CREATING STRUCTURE FROM DISORDER:
USING FOLKSONOMIES TO CREATE SEMANTIC METADATA

Hend S. Al-Khalifa, Hugh C. Davis
Learning Technology Research Group, ECS, The University of Southampton, Southampton, UK
{hsak04r/hcd}@ecs.soton.ac.uk

Lester Gilbert
Learning Technology Research Group, ECS, The University of Southampton, Southampton, UK
lg3@ecs.soton.ac.uk

Keywords: Semantic Metadata, Folksonomies, Collaborative Tagging, Social Bookmarking, Learning Resource.

Abstract: This paper reports on an on-going research project to create educational semantic metadata out of folksonomies. The paper describes a simple scenario for the usage of the generated semantic metadata in teaching, and describes the 'FolksAnnotation' tool which applies an organization scheme to tags in a specific domain of interest. The contribution of this paper is to describe an evaluation framework which will allow us to validate our claim that folksonomies are potentially a rich source of metadata.

1 INTRODUCTION

Sue is teaching a course on Cascading Style Sheets (CSS) as part of the web development course in her institute. In her daily quest for finding suitable learning resources to support her curriculum, she uses the del.icio.us bookmarking service to hunt for resources instead of spending her time Googling.

Sue believes that del.icio.us contains links to massive amounts of useful materials that can be used in an educational context, and will be of great help to her.

There is no semantic metadata in del.icio.us to describe the educational purpose of these materials, but for Sue this lack of metadata is not a major problem, because she has the appropriate tool to generate this missing information. So, she fires-up the FolksAnnotation tool, a desktop application, which works as an interface to the del.icio.us bookmarking service, to convert people's tags into more structured and meaningful metadata records. One added benefit to the generated metadata records is that they comply to a pre-defined CSS ontology.

By using this tool, Sue removes the hurdle of visiting the designated bookmarked website or even going through all the tags that people have generated to know what the site is about. Moreover, she can use the generated metadata records in her course database portal.

In another scenario, Sue uses the structured metadata created from the FolksAnnotation tool to populate her course portal database. The portal helps her students and other teachers alike, to search for CSS resources and to get more 'intelligent' results.

2 BACKGROUND

The growing popularity of folksonomies and social bookmarking services has changed how people interact with the Web. Many people have used social bookmarking services to bookmark web resources they feel most interesting to them, and folksonomies were used in these services to represent knowledge about the bookmarked resource. Next a brief overview of the two named concepts is given.

2.1 Folksonomies

The word folksonomy is a blend of the two words 'Folks' and 'Taxonomy'. It was first coined by the

information architect Thomas Vander Wal in August of 2004. Folksonomy as Thomas (Vander Wal, 2004) defines is: "… the result of personal free tagging of information and objects (anything with a URL) for one's own retrieval. The tagging is done in a social environment (shared and open to others). The act of tagging is done by the person consuming the information."

From a categorization perspective, folksonomy and taxonomy can be placed at the two opposite ends of categorization spectrum. The major difference between folksonomies and taxonomies are discussed thoroughly in (Quintarelli, 2005) and (Shirky, 2005).

Taxonomy is a top-down approach. It is a simple kind of ontology that provides hierarchical and domain specific vocabulary which describes the elements of a domain and their hierarchal relationship. Moreover, they are created by professional people, and require an authoritative source.

In the contrary, folksonomy is a bottom-up approach. It does not hold a specific vocabulary nor does it have an explicit hierarchy. It is the result of people own vocabulary, thus, it has no limit (i.e. open ended), and tags are not stable nor comprehensive. Moreover, folksonomies are generated by people who have spent their time exploring and interacting with the tagged resource (Wikipedia, 2006).

2.2 Social Bookmarking Service

Social bookmarking services are server-side web applications; where people can use these services to save their favorite links for later retrieval. Each bookmarked URL is accompanied by a line of text describing it and a set of tags (aka folksonomies) assigned by people who bookmarked the resource (as shown in Figure 1).

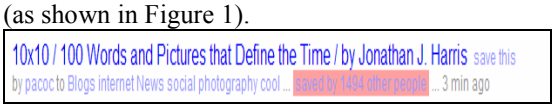

Figure 1: Excerpt from the del.icio.us service showing the tags (Blogs, internet, ... ,cool) for the URL of the article by Jonathan J. Harris, the last bookmarker (pacoc, 3mins ago) and the number of people who bookmarked this URL (1494 other people).

A plethora of bookmarking services do exists (e.g. del.icio.us, Furl, Spurl and del.irio.us); however, del.icio.us is considered one of the largest social bookmarking services on the Web. Since its introduction in December 2003, it has gained popularity over time and there have been more than 90,000 registered users using the service and over a million unique tagged bookmarks (Menchen, 2005; Sieck, 2005). Visitors and users of the del.icio.us service can browse the bookmarked URLs by user, by keywords (aka tags or folksonomies) or by a combination of both techniques. By browsing others bookmarks, people can learn how other people tag their resources; thus, increasing their awareness of the different usage of the tags. In addition, any user can create an inbox for other users' bookmarks, by subscribing to the other user's del.icio.us pages. Ditto, users can subscribe to RSS feeds for a particular tag, group of tags or other users.

3 RESEARCH MERITS

The FolksAnnotation tool applies an organization scheme to people's tags in a specific domain of interest (i.e. teaching CSS). Thus, the folksonomy tags in our system are modeled not as text keywords but as RDF resources that comply to pre-defined ontologies. This provides two benefits:

Benefit 1: While the folksonomy approach retrieves documents by using 'bag of words', property-value pairs enable more advanced search such as question answering, reasoning as well as document retrieval. So our approach will provide a property-value relationship that is semantically rich and allow for more 'intelligent' search such as: Search by Difficulty, Search by Instructional level and Search by Resource type.

Benefit 2: Typical semantic annotation tools depend on an intermediate process called Information Extraction (IE) to extract the main concepts from the annotated document before relating them to the designated ontologies. The IE process is a very complex phase in the semantic annotation lifecycle, and encompasses many advanced techniques from the natural language processing domain. Moreover, the processing time required to accomplish the IE task is significant. So, instead of using IE process as an intermediate phase for extracting knowledge from documents, why not rely on people's generated metadata? Therefore, by using folksonomies as knowledge artifacts in the process of semantic annotation, we ensure that we have used a cheap and rich source of metadata generated by people's collective intelligence.

4 IMPLEMENTATION

The implementation of the FolksAnnotation tool has been previously reported in (Al-Khalifa and Davis, 2006), however, a briefly discussion about the implemented tool and the portal that uses the generated semantic metadata needs to be highlighted to setup the stage for the evaluation framework (section 5).

4.1 The FolksAnnotation tool

Is a stand-alone application that takes as an input a del.icio.us URL of a bookmarked resource, and generates in the background the appropriate semantic metadata in an RDF format. The tool was built using Java SWT library and uses Jena API for ontology manipulation and inference.

The tool consists of two components (as shown in figure 2): the Normalization pipeline, and the Semantic Annotation pipeline. Next, a detailed description of the two processes is discussed

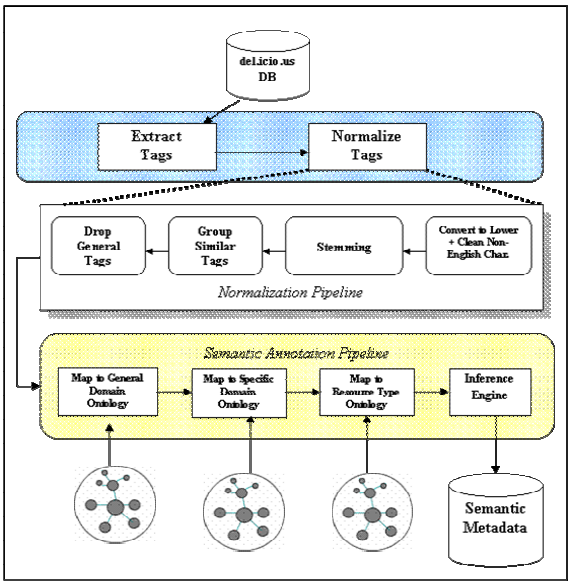

Figure 2: System Architecture of the 'FolksAnnotation' Tool.

4.1.1. The Normalization pipeline

This process starts by fetching a bookmarked resource from the del.icio.us bookmarking service, so that the tag extraction process starts extracting viable information from the web page of the bookmarked web resource. This information includes: Web Resource Title, URL, Number of people who bookmarked the resource and the list of all tags assigned to the bookmarked resource.

All tags assigned to a web resource in the del.icio.us service are extracted and then normalized using several techniques. First, tags are converted to lower case so that string manipulation (e.g. comparison) can be applied to them easily. Secondly, non-English characters are dropped; this step is to insure that only English tags are present when doing the semantic annotation process. Thirdly, tags are stemmed (e.g. converting plural to singular) using a modified version of Porter Stemmer, then similar tags are grouped (e.g. inclusion of substrings). Finally, the general concept tags (e.g. 'programming', 'web', etc) in our domain of interest are eliminated. The process of normalization is done automatically and it is potentially useful to clean up the noise in peoples' tags. Table 1 and Table 2 depict this process by giving an example of tags before and after normalization.

Table 1: Tags used to annotate a sample web resource (http://apples-to-oranges.com/blog/examples/cssgraphs.html, Date accessed May 12, 2006 at 10:00 PM GMT) stored in the del.icio.us service (before normalization). The numbers refer to the frequency of occurrences.

123 css	18 gui	7 howto	3 stats
56 design	14 html	5 tips	2 bargraph
47 graphs	12 webdev	5 usability	2 example
46	10 reference	5 graphing	
webdesign	9	3 bar	
28 graph	development	3 coding	
27 web	8 cool		

Table 2: Tags after applying the normalization process.

123 css	10 reference	5 usability	2 example
80 graph	8 cool	5 bargraph	
18 gui	7 howto	3 code	
14 html	5 tip	3 stats	

4.1.2. Semantic Annotation Pipeline

The semantic annotation process is the backbone process that generates semantic metadata using the three ontologies. The process attempts to match folksonomy terms (after normalizing them) from the bookmarked resource against terms in the ontology (which it will work as a controlled vocabulary) and only selects those terms that appear in the ontology.

The inference engine is responsible for associating pedagogical semantics to the annotated web resource. In our system we define two pedagogical semantic terms. 'Instructional level' can be basic, intermediate or advanced and refers to where the concept fits within the domain being studied. 'Difficulty' can be easy, medium or hard, and describes how conceptually difficult this

resource will be to understand within the domain and instructional level concerned.

These two pedagogical values are generated from a set of inference rules so long as enough information is available in the basic semantic descriptors. For example, given a web resource within the domain of 'CSS' tagged with a folksonomy value of 'font' the inference engine would trigger the rule that states "if a web resource has a tag value of 'font' then its difficulty will be 'easy' and its instructional level will be 'basic'".
After finishing the annotation process, each item of the generated semantic metadata is saved in a database (e.g. a triple store) for later query by a dedicated portal.

4.2 The Portal

Is a web-based application that provides miscellaneous facets to access the generated semantic metadata. The application was implemented using Tomcat servlet engine 5.5 that runs JSP pages and used Jena 2 API for ontology manipulation.

5 EVALUATION FRAMEWORK

Figure 3 shows the overall evaluation steps needed to verify our research claims. The evaluation is divided into three parts: Metadata Assignment Evaluation, Metadata Performance Evaluation and other statistical evaluation.

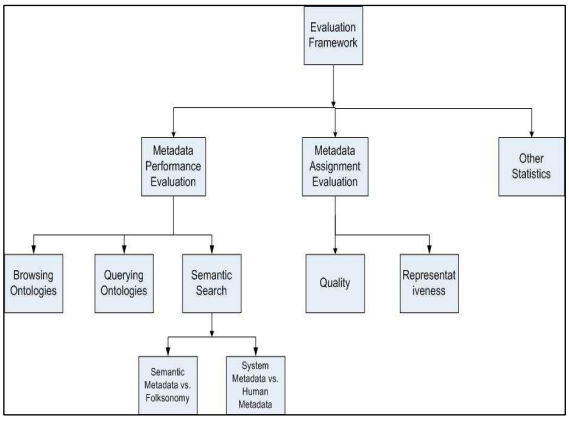

Figure 3: The proposed evaluation framework.

5. 1 Metadata Assignment Evaluation

This evaluation stage is necessary to evaluate the quality and the representative-ness of the generated semantic metadata. This can be done using qualitative evaluation techniques.

Metadata quality, as a qualitative evaluation technique, is defined by (Guy et al., 2004) "… supports the functional requirements of the system it is designed to support." Therefore, to evaluate the functional requirements of this research a set of Metadata quality questions need to be answered, which are:

1. Are the semantics of the descriptors clear and unambiguous?
2. How well does the metadata describe the resource?
3. How accurate is the generated metadata represent the web resource?

To answer these questions, a questionnaire will be designed and projected to a group of subject domain experts to rate the appropriateness of the metadata assigned. The questionnaire will measure how well the user believes the metadata predicts the actual contents of the web resource.

5.2 Metadata Performance Evaluation

Another corner stone in the evaluation mechanism is to evaluate the performance of the metadata. This implies the following questions:

1. Can the resources be accessed in different ways? i.e. not only by search.
2. Is searching by the generated semantic metadata is better than searching by folksonomies?
3. How well does automatic metadata perform compared to manual metadata?

A very well-know measurement of the success for the metadata performance in search is the Recall value (Converge measurements).

$$\text{recall} = \frac{|\{\text{relevant documents}\} \cap \{\text{retrieved documents}\}|}{|\{\text{relevant documents}\}|}$$

This measurement will be used in the Semantic Search versus Folksonomy search sub-evaluation phase.

5.2.1 Evaluation of metadata performance

In this preliminary evaluation, we will try to answer question one and two in the Metadata Performance Evaluation phase.

To measure the performance of the generated metadata, we have developed three different ways to access and retrieve the annotated web resources which include: Ontology Browsing, Ontology Querying and Semantic Search.

Ontology browsing and ontology querying add two flexible ways to reach; retrieve and search for annotated learning resources. Since the ontologies are created in a hierarchical taxonomic nature, they can be directly projected to the user as views.

1) Ontology Browsing

In this option, the user can retrieve learning resources either by browsing the concepts in the web design ontology, CSS ontology, or the resource type ontology. When a concept is selected in either ontologies all resource resembling the selected concept are retrieved along with their full description.

Figure 4 shows the user interface depicting ontologies as views. When a concept is selected by clicking on a link listed in the view an ontology-based search is initiated and shows all results returned to the user, based on the selection made.

The browsing algorithm works by reasoning over the data. Such that when a concept is selected the algorithm searches the knowledge base for all resources related to the concept.

One benefit of using the view-based search paradigm is that users can have a grand vision of all concepts provided by the domain and select concepts that represents what they are looking for.

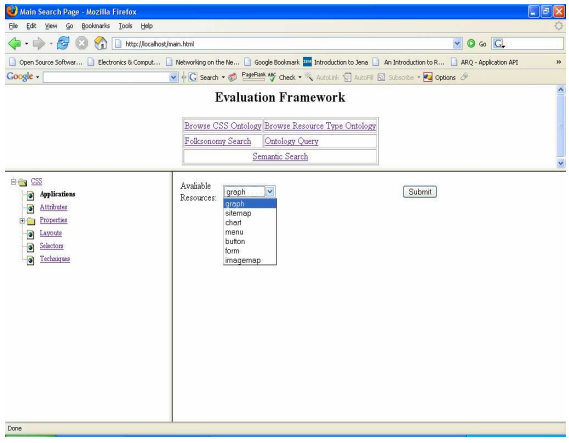

Figure 4: Browsing the CSS ontology; the left pane shows the ontology view while the right pane shows the returned results initiated by the selection made on the left pane.

2) Ontology Querying

To further enhance the experience of searching for CSS resources. A query interface has been implemented, which enables the composition of different queries to access the knowledge base. The user is presented with a set of query filters to choose from, as shown in figure 6. These include query by: resource type, difficulty, instructional level, subject, technique and application.

3) Semantic Search

To really test the performance of the generated metadata, a rigorous test needs to be applied to the semantic metadata. This includes two types of test: semantic search versus folksonomy search and folksonomy semantic metadata versus human expert semantic metadata.

A) Semantic Search versus Folksonomy search

To evaluate the performance of the generated semantic metadata, we have embarked on an evaluation procedure adopted from (Li et al., 2005), where they compared keywords against semantic topic search. However, in our system we have compared the performance of folksonomy search against semantic topic search to see which search results in more relevant records.

We used the option of semantic search in our portal to allow us to search CSS topics (e.g. BoxModel, Layout, Navigation, Positioning and Typography) in two ways. For the first search we queried the folksonomy for the chosen topic and in the second search we conducted a semantic search on the CSS ontology for the same topic.

In some cases the number of resources returned by the semantic search is higher as the semantic search benefits from the relationship between topics in the CSS ontology, in this case the 'related_to' relationship which links between related concepts. For instance, when someone searches for the topic 'positioning', all resources that have as their subject the word 'positioning' plus all related resources will be retrieved. Table 3 shows the result obtained when searching for the (positioning and navigation) topics in the CSS ontology.

These results demonstrate that the semantic search outperforms folksonomy search in our sample test, this is because folksonomy search, even if the folksonomy keywords were produced by humans, is analogous to keyword search and therefore limited (Motta & Sabou, 2006).

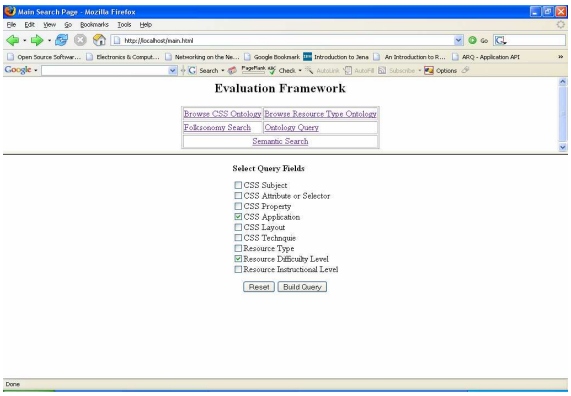

Figure 5: Ontology query filters selection.

Table 3: The Relevance Result between Folksonomy Search and Subject Search Using the CSS ontology.

CSS Topic	*Positioning*	
	Folksonomy Search	**Semantic Search**
Number of records found	3	4
Number of records relevant to topic	3/4	4/4
CSS Topic	*Navigation*	
	Folksonomy Search	**Semantic Search**
Number of records found	1	2
Number of records relevant to topic	1/2	2/2

B) Automatic Metadata versus Manual Metadata
In this sub-evaluation stage, we intend to ask a subject expert in the domain of CSS to annotate a set of CSS resources given our ontologies and then feed the annotated resources to our portal and perform semantic search.

The rational of this evaluation step is to check whether the automatic generated metadata using people's tags is more or less the same as an expert assigned metadata.

5.3 Other Evaluation Factors and Statistics

The researchers are planning to evaluate the effectiveness of the various stages in the FolksAnnotation tool; which includes:

The evaluation of the effectiveness of the normalization process, i.e. the size of the tag set before and after normalization.

What are the tags that are not used, why they have not been used and how can they be used?

The relation between the number of people who bookmarked a web resource and the granularity of the generated metadata.

6 CONCLUSIONS

In this paper we have reported on the status of an ongoing research to investigate the possibility of using folksonomies as a media for semantic annotation. Our aim in this research was to show that semantic metadata can be potentially generated using folksonomies guided by domain ontologies. And to some extent we tried to show that part of our claim is valid by reporting on the results of the possible evaluation steps we have embraced.

REFERENCES

Al-Khalifa, H. S. and Davis, H. C. (2006). FolksAnnotation: A Semantic Metadata Tool for Annotating Learning Resources Using Folksonomies and Domain Ontologies. Proceedings of the Second International Conference on Innovations in Information Technology. IEEE Computer Society, Dubai, UAE

Guy M., A. Powell and M. Day. (2004). Improving the Quality of Metadata in Eprint Archives. Ariadne Issue 38. URL: http://www.ariadne.ac.uk/issue38/guy/intro.html

Li, J. Z., Gasevic, D., Nesbit, J. C., & Richards, G. (2005). Ontology Mappings Enable Interoperation of Knowledge Domain Taxonomies. Paper presented at the 2nd LORNET international annual conference, November 16-18, 2005, Vancouver, Canada.

Menchen, E. Feedback, Motivation and Collectivity in a Social Bookmarking System. in Kairosnews Computers and Writing Online Conference. 2005.

Motta, E. and M. Sabou. Language (2006) "Technologies and the Evolution of the Semantic Web." In Proceedings of LREC , Genoa, Italy, 24-26 May.

Quintarelli, E. (2005) Folksonomies: power to the people. in ISKO Italy-UniMIB meeting. 2005. Milan, Itlay.

Shirky, C. (2005). Ontology is Overrated: Categories, Links, and Tags. Accessed on March 27, 2006. Available on line http://shirky.com/writings/ontology_overrated.html

Sieck, S., connotea and citeulike: "folksonomies" emerge within scholarly communities, E. Insight, Editor. 2005

Vander Wal, T. (2004). Folksonomy definition and wikipedia. Accessed on April 29, 2006.Available online http://www.vanderwal.net/random/category.php?cat=153

Wikipedia. Folksonomy. Accessed on March 26, 2006.Available on line http://en.wikipedia.org/wiki/Folksonomy