Ontology Extraction from Social Semantic Tags

Csaba Veres
InfoMedia, Univ. of Bergen
POB 7802, 5020 Bergen
+47 55588154
csaba.veres@uib.no

Weiqin Chen
InfoMedia, Univ. of Bergen
POB 7802, 5020 Bergen
+47 55584143
weiqin.chen@uib.no

Andreas Opdahl
InfoMedia, Univ. of Bergen
POB 7802, 5020 Bergen
+47 55584140
andreas.opdahl@uib.no

Kristian Johansen
InfoMedia, Univ. of Bergen
POB 7802, 5020 Bergen
k.johansen@student.uib.no

ABSTRACT
We are developing an approach to organizing bookmarks and other information resources by annotating them with *metadata* in the form of *synsets* taken from WordNet. (A synset is a unique sense of a word). Collections of annotated bookmarks can be semantically enriched by adding *hypernyms* and *hyponyms* from WordNet, where the additional synsets provide a maximally informative summary of the user defined annotations. The choice of maximally informative nodes is currently obtained through an interactive visualization. The ultimate aim is to automatically produce a lightweight ontology of any annotated data source.

Categories and Subject Descriptors
H.3.4 [**Information storage and retrieval**]: Systems and Software – *Semantic Web, Social Networking, Web 2.0.*

General Terms
Management, Standardization.

Keywords
Ontology, folksonomy, WordNet.

1. INTRODUCTION
Ontologies present a useful technology for describing the knowledge in a domain of interest. They facilitate the discovery of existing information in that domain, and for exposing hitherto hidden knowledge, using inference over the known facts [1]. But developing an ontology for a domain of interest is a difficult and potentially costly task, which has hindered their deployment as a ubiquitous technology for information management. There are attempts to (semi) automatically generate ontologies from the contents of the information spaces, but the utility of these techniques is questionable [4].

These difficulties have allowed a competing technology to flourish in recent years. The exploration and management of information spaces has to some extent been usurped by the popular technique of social tagging [3]. The tag spaces or folksonomies that emerge from collaborative tagging facilitate refindability and discovery. However, folksonomies have subsequently presented with problems of their own. Issues such as ambiguity and indeterminacy of tags have made them less useful for large scale information management. In addition, they do not have the richness of ontologies and cannot support inference. Our work seeks a marriage of these technologies. First we use semantic tags for more precise user annotations, and then we use the semantic tags to generate enriched ontologies of the information space.

2. DESCRIPTION OF WORK
We are developing a refinement of content management with *user tagging,* that uses semantically unambiguous annotations instead of naive user keywords. We achieve this by linking our annotations to existing semantic metadata standards through URIs. In this paper, we focus on the electronic lexical database WordNet [2] as an example of a metadata standard. Our basic approach is to annotate information resources with WordNet synsets, which are disambiguated senses of words.

In our current experiments we are using an existing collection of 73 http bookmarks that had already been semantically annotated with 111 WordNet *synsets*. Semantic annotations in this case refine naive user keywords by disambiguating each keyword with a reference to a valid synset in WordNet. For example, a bookmark to http://www.bats.org.uk/ can be be annotated with the keyword "bat" and the URI http://www.w3.org/2006/03/ wn/wn20/instances/bat-noun-1, whereas http://baseball.com/ can be annotated with "bat" along with http://www.w3. org/2006/03/wn/wn20/instances/bat-noun-2. The naive (non-semantic) alternative is to annotate both URIs only with the keyword "bat", so that searches for "bat" return bookmarks both about mammal bats and sports bats, along with all the other meanings of "bat". In addition to distinguishing between homonyms, semantic annotations can even take care of synonyms, so that two keywords, such as "bat (in the mammal sense)" and "chiropteran", are both linked to the same URI (in this case http://www.w3.org/2006/03/wn/ wn20/instances/bat-noun-1).

The use of semantically precise tags allows us to organise the metadata annotations in ways that are not possible with naive keywords. For example with the WordNet *synsets*, we can easily compute the transitive closure of the *hypernyms* of each tag. Unfortunately if we do this for all 111 tags, we end up with a very rich and confusing taxonomy filled with concepts of various degrees of usefulness. (WordNet is a very comprehensive lexical database, which includes terms that may not enjoy common use. For example, *hypernyms* of "car" include "self-propelled vehicle", and "instrumentality"). In this research we are developing algorithms to trim the tree to leave only the most useful nodes. Here we describe a very simple algorithm to do this. After we create chains of WordNet *hypernyms* for each of tag, we mark as irrelevant all the *synsets* that do not have two (this is a parameter) or more different Bookmarks in their hyponym chains. We also mark as irrelevant the *hypernyms* located at height six (also a parameter) or more in some chain, that is, we remove the top level

nodes. What we are left with is a collection of *hypernyms* of the original *synsets* that satisfy two criteria. Firstly, each remaining *hypernym* is a decision point that returns a smaller subset of Bookmarks than its *hypernym* (if any); otherwise, they would not add semantic structure to the existing annotations. Secondly, none of the remaining *hypernyms* are too abstract; otherwise, they would not be useful when browsing and visualizing the bookmark collection. Figure 1 helps to illustrate the results of our algorithm.

The green nodes represent the *synsets* that were used to annotate the resources. Every other node is a WordNet *hypernym*. Grey nodes are trimmed because each one subsumes just one other node. Blue nodes are not trimmed because they each subsume two or more nodes. Black nodes are trimmed because they are less than six levels removed from the top level. Each of these numbers is a variable that can be adjusted trough the dialog box. It is also possible, for example to keep only nodes that subsume three or more other nodes, and are more than four levels from the top.

This simple procedure has already given some interesting results. For example the tags "game", "teaching" and "shopping" resulted in the inferred *informative hypernym* "event", and the tags "browser", "spreadsheet" in "program". The current algorithm is a very simple one, and one area for future research is to investigate some information theoretic approaches to finding maximally informative nodes.

The addition of these informative *supernodes* is useful in organizing the content of the information store. In a sense, well chosen *supernodes* act as emergent categories that structure the content of the store. These categories are not fixed, but emerge to reflect the contents of individual stores. However, the categories are not wholly unconstrained, because they emerge from the rich vocabulary of the mental lexicon. As such, they reflect an understanding which is shared by the entire language speaking community. The category structure can be used to find individual resources within a store, but it can also be used to compare different stores by comparing the category structures.

The emergent category structures can be greatly enhanced by enriching the inferred nodes with relation data available in WordNet. This includes part-of/has-part (car - accelerator, airbag, …), domain terms (car - tunnel, passenger, prang, …), co-ordinate terms (car - go kart, motorcycle, cart, …), and so on. There are several ways to implement this enriching process. One is to augment the base tags before the trimming algorithm, and the other to augment the remaining nodes after the trimming is complete. It is not clear if these will give the same result, but both will result in a lightweight ontology with classes and relations.

3. CONCLUSION

In conclusion, we aim at developing a simple but powerful procedure for developing a domain ontology automatically, based on the considered annotation of domain resources by every day users. The ontology can be used to retrieve resources, and to characterize the knowledge stored in a particular information source. The approach is completely generic. We used bookmarking as an illustrative example, but the technique will work for any information store annotated with user generated semantic tags.

4. REFERENCES

[1] Gruber, T. 1995. Toward Principles for the Design of Ontologies Used for Knowledge Sharing. *Int. J. Human-Computer Studies* 43, 5-6, 907-928.

[2] Miller, G.A. 1995. WordNet: A Lexical Database for English. *Communications of the ACM* 38, 11, 39-41.

[3] Shirky, C. 2005. *Ontology is Overrated -- Categories, Links, and Tags*. In Clay Shirky's Internet Writings.

[4] Zheng, H-T., Borchert, C., Kim, H-G. 2008. A Concept-Driven Automatic Ontology Generation Approach for Conceptualization of Document Corpora. In *Proceedings of Web Intelligence and Intelligent Agent Technology*. WI-IAT '08. IEEE/WIC/ACM. vol.1, 352-358.

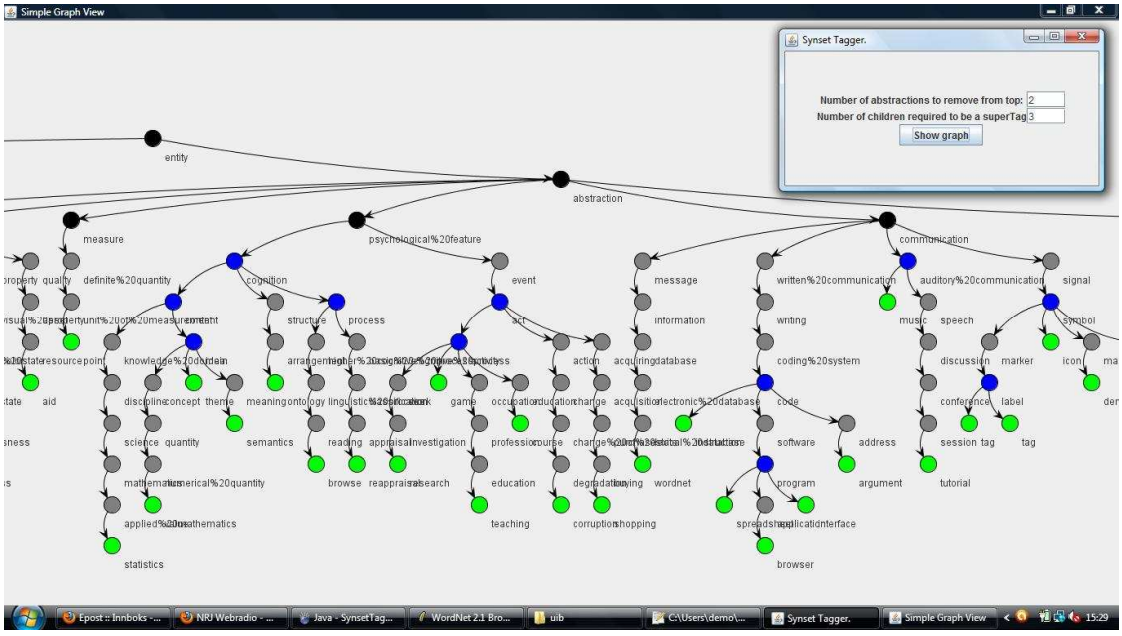

Figure 1. Inferred category structure