FAsTA: A Folksonomy-Based Automatic Metadata Generator

Hend S. Al-Khalifa and Hugh C. Davis

Learning Societies Lab
ECS, The University of Southampton
SO17 1BJ, Southampton, UK
hsak04r/hcd@ecs.soton.ac.uk

Abstract. Folksonomies provide a free source of keywords describing web resources, however, these keywords are free form and unstructured. In this paper, we describe a novel tool that converts folksonomy tags into semantic metadata, and present a case study consisting of a framework for evaluating the usefulness of this metadata within the context of a particular eLearning application. The evaluation shows the number of ways in which the generated semantic metadata adds value to the raw folksonomy tags.

1 Introduction

Folksonomy, a term coined by Thomas Vender Wal in 2005, is a mechanism to describe web resources using people's own vocabulary. As defined by an article in Wikipedia[1] folksonomy is *"... an Internet-based information retrieval methodology consisting of collaboratively generated, open-ended labels that categorize content such as Web pages, online photographs, and Web links*."

Users have their own perspective when tagging a resource; they may add new contextual dimensions, for example to suggest its application or its relationship to neighboring domains. This effect has been witnessed in our domain of study (Web Design with Cascading Style Sheets 'CSS'), where people tag resources appearing in that domain with extra contextual dimensions such as the application of a web resource, its type and other parallel domains for instance 'PHP programming'.

Clearly, folksonomies are a potential source of useful metadata. As Peterson [1] said "*The overall usefulness of folksonomies is not called into question; just how they can be refined without losing the openness that makes them so popular*". In our work, rather than attempting to refine the tagging process we have taken the open vocabulary tags and mapped them against domain ontologies in order to derive structured semantic metadata from the folksonomies. This paper describes our tool, its evaluation and shows that folksonomies contain acceptable indexing words that can create semantic metadata with added value.

[1] http://en.wikipedia.org/wiki/Folksonomy (27thMarch 2007)

2 Methodology

The semantic metadata elements used to describe CSS web recourses were constructed by mixing elements from the IEEE LOM standard and elements specific to the domain of CSS, in other words, creating a domain specific application profile from IEEE-LOM. The application profile consists of 15 elements, which include: *Title, Description, Keywords, Resource Type, Recommendation, Property, Selector, Unit, Attribute, Technique, Application, Subject, Layout, Difficulty level and Instructional level*.

In order to produce the CSS semantic metadata from folksonomy tags, we have implemented a tool that extracts tags form URLs talking about CSS in del.icio.us and utilizing these tags in the process of semantic metadata generation. Herein, we briefly present our tool, namely the FolksAnnotations Tool Architecture (FAsTA) and its components, however, for a full tool description the reader is referred to [2].

The main two processes used in FAsTA are: the Tags Extraction and Normalization pipeline and the Semantic Annotation pipeline.

The Tags Extraction and Normalization pipeline starts by fetching a bookmarked web resource from the del.icio.us bookmarking service, then the tag extraction process begins by extracting folksonomy tags from the web page of the bookmarked web resource. The extracted tags are then passed to the normalization process which performs a series of filters to clean the tags. The filters are preformed sequentially in the following order:

- **Lower-case filter**: Tags are converted to lower case,
- **Non-English filter**: Non-Roman Alphabet are dropped; this step is to insure that only English tags are present when doing the semantic annotation process,
- **Stemming filter**: stem tags using a modified version of the Porter Stemmer (http://www.tartarus.org/~martin/PorterStemmer/),
- **Tags sense Disambiguation filter**: stemmed tags are passed to this module to remove ambiguous tags, i.e. polysemy.
- **Grouping filter**: similar tags are grouped (e.g. inclusion of substrings),
- Finally, the **removal filter**, where the general concept tags in our domain of interest (e.g. programming, web, etc) and ambiguous tags are eliminated.

The process of normalization is done automatically and it is potentially useful to clean up the noise in people's tags. The normalized tags list is then passed to the semantic annotation process, where each normalized folksonomy tag is mapped to a corresponding ontological instance in one of the three ontologies, which are: the Web Design Ontology, the CSS Subject Ontology and the Resource Type Ontology [2]. This process will attach ontology instances as descriptors for a web resource.

3 Evaluations and Results

To evaluate the output of our prototype tool, many evaluation aspects need to be considered, including the usefulness, the quality and the representativeness of the generated metadata semantics.

Barritt and Alderman [3] determines the usefulness of metadata from two viewpoints: validity, i.e. creating valid metadata for every learning resource, and searchability, having the search tools in place to use that metadata. Guy et al. [4] defines metadata quality as "… *supports the functional requirements of the system it is designed to support*." Thus, to stipulate the 'functional requirements' of the current work, we have considered that the semantic metadata needs to have no errors and the semantic descriptions need to correctly reflect the nature of the described web resource. Finally, the representativeness of a semantic metadata can be thought of as how well the metadata descriptors describe the semantics of the given domain, in this case the domain of Web design with CSS.

Therefore, to evaluate these different aspects, we have implemented an evaluation framework that consists of the following procedures:

- Metadata assignment evaluation, which consist of:
    - Metadata Representativeness.
    - Metadata Quality and Validity.
- Identifying niche tags in 'The Long Tail': this procedure investigates whether distinguishable values of the semantic metadata elements come from rare tags residing in 'The Long Tail'.

3.1 Metadata Assignment Evaluation

The metadata assignment evaluation stage is necessary to evaluate the quality, validity and representativeness of the generated semantic metadata record.

To verify these requirements, we used a blend of quantitative and qualitative evaluation techniques. Thus, to evaluate the previous requirements a set of questions need to be answered, which are:

- Are the semantics of the descriptors *clear* and *unambiguous*?
- How well does the metadata *describe* the resource?
- How accurate is the generated metadata *represents* the web resource?

To answer these questions, a questionnaire was designed and distributed to a group of subject domain experts to rate the appropriateness of the descriptors and the validity of the assigned metadata. The questionnaire also measured how well the respondent believes that the metadata predicts the actual contents of the web resource. The questionnaire was distributed to two target populations (web designers and experts in the field of learning technologies and metadata, i.e. 'specialists'.). The web designers' community was reached using mailing lists that reside at Yahoo Groups or other focused groups such as css-discuss.org. The total response from the web designers group was 29 respondents. The specialist group was reached by distributing the questionnaire to the CETIS-Metadata mailing list and to colleagues from the Learning Societies Lab Research Group (LSL) at the University of Southampton, UK. The total number of respondents from the specialist group was 19.

3.1.1 Metadata Representativeness

Two questions in the questionnaire were designed to capture the respondents view on the representative-ness of the metadata elements. The first question handles the descriptors of CSS web resources and the second question handles the required fields needed to search for CSS web resources. The respondents were asked to rate (based on a scale from 1 to 5 where 1 represents 'useless' and 5 represents 'very useful') how useful each metadata element was to describe and search for web resources in the domain of teaching web design with CSS.

For the question asking about '*how useful are the metadata descriptors used to **describe** a CSS web resource*'. The overall statistics for the web designers' group responses show that the mean of the metadata elements are all above average, except for one element which is slightly below midpoint. However, the standard deviation for all elements is quite high, which indicates the varied view between respondents.

On the other hand, the overall statistics for the specialist group responses show that the mean of the metadata elements are all above average with a quite high standard deviation for all elements, except for two elements which indicated an agreed view in their importance between respondents.

For the question asking about '*how useful are the metadata descriptors used to **search** for a CSS web resource*'. The overall statistics for the web designer's group responses show that the mean of the metadata elements are all above average, except for one element, again, which is slightly below midpoint. However, the standard deviation for most elements is quite high, which indicates the varied view between respondents, expect for two elements which indicates some consistency on the respondents rating towards these two elements. By comparing the means of all elements, it is apparent that most elements are equally likely useful descriptors for retrieving/searching for a CSS web resource.

In contrast, the overall statistics for the specialist group responses show that the mean of the metadata elements are all above average, except for three elements which were slightly below midpoint. However, the standard deviation for half of the elements was quite low, which indicates consistency in the respondents' view of these elements.

3.1.2 Metadata Quality and Validity

The questionnaire was also designed to include a question about the quality and validity of a random sample of three CSS web resources metadata records. These three automatically generated semantic metadata records were selected based on their coverage of the various aspects of the CSS metadata descriptors. Therefore, the three metadata records were exposed to both groups (web designers and specialist) to rate them based on a metric produced by Greenberg [5] to evaluate the quality and validity of metadata elements. The evaluation is based on a three-tier scale, which are: Good, Fair and Reject.

The results of the quality and validity for each metadata element of the three resources were assessed for each element. Thus, for the three annotated web resources both the web designers group and the specialist group agreed in giving the following metadata elements: *Title, Resource type, Subject, Application, Technique, Property, Attribute and Layout*; either a 'Good' or 'Fair' rate. However, the two groups diverge in their opinion of the rest of the metadata elements which are: *Description*, *Keywords* and *Selector*. In the specialist group they rate these elements as 'Fair', 'Good' and 'Fair' respectively; while, the web designers group has rated them as 'Reject'.

3.2 Exploring The Long Tail

As we were evaluating our generated semantic metadata, we observed that most fine grained semantics of the CSS domain came from minority tags. Thus, some niche folksonomy tags from the CSS ontology create a finer-grained indexing for a web resource. This observation helped us to form the following hypothesis: "*Fine-grained metadata values come from The Long Tail*".

The Long Tail, as defined in Wikipedia[2]: "*…The long tail is the colloquial name for a long-known feature of statistical distributions ... In these distributions a high-frequency or high-amplitude population is followed by a low-frequency or low-amplitude population which gradually "tails off."* "

To verify our hypothesis we analyzed the distribution of the list of tags used to semantically annotate web resources in our data set. One observation we found when compiling the list of tags used to create the semantic metadata was that the distribution of all tags that are used for semantically annotating a web resource always yields a long tail shape, as shown in Fig. 1. Notice that the tags 'list' (1 time), 'menu' (2 times), 'button' (9 times) and 'rollover' (10 times), are niche instances from the CSS ontology and at the same time fall in 'The Long Tail' region.

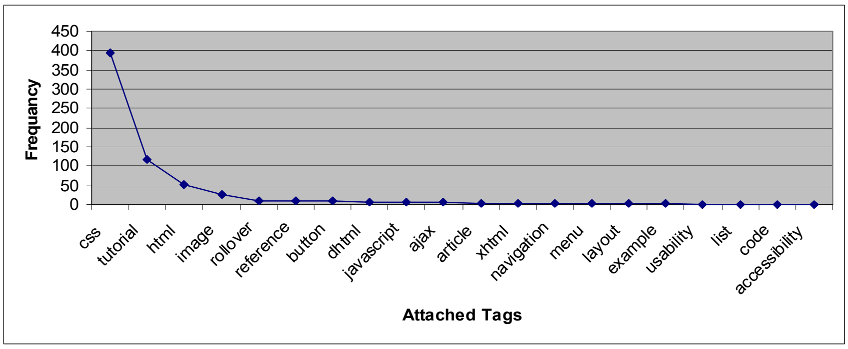

Fig 1. The Long Tail shape for the tags used to semantically annotate the "What Are CSS Sprites? A Quick Example: Button Rollovers" web resource.

Consequently, we examined the graph of each web resource tags list to determine the tags that fall within 'The Long Tail' portion and found that from the 100 annotated web resources 80% have one or more niche-tags. The average portion of niche-tags for all web resources was 16% with a standard deviation of 11.77%. This implies that on average 16% of the used tags for each resource will be a niche-tag. This finding verifies our claim about the source of the fine-grained metadata values.

[2] http://en.wikipedia.org/wiki/The_Long_Tail (27th March 2007)

4 Conclusion and Future Work

In this paper we have showed how we successfully managed to convert folksonomy tags into useful semantic metadata. In previous work [6] we have compared the semantic metadata generated to the keywords extracted using context based keyword extraction technique, and demonstrated the improved value of the folksonomy tags.

In this work we have described a framework to evaluate and demonstrate the usefulness, the quality and the representativeness of the generated semantic metadata. Based on our evaluation framework, our findings can be summarized in three points:

1. Folksonomy tags demonstrated that they are 'good enough' source for creating semantic metadata. This might be attributed to the latent (implicit) semantics embedded in the tags used to describe web resources. The observed latent semantics helped us to build the appropriate ontologies that captured folksonomy semantics and converted folksonomy tags to semantic metadata.
2. Folksonomy tags showed the power of aggregating people's intelligence which helped in producing meaningful metadata. This was done without requiring their consensus in choosing the tags.
3. We have shown that useful fine grained metadata values in our case study came from The Long Tail. These values played a prominent role in distinguishing the metadata of a given web resource from other equivalent resources.

Finally, there are many potential extensions that could enhance the tool performance and output. The extensions could include: expanding the semantic metadata and ontologies, improving the normalization pipeline, and performing further evaluation procedures such as a comparative study to compare our tool performance against other automatic metadata generators.

References

[1] Peterson, E., *Beneath the Metadata: Some Philosophical Problems with Folksonomy.* D-Lib Magazine, 2006. **12** (11).

[2] Al-Khalifa, H. S. and Davis, H. C. Replacing the Monolithic LOM: A Folksonomic Approach. In Proceedings of ICALT 2007. (in press), Niigata, Japan.

[3] Barritt, C. and F. Alderman, *Creating a Reusable Learning Objects Strategy.* 2004, San Diego: Pfeiffer.

[4] Guy, M., A. Powell, and M. Day, *Improving the Quality of Metadata in Eprint Archives.* Ariadne, 2004. **January**(38).

[5] Greenberg, J., *Metadata extraction and harvesting: A comparison of two automatic metadata generation applications.* Journal of Internet Cataloging, 2005. **6**(4): p. 59-82.

[6] Al-Khalifa, H.S. and H.C. Davis, *Exploring The Value Of Folksonomies For Creating Semantic Metadata.* International Journal on Semantic Web and Information Systems (IJSWIS), 2007. 3(1) pp. 13-39.