms

Design, Human Factors

Keywords

Ontology Mapping, RDFa

1. INTRODUCTION

Semantic Web technologies have enabled self-describing data to be published on the web in a standardized way. RDFa is rapidly "triplifying" the web [6]. It allows web authors to publish and link their data by embedding RDF triples alongside visual contents in the page in an accessible way.

So a large volume of RDFa data is being published, but who is consuming it and for what use cases or patterns? Many applications have been developed following W3C's RDFa specification [1]. Search engines can interpret and display RDFa and other semantic markups. A web author may write a script to fetch data related to their own content in a mashup. However the use case we focus on here is web mining and knowledge merging of RDFa data.

Stumme et al. have discussed how semantically-structured data can help improve results of traditional web mining [9]. As the web becomes more triplified, mining semantic data like RDFa will become an important addition to mining techniques tackling unstructured data on the web. Here we focus on mining of structured data from the open web and subsequent merging into local knowledge repositories based on their own independent schemas for applications like business intelligence. The merging task, often called ontology alignment or semantic mapping, has been seen to be exclusively for knowledge engineers. Hence a research challenge is to provide support for enterprise users. In this scenario, the existing mining infrastructure is complimented with efficient end user-centric merging and mining tools. Our approach is likely to be less brittle and more efficient than deploying both traditional web mining tools and semantic mapping tools in a loosely connected tool chain, especially since most current semantic mapping tools are unsuitable for enterprise users and hence remain a roadblock for widespread adoption.

Mapping between two or more differently-structured models has always been a challenge. However, when it comes to data published on the web using RDFa, syntax-based transformations will likely fail due to RDFs serialization flexibility and human-crafted transformations are too expensive to create. Automated approaches to semantic mapping have been extensively explored. Tools like the Alignment API [4] can generate sets of candidate mappings. These candidates, however, still need validation by a human, especially for domain-specific models or the loose formality of linked data. The use of Semantic Web technologies themselves can also be a challenge for those not familiar with formal knowledge representation. Optimization of the user-interaction required to validate a given candidate correspondence set, in terms of navigation support, appropriate candidate selection, inferring the consequences of implicit and explicit user feedback to date or background knowledge all have a role to play in minimizing the cost of producing mappings and increasing usability.

This paper explores the requirements for, and proposes an initial design for a tool that helps minimize human effort in mining and merging structured data published on the web into local knowledge repositories. Section 2 describes the initial use cases that such a tool should support. In section 3 we look at the currently available tools. In section 4 we present the design of a proof-of-concept tool as an extension for a web browser. In the last section we have some concluding remarks and discuss future work.

2. TECHNICAL APPROACH

To the best of our knowledge, there have not been any application that features both RDFa mining and ontology matching with an end user-centric approach. For the initial prototype, our tool does not address retrieval but is aimed at supporting efficient user-based feedback on RDFa-based pages for merging with a local knowledge-base. Hence, an agent must browse to or locate a desired RDFa-enabled page. The tool then extracts RDF triples from that page (describing individuals), locates any referenced vocabularies or ontologies on the web of data, builds an internal model of the individual and schema information described by the page, generates initial automated candidate correspondences between the target local schema and the constructed schema,

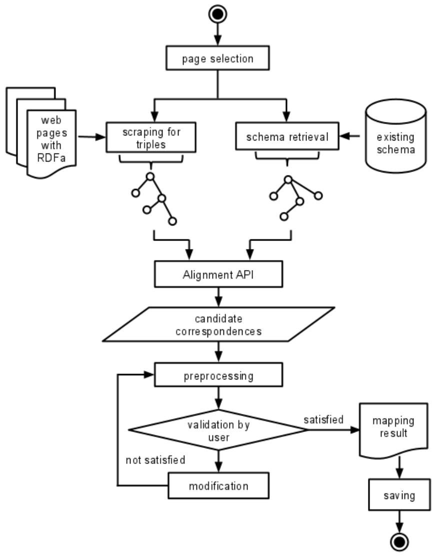

Figure 1. RDFa Mapping/Mining Task Process

preprocesses the candidates to minimize the human intervention required and to provide appropriate suggestions for users of potential mappings derived from the current page, allows the users to accept, modify or abandon the automated results, and save the resultant mapping set for future use., for example in a mining session. This process is described in figure 1 below.

3. RELATED WORK

Fuzz [3] and RDFa Developer [8] are probably the most full-featured and up-to-date web browser extensions for extracting and reusing triples from RDFa-enabled web pages.

There has also been some significant work on assisting users in ontology mapping. CogZ [5] and Prompt [7], extensions of the Protégé Ontology Editor, provide a graphical user interface for mapping. They present and two schemas to users side-by-side and allow users to filter and "link" terms.

Towards a user-centric approach, we have in the past proposed a process of ontology mapping that relies on feedback from the user [2]. The process includes a stage of setting up mapping presentation before displaying to the user.

4. DESIGN & ARCHITECTURE

As RDFa is embedded in web pages, it may be useful for the tool to deliver a user experience similar to web browsing. It also may be easier to create an initial prototype tool as a web browser extension. We describe in Figure 2 the architecture of our initial prototype tool that reflects the technical approach described above. Triples are extracted from an RDFa-enabled web page using a parser and stored locally as RDF/XML. The stored data and the target schema are sent to Alignment API. Candidate match results from the Alignment API are then obtained and stored for display and modification.

Results display is optimized as described in the use cases. This helps the user to prioritize items, maximize context and work on important correspondences first. Ease of use for enterprise users is also an essential goal. We initially propose three options for users to validate mapping results: (1) equivalent, (2) not equivalent, and (3) not relevant. "Not equivalent" terms will be manually mapped by the user, while "not relevant" terms are ignored. The mapping set so generated can then be fed into a mining/extraction process.

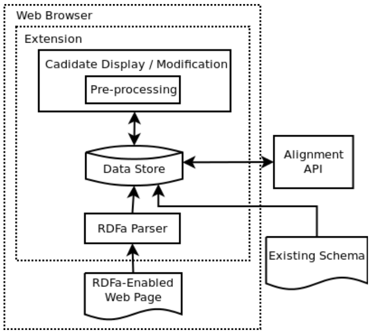

Figure 2. User Centric Mapping/Mining Tool Architecture

5. CONCLUSIONS & FUTURE WORK

In this paper, we have explored consuming RDFa with the goal of merging it with local knowledge repositories in a user-centric way. Prior work on semantic mapping shows that some human effort is useful in the mapping process. A tool can help to reduce workload by suggesting correspondences that are likely to be valid. It can further ease the task by applying intelligence to the process of selecting and presenting correspondences for validation. At this stage we have performed requirements analysis and proposed a prototype tool design as a web browser extension.

Implementation and evaluation of the concept are yet to be done. A user trial is planned. This will focus on users that are non-technical domain experts and compare our results to commercial web scraping tools. We can determine the usability of the tool through qualitative surveys and quantitative measurements such as time spent and number of mouse clicks required to complete the task. We will also compare the mapping results from that using the tool with a pre-defined gold standard.

6. REFERENCES

[1] Adida, B., Birbeck, M., McCarron, S., Pemberton, S. 2008. RDFa in XHTML: Syntax and Processing. http://www.w3.org/TR/rdfa-syntax.

[2] Conroy, C., O'Sullivan, D., and Lewis, D. 2008. Ontology Mapping Through Tagging. In *Proceedings of the 2008 International Conference on Complex, Intelligent and Software Intensive Systems (CISIS '08)*. IEEE Computer Society, Washington, DC, USA, 886-891.

[3] Digital Bazaar. 2008. Fuzz. http://rdfa.digitalbazaar.com/fuzz

[4] Euzenat, J. 2004. An API for Ontology Alignment. *The Semantic Web – ISWC 2004*, 698-712.

[5] Falconer, S.M., and Storey, M.A. 2007. A Cognitive Support Framework for Ontology Mapping. In *Proceedings of the 6th ISWC*, Busan, Korea, November 2007, Springer-Verlag, Berlin, Heidelberg, 114-127.

[6] Mika, P. 2011. Microformants and RDFa Deployment across the Web. http://tripletalk.wordpress.com/?p=59.

[7] Noy, N.F., and Musen, M.A. 2003. The PROMPT Suite: Interactive Tools for Ontology Merging and Mapping. *Int. J. Hum.-Comput. Stud.* 59, 6 (December 2003), 983-1024.

[8] Pozueco. J. 2010. RDFa Developer. http://rdfadev.sourceforge.net

[9] Stumme, G., Hotho, A., and Berendt, B. 2006. Semantic Web Mining: State of the Art and Future Direction. *J. Web Semant.: Sci. Serv. Agents World Wide Web 4*, 2 (June 2006), 124-143.