Leveraging Emergent Ontologies in the Intelligence Community

Jim Starz, Jason Losco, Brian Kettler, Rachel Hingst, and Chris Rouff

Lockheed Martin Advanced Technology Laboratories

jstarz@atl.lmco.com

Abstract - The vision of a Semantic Web of intelligence knowledge has yet to be fully realized, in part because of the tough challenges of ontology engineering and maintenance. Recent developments on the World Wide Web and IC intranets demonstrate that individual users are willing to supply structured information conforming to de facto standards. This can be most prominently seen in "peer produced" folksonomies and knowledge bases such as Wikipedia and Intellipedia, its cousin. Though these structures lack the machine reasoning potential of highly engineered ontologies, for many purposes they are "good enough". This paper describes Contrail, a prototype information management application, that leverages an "emergent" ontology from Wikipedia to model a intelligence analyst's context and exploit that model to aid information retrieval, refinding, and sharing

I. INTRODUCTION

The widespread adoption of Semantic Web and other ontology-based applications in the intelligence community (and indeed the wider web) is that quality ontologies are difficult to build, maintain, and exploit. Ontology engineering requires significant subject domain expertise and knowledge engineering skills. For all-source and other kinds of analysts, such ontologies span a broad range of subject domains, which are constantly evolving.

Wikipedia and Intellipedia are approaches to capturing this broad range of knowledge from the community without requiring pre-built ontologies. These knowledge bases are not without structure. A prominent example is the World Wide Web's Wikipedia, which contains over fifteen million pages. The structure for pages of the same type are very similar, illustrating that people are willing to provide structure in the form of lightweight ontology-like information. This similarity is discussed in the work on Wikitology [4] and dbpedia [1].

While such "ontologies" might not support formal automated reasoning system well, they can support other useful applications. Our research investigated leveraging emergent ontologies for the purposes of representing user models of analysts. The work used an ontology derived from Wikipedia. This paper describes our prototype application, its use of Wikipedia, and some preliminary results.

II. THE CONTRAIL TOOLS

The Contrail tools help analysts find, organize, re-find, and share unstructured and semi-structured information obtained from the Web (or Intelink), email, documents, and other sources [2]. While our focus is on intelligence analysts, these tasks are those of many knowledge workers. Contrail has been evaluated in several experiments with real intel analysts on open source intelligence tasks.

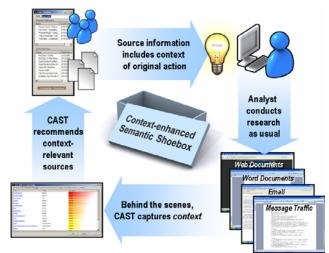


Fig. 1. High-Level Concept of Operations for Contrail Tools

Fig. 1 shows the high-level concept of operations for the Contrail tools as an analyst does her research online, she finds relevant items through web browsing, web searches, reading email, etc. Through instrumentation and logging services, Contrail is notified of these "information keeping actions", such as the bookmarking of a web page. Contrail then performs a semantic analysis of each kept information item's content using text analytics and other methods. Using the results of this analysis, Contrail updates its model of the analyst's context and stores a copy of the kept item in her Semantic Shoebox. A user's Semantic Shoebox can be thought of as a semantically grounded container for

accumulated pieces of information. Contrail supports the sharing and retrieval of kept items from other analyst's shoeboxes. The contextual knowledge appended to these items by Contrail helps one analyst quickly understand the potential relevance and pedigree of an item retrieved from another analyst's shoebox.

The Contrail Refinder tool, shown in Fig. 2, presents a more comprehensive view of a Semantic Shoebox and displays a variety of information (textually and graphically) associated with a kept item including its metadata, content, and context tags. A user may do a one button search to display those items most relevant to his current context. Contrail also presents context-relevant recommendations for stored items and potential collaborators in a desktop sidebar.

At the core of Contrail is its Context Aggregator which maintains and updates the user's context at each keeping action. Concepts and their instances (specific people, organizations, locations, etc.) are extracted from the text of the kept item using a commercial entity extractor. A spreading activation algorithm is used to find related concepts in a knowledge base (KB). These related concepts might not be explicitly mentioned in the text itself. Extracted and related concepts are thus associated with an activation level and the most active concepts represent the user's current context. Contrail's KB, grounded in handbuilt OWL ontologies extending the SUMO [3].

This approach worked well, as judged in experiments with analysts who periodically reviewed Contrail's model of their contexts. Contrail's use of an ontologically-grounded

knowledge base of concepts, however, presented significant ontology engineering and maintenance challenges, as well being limited by the underlying entity extractor used. These challenges – all potential barriers to Contrail's deployment – included the potential breadth required for ontologies and the handling of new concepts and entities in these dynamic domains.

III. USING WIKIPEDIA

To alleviate these issues, we have replaced the static ontology based context representation with one based on Wikipedia. We used IR based techniques to relate documents with pages in Wikipedia and associated a score with each relationship. One significant benefit of this approach is the elimination of the need for knowledge engineering to update the "ontology." Wikipedia serves as a publicly maintained emergent ontology, allowing for user context to shift as the world changes.

Specifically, keeping actions performed by the users associate their interests in particular documents or snippets of text. Based on this text, we query a Lucene index of Wikipedia to obtain pages that may be of interest to the user. A weighted merge of the query results is performed with their existing contextual information to form their updated user model.

It should be noted that given the scale of Wikipedia, such queries are very resource intensive. Despite this challenge, the results from leveraging the emergent ontology from Wikipedia appear promising.

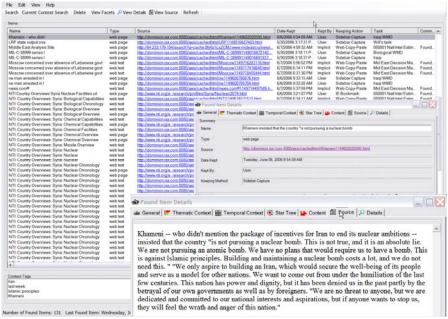


Fig. 2. Contrail Refinder (Item Browser, Item General Details, and Item Source Details screens)

IV. EVALUATION

Initial informal experimentation using this new approach for user modeling has shown significant improvements over using a traditional static ontology in representing user context. The new approach improves finding documents and collaborators. There was also anecdotal evidence that the biggest advantage occurred when new concepts and instances were present in the emergent ontology that could be immediately leveraged. An example of the differences is shown below.

TABLE 1
Example of context terms from static ontology and Wikitology derived terms

Static Ontology	Wikitology
Indonesia	United Malyas Nat. Org.
Malaysia	Ketuanan Melayu
Singapore	Mahatir bin Muhamed
June	Islam in Malaysia
2002	Anwar Ibraham

The Wikitology approach consistently provided more specific terms that may not easily be found in an ontology or by text analytics packages. Using the old approach, we found general terms would dominate the user context. The breadth of Wikipedia does add the potential for significant noise, such as pages about specific dates. Though Wikipedia is relatively comprehensive, for specific domains pages may not exist. For emerging concepts, it is critical to mirror Wikipedia and update the index regularly. The results of this evaluation will be documented in a future research paper.

V. FUTURE WORK

Our research agenda includes further investigations to determine new applications where emergent ontologies can be applied. This investigation will include tools leveraging these ontologies for enhanced semantic authoring. We also plan to investigate the extraction of rules from patterns in emergent ontologies. A major focus area will be handling the significant scale and rapid updates of Wikipedia. Both of the aspects provide significant challenges and opportunities. Finally, we plan to make additional extensions to the Contrail suite of tools to extend the representation of user models.

VI. CONCLUSION

In the large distributed nature of the World Wide Web, leveraging massive convergence in terminology and structure can be highly useful. While these structures may not replace formal ontologies, they can be appropriate for certain applications and they can help bridge a gap to more formal structures. We have demonstrated that the use of the

ontological structure of Wikipedia for representing context has advantages over human-engineered ontologies for at least one application and likely many others.

ACKNOWLEDGEMENTS

Many of the concepts applied in this paper were motivated by conversations with Tim Finin of the University of Maryland at Baltimore County.

REFERENCES

- [1] S. Auer, C. Bizer, J. Lehmann, G. Kobilarov, R. Cyganiak, Z. Ives: DBpedia: A Nucleus for a Web of Open Data. In Aberer et al. (Eds.): The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, November 11–15, 2007. Lecture Notes in Computer Science 4825 Springer 2007, ISBN 978–3-540–76297–3.
- [2] B., Kettler (2008). Putting Knowledge in Context to Facilitate Collaboration. In Proceedings of the 2008 International Symposium on Collaborative Technologies and Systems (May 19-23, 2008 in Irvine, CA). IEEE, 313-320.
- [3] I. Niles, and A. Pease. 2001. Towards a standard upper ontology. In Proceedings of the international Conference on Formal ontology in information Systems - Volume 2001 (Ogunquit, Maine, USA, October 17 - 19, 2001). FOIS '01. ACM, New York, NY, 2-9.
- [4] Z, Syed et al., "Wikipedia as an Ontology for Describing Documents", In Proceedings, Proceedings of the Second International Conference on Weblogs and Social Media, March 2008.
- [5] M. Williams and J. Hollan. (1981). The Process of Retrieval from Very Long-Term Memory. Cognitive Science 5: 87-119.