Tagging Semantics: Investigations with WordNet

Michael J. Cole  
Rutgers University  
4 Huntington Street  
New Brunswick, NJ 08901  
+1 917 774-2653  
mcole@scils.rutgers.edu

Jacek Gwizdka  
Rutgers University  
4 Huntington Street  
New Brunswick, NJ 08901  
+1 732-932-7500 ext. 8236  
jcdl2008(at)gwizdka(dot)com

ABSTRACT
The content of a tag sequence references both a user's concepts and the user's conceptualization of an information object. The tagging history of 823 users of the Delicious social tagging service is analyzed using WordNet. Three semantic measures of the tagging content are developed: the level of category references, the changes in category level for each noun as the tagging sequence unfolds, and the scope of concept coverage as the compactness of the WordNet subgraph for the noun senses. Observed patterns of concept reference as a function of sequence position hint at dynamic properties of the tag production process by marking a trace of cognitive activity. If tagging is object categorization, these measures provide a view of the personal categorization behavior of non-professionals and illuminate biases in the production of 'folksonomies' due to tag production processes.

Categories and Subject Descriptors
H.3.1 [Content Analysis and Indexing]: Indexing methods

General Terms
Human Factors, Theory.

Keywords
Tagging, folksonomy, categorization, semantics.

1. INTRODUCTION
Users producing tags appear to carry out a variety of intentions, but a main activity of social tagging is taken to be an indexing or categorization activity allowing a user to navigate a personal ontology for the purposes of refinding or future use of the information object. This direct knowledge of the user's categorization makes it interesting to ask about the level of object abstraction for this indexing activity. The content of a tagging instance can be considered as a region of the WordNet space. The subgraphs in the region cover each interpretation of the tag sequence, enabling an exploration of tagging semantics by selecting interpretations based on hypotheses about the user's intention that caused the selection of those tags.

We consider user tagging instances using two perspectives to calculate properties of the dynamics of user tagging instances, for example, whether the sequence of noun references become more or less specific as the user produces the descriptive tags. If tag sequences are traces of cognitive activity in user thoughts about the information object, observed patterns may provide hints about user intention and user access to mental representations of the information object.

2. Methodology
Words in each tag sequence are associated with particular synsets in the WordNet semantic space using two assumptions. In one, the word meaning is fixed with the naive assumption that each tag word is the most common sense of the word. In the other, every combination of WordNet senses for the words in the tag sequence are considered and the combination yielding the most compact graph in the semantic space is selected. We then use the calculated tag meanings from these two content representations to examine the sequence of tags in each tagging instance. In particular, we count the relative hierarchical position of the tags in the WordNet synset hierarchy, which corresponds roughly to the specificity of the word meaning. These two perspectives may be thought of as looking at tag production as tags selected to be most widely understood vs. producing tags from the perspective of a single concept.

3. Results and Conclusion
Users are found to be strongly biased towards specific references in tag meaning. While there are no strong patterns of increasing or decreasing levels of conceptual reference in the tag sequences when all users are considered, some users display consistent biases to start at a conceptual level of reference and then elaborate with tags that are of greater generality or specificity compared to the initial tag.

This methodology to explore tag semantics is promising and can be extended to compare user tagging intentions and explore user recovery or understanding of (re)encountered tags associated with information objects, for example in library catalogs or on web pages.

4. ACKNOWLEDGMENTS
This work has been partially supported by IMLS grant LG-06-07-0105-07.