Deriving Ontology from Folksonomy and Controlled Vocabulary

Introduction

Popular online tagging websites, such as Flickr, Technorati, and Del.icio.us, allow users to tag objects freely without constraints of any controlled vocabulary. The word “folksonomy” has been used to describe this type of grass-root taxonomies, which provides a rich source for building ontologies.

Research has experimented with building ontology purely from user-generated tags, and the approaches have been focused primarily on statistical methods. Schmitz (2006) conducted a study of inducing an ontology from Flickr tags, which used statistical methods to detect subsumption pairs based on co-occurrences of tags. Heymann Carcia-Molina (2006) established a hierarchical taxonomy based on tags from Delicious and CiteULike respectively by using cosine similarity of tag vectors.

Although a good source for building ontology, folksonomy has its disadvantage in representing object content. For example, the tags are subjective and their hyponyms are usually not indicated (Mathes, 2004). Relations between tags are unknown and useful information is missed in tags, therefore the unorganized status of tags affects retrieval of objects negatively. Folksonomies alone are not sufficient for building a comprehensive and high-quality ontology. Researchers have suggested using other sources such as WordNet and Wikipedia to assist ontology building from folksonomies (Damme et al., 2007).

Contrary to folksonomies, controlled vocabulary is characterized by rigid structures and slow responsiveness to new terminology. But its systematic organization and careful formulation of terms and relationships would be complementary to the disadvantages of folksonomies. Online lexical resources and gazetteers as instances of controlled vocabulary appear to be promising approaches in using folksonomies for generating ontologies (Schmitz, 2006; Damme et al., 2007). While researchers are speculating this method, little research has been done to actually implement it to build ontology. In addition, building ontologies from a combination of both folksonomies and controlled vocabulary is rarely mentioned in previous works. This study is intended to fill the gaps between methodologies in using folksonomies to produce ontologies.

Assumption

Tags are keywords used to index objects and one object (such as a photo or a webpage) may be tagged by one or more tags. Each tag may be associated with other tags through co-occurrences. We assume that frequent co-occurrences do not happen by accident in statistical terms. If one tag frequently co-occurs with another tag, then there should be some relationship between them. The co-occurring tags of one tag are called “related tags”. We emphasize related tags in this study because related tags can offer useful semantic information about tag relationships.

Methodology

This study uses an approach of combining user-generated tags and controlled vocabulary to develop an ontology on landscape. The controlled vocabulary is used as the backbone (Mani et al., 2004) for the ontology. We then expand the knowledge structure by adding more entries from user-generated tags. The existing terms in the controlled vocabulary serve as classes in the ontology, with clear hierarchy of classes and subclasses (Qin & Paling, 2001). Based on the tags, we build a micro-hierarchical system for each class. The micro-hierarchies may be deployed in two ways: one is used as updates for the controlled vocabulary and the other is used as assemblages to form a portable ontology. This is the overview of the method, and the method of building hierarchy for subclasses is presented below.

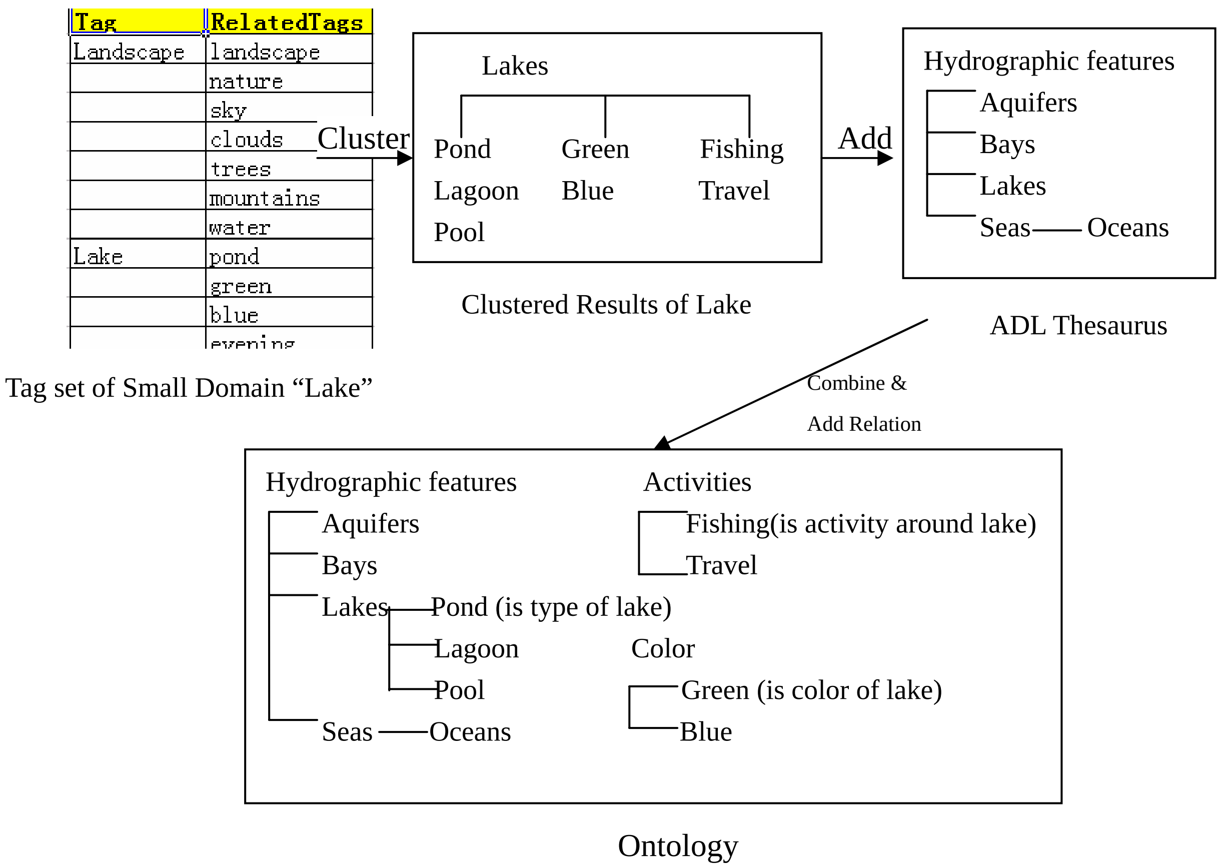
Tags can be represented by their related tags. For a tag T, its related tags are rt_1, rt_2, …, and rt_n, which forms a vector for the tag T(rt_1, rt_2, …, rt_n). All the tags in the domain (the subclass domain) can be converted to the vector format. Then we further build a hierarchy for all the tags based on the vector similarity scores. Clustering techniques is applied in extracting hierarchy from the tags. We compute the vector similarity between tags and cluster the tags to form a hierarchy. Relations will be added to describe the relations between classes of the hierarchy manually.

Data and Implementation

For the pilot study, we chose the field of landscape images as the ontology domain, and used Flickr tags and Alexander Digital Library (ADL) geographic feature type thesaurus as controlled vocabulary. Flickr users tag photos with keywords and the system has cumulated more than a million photos tagged by “landscape.” The ADL thesaurus was extracted into a database and the level of each term in the hierarchy is indicated. For example, “gulfs” is labeled level 2, with its upper category “hydrographic features” as level 1.

The tag data were collected from the related tags from the Flickr API flickr.tags.getRelated. For example, “sky” is one of the related tags of “landscape,” and “landscape” can be represented by vector (nature, sky, clouds, trees, …, evening). Figure 1 demonstrates how the “lakes” subclass of ADL thesaurus is expanded into a hierarchy.

The data set of tags is built upon tag “lakes” by collecting the related tags, related tags of the related tags, and so on, until the number of tags reached the threshold 100. The data set can be viewed as tags in the small domain of lake. With the related tags provided by Flickr API, we can obtain vector representation of each tag in our database. Clustering is applied on the vectors to explore hierarchy among the tags. With the updated hierarchy, relations are added between classes. The workflow and mock-up result are shown in Figure 1.



Tag	RelatedTags
Landscape	landscape
	nature
	sky
	clouds
	trees
	mountains
	water
Lake	pond
	green
	blue
	evening

Figure 1. A mock-up result of ontology generation from tags

Future Plan

Knowledge capture is the bottleneck problem in intelligent information systems. Developing ontologies from user-generated tags in combination of controlled vocabulary is our contribution to solving this bottleneck problem.

In this poster proposal, we proposed a method of building ontology from folksonomies and controlled vocabulary. A pilot study was conducted in the field of landscape with partial completion. The next phase of research is to create a training data set by building knowledge base that contains tags, controlled vocabulary, and reasoning rules for automatic ontology generation and relation assignment. While the training data set involves manual work due to its complexity and lack of prior data in this regard (Qin & Paling, 2001; Ding & Foo, 2002), we expect to automate this process in future research.

References

Alexandria Digital Library Feature Type Thesaurus. Retrieved October 29, 2007 from http://www.alexandria.ucsb.edu/gazetteer/FeatureTypes/ver070302/index.htm

Chung, C.Y., Lieu, R., Liu, J., Luk, A., Mao, J, & Raghavan, P. (2002). Thematic mapping-From unstructured documents to taxonomies. CIKM '02, Virginia, USA.

Damme, V.C., Hepp, M., & Siorpaces, K. (2007). FolksOntology: An integrated approach for turning folksonomies into ontologies. ESWC 2007 "Bridging the Gap between Semantic Web and Web 2.0" workshop.

Ding, Y., & Foo,S. (2002). Ontology research and development. Part1-A review of ontology generation. Journal of Information Science, 28(2), 123-136.

Heymann, P., & Carcia-Molina, H. (2006). Collaborative creation of communal hierarchical taxonomies in social tagging systems. Technical Report 2006-10, Department of Computer Science, Stanford University.

Mani, I., Samuel, K., Concepcion, K., & Vogel, D. (2004). Automatically inducing ontology from corpora. 3rd International Workshop on Computational Terminology, COLING'2004, Geneva.

Mathes, A. (2004). Folksonomies-Cooperative classification and communication through shared metadata. Retrieved Octerber 29, 2007, from http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html

Qin, J., & Paling, S. (2001). Converting a controlled vocabulary into an ontology: The case of GEM. Information Research, 6(2). Retrieved October 29, 2007, from http://informationr.net/ir/6-2/paper94.html

Schmitz, P. (2006). Inducing Ontology from Flickr Tags. WWW 2006, Edinburgh, UK.