Towards Better Understanding of Folksonomic Patterns

Hend S. Al-Khalifa & Hugh C. Davis
Learning Societies Lab
ECS, Southampton University
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 their semantics spans multiple contextual dimension. In this paper, we present a pragmatic experiment that analyzes folksonomies using three classification categories: *Personal*, *Factual* and *Subjective*, in order to gain more understanding of the types of tags used in the social tagging process. The rational for this work was to measure the potential portion of folksonomy tags that might be helpful when considering the creation of structured metadata.

Categories and Subject Descriptors
H.3 [**Information Storage and Retrieval**]

General Terms
Design, Experimentation.

Keywords
Folksonomy, Collaborative tagging, Social bookmarking.

1. INTRODUCTION
Lately, collaborative tagging has been gaining momentum in library science, information systems and related fields. Social bookmarking services such as del.icio.us, where people bookmark and share their favorite web links, are becoming a plentiful source of cheap metadata that are called folksonomy.

Folksonomy, a term coined by Thomas Vender Wal in 2005, is a mechanism to describe web resources using people's own vocabulary. Or as defined in Wikipedia (from http://en.wikipedia.org/wiki/Folksonomy, accessed 1st April 2007) 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.*"

Contrary to traditional categorization systems, folksonomy is a user-generated labeling system, where people's chosen vocabulary is used rather than librarian or author-generated vocabulary. Thus, folksonomies are becoming a focal point for web resource discovery.

Users have their own perspective when describing web resources. They may describe a resource by its type, discipline, content or they may even add new contextual dimensions to it to visualize its application or relation to other neighboring domains.

To distinguish between folksonomy tags and contextual keywords extracted by context-based keyword extractor, we have in [5] compared folksonomy tags to keywords extracted using a context based keyword extraction technique, and demonstrated the improved value of the folksonomy tags over context-based keywords. However, to gain better understanding of the types of tags generated by the users of social bookmarking services, this paper describes our attempt to explore the semantics of folksonomy tags by classifying them into three genres and providing an in depth analysis of the tags falling in the three categories. Also this attempt will help us to shed some light on the potential use of folksonomy tags in metadata creation.

Our experimental dataset was selected to be from the popular social bookmarking service del.icio.us with an emphasis on the domain of Cascading Style Sheets (CSS).

2. METHODOLOGY
In order to manually analyze folksonomy tags, we first implemented a tool that fetches tags associated with a bookmarked web resource in del.icio.us and then clean the noise in these tags and set them ready for analysis. The tool, as shown in Figure 1, consists of an extraction module and a normalization pipeline module.

The extraction module is responsible of fetching a bookmarked web resource along with its associated tags and passing the extracted tags to the normalization pipeline for cleaning.

The normalization pipeline contains four filters:

1. **Lower-case filter**: Tags are converted to lower case so that string manipulation (e.g. comparison) can be applied easily,
2. **Non-English filter**: Non-Roman alphabets are dropped; this step is to insure that only English-like tags are present when doing the analysis,
3. **Stemming filter**: Tags are stemmed (e.g. plurals converted to singular) using a modified version of the Porter Stemmer.
4. **Grouping Similar Tags filter**: identical tags and substrings are grouped and their occurrence is counted.

HT'07, September 10-12, 2007, Manchester, United Kingdom.

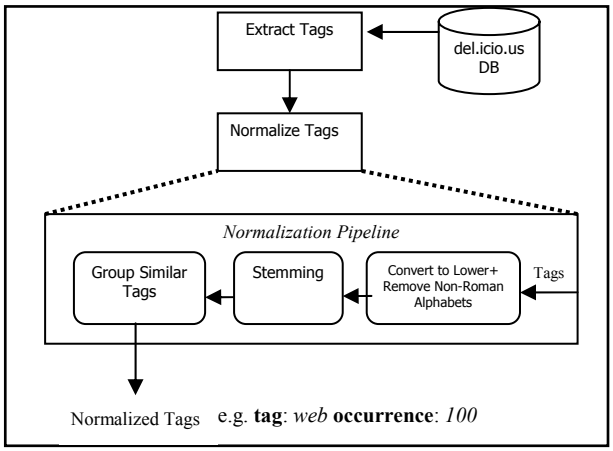

Figure 1. The normalization pipeline.

2.1 Data Set

Our data set consist of a sample of 100 randomly selected bookmarked web resources from the domain of CSS with a total number of 72,458 posts (i.e. a post is a resource tagged by a user). The total number of tags before normalization was 245,892, and the number of tags after normalization was 10,900. These 10,900 folksonomy tags were manually inspected and classified based on the Sen et al. [1] classification, as will be described in the next section.

2.2 Classification Scheme and Heuristics

To evaluate the folksonomy tags a classification scheme from Sen et al., which was adopted and modified from Golder and Huberman [2] to categorize folksonomy words, was used.

Sen et al. have classified folksonomy tags into three groups:

- **(P)ersonal tags**: "have an intended audience of the tag applier themselves. They are often used to organize a user's own resources (self-reference, task organization, time management) e.g. 'myblog'.
- **(S)ubjective tags**: express people opinions related to a web resource e.g. 'cool', and
- **(F)actual tags**: identify 'facts' about the described web resource such as people, places, or concepts e.g. 'tutorial'."

To use the Sen et al. classification, we modified it by adding some additional heuristics, which are:

- Tags occurrences were used as an indicator of the agreed meaning of it; therefore, lower tags occurrence indicates personal use.
- Compound tags and vague abbreviations are considered personal, since no one knows what do they mean, or why they were formed in this shape. In addition, their tag occurrence is at its minimum.
- Misspelled tags are not counted in the classifications.

2.3 Analysis Results

Figure 2 shows the overall distribution of the inspected folksonomy tags after classifying them into three categories: 34% were personal tags, 62% were factual tags and 4% were subjective tags.

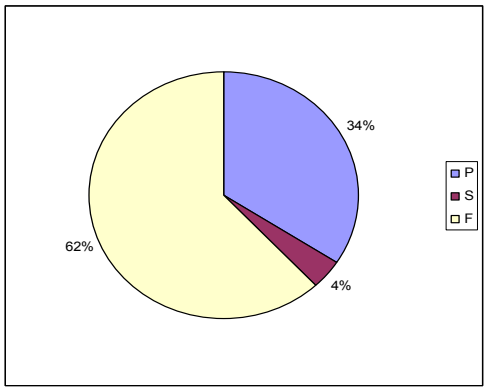

Figure 2. Classification of Folksonomy tags.

From our manual inspection we also found that tags falling in one of the previously mentioned categories can be further classified into an ad hoc classification which includes: abbreviations, acronyms, complete words (singular/plural variations), compound words or typos and non-English words as shown in Table 1.

Classification	P	S	F
abbreviations & acronyms	10%	-	1%
complete words	27%	86%	96%
compound words	57%	13%	3%
typos & non-English words	6%	1%	-

Table 1: A rough estimation of the ad hoc classification

Thus the inspected folksonomy tags stressed the previously mentioned ad hoc classification and helped in revealing more concise general pattern in people's compound tags; despite the dynamic nature of the compound tags.

The devised general pattern for the compound tags can be formulated as follows:

$$[P^*,S^*,F^*]_{3!}$$

P, S and F each of which can be either an abbreviation, acronym or complete word (singular and/or plural variations). (3!) indicates the possible number of permutations between the three categories (hence 6 possible orders) and (*) indicates that there are zero or more possible occurrences of the category.

To illustrate this pattern some examples of actual compound tags from our data set are presented in the following form ([*pattern*], *example*):

([F,F,F], bugscrossbrowser), ([F,P], cssotherproject), ([F,F], csslib), ([F,S], csswelldone), ([S,F,F], nicesiteportal), ([S,F], goodexample), ([P,P/F,P],personaltoolbarfolder), ([S,F,S], naughtyexerciseevaluation).

Finally, spelling errors constituted around 6% of the entire folksonomy tags. By examining the types of errors that people have generated, we have found that some users inserted an extra character by mistake when typing the tag e.g. (termplate), or switched the places of characters e.g. (hmtl), or even missed a character e.g. (tutoria). These misspelled tags are usually used only by the person who created them and they do not gain much attention from other del.icio.us users. Next, a detailed analysis of the folksonomy tags that falls in the main three categories is discussed in depth.

2.3.1 (P)ersonal tags

These are tags that have an intended audience. They are often used to organize a user's own resources, and can be roughly classified into: self-reference tags, task and time management tags and others [4].

Self-reference tags classification, includes any tag that has to do with the user own interest. Such as dates e.g. (January, monthly and night), names e.g (tojack) and own reference e.g. (mylink, mysite and myblog). These tags usually appear once or twice among all tags in a given bookmarked web resource.

On the other hand, the most frequent task and time management tags were 'howto', 'tip', 'toread', 'work', 'todo' and their varieties such as 'readlater', 'todescribe', 'tostudy', etc. These tags tend to function as reminders and to-do lists to manage someone's future activities.

Foreign tags were also spotted, these tags use Roman Alphabets in their writing; thus it can not be removed in the normalization pipeline. Such examples include the frequent Spanish tag 'herramienta' which means 'tool' and the Portuguese tag 'Artigo' which means 'Article'.

Also there are some occurrences of prepositions in the tags list such as 'for', 'with', 'and', 'one' and 'in'. These prepositional tags might be inserted unintentionally by a user who is thinking that the del.icio.us service deals with sentences/phrases as whole tags. A quick examination of the tags list that contains these prepositions justifies our assumption.

Finally, compound tags with minimum tag count are considered personal tags, since no other del.icio.us user has used them. These tags constitute around 30% of all personal tags. By the same token, abbreviations are considered personal tags since no one knows what are their intended meaning more than the person who created them. These tags constitute 6% of all personal tags.

2.3.2 (S)ubjective tags

These are tags that express people's opinions on the bookmarked web resource. Although these tags constitute a small portion of the inspected folksonomy tags (i.e. 4%), an in depth inspection was carried out to analyze them.

Two classifications were observed in subjective tags: either the subjective tags were compound or informal. The compound tags consisted of a subjective qualifier with either a factual or personal tag, e.g. 'beautifulsite', 'goodfor'. Informal subjective tags include words that are produced by the user's own vocabulary such as 'Kool' or 'kickass'.

2.3.3 (F)actual tags

These are tags which identify 'facts' about the described web resource such as people, places, or concepts. A more specific rough classification can be: *web resource title/URL/author, synonyms (either near or far), rights/language, compound tags, generic, acronyms, spelling variation and other areas of application or usage*.

Usually del.icio.us users use the explicit title of a web resource, the author or words appeared in the URL to bookmark the web resources. This pattern might be used because it is easy to remember a bookmark category by its title, author or URL. To give an example most users who bookmarked articles from A-List-Apart (*alistapart.com*) website have used tags such as: 'ala', 'alistapart' and 'zeldman' (a popular author in the website).

Another notable category was the use of synonym tags (with their two types near & far). Near synonyms mean that the average person can use the tag (i.e. usual vocabulary), and far synonyms mean that only the elite user can use the tag. As an example, the tag 'library' is an instance in the resource type ontology, which means a collection of things. On one hand, 'database' and 'collection' are two near tags that can be used as synonyms by the casual del.icio.us user, this is evident when more than one user uses these tags,. On the other hand, the tag 'Grid' is considered a far synonym because the average user can not predict it as a straight forward synonym; this was evidence from the number of users who used this tag; i.e. used by only one user.

The *rights* tags are used to indicate the privilege to use a web resource, e.g. 'free', 'opensource', 'freedom', etc. These tags constitute potential useful information to populate the rights element in a typical metadata record. Likewise, the *language* tags indicate the language of a web resource. The language tag comes in different forms (complete words, abbreviations, spelling error, etc.), e.g. 'English', 'langen', 'en', etc. Also this type of tags can be useful to populate the language element in a typical metadata record.

Compound tags took a good share of all the tags in the factual category. These are different from the compound tags mentioned in the personal category, such that more than one del.icio.us user have used them. The use of compound tags might be due to people who are trying to preserve the maximum amount of facts about a web resource in one tag. It also shows that people are mixing generic tags with more specific ones to qualify them, e.g. 'cssarticle'.

Generic, acronyms and spelling variation, although there were not much of these tags, yet, they formed a noise in the tags lists.

Finally, sometimes factual tags refer to other areas of application or usage of the tags. For example in a web resource that talks about the 'shadow' technique in CSS, one tagger have used the tag 'dreamweaver', this might indicate that this technique can be used in the Dreamweaver software, however, by inspecting the content of the web resource, it did not mention any thing about the software package or its usage in CSS.

2.3.4 Discussion

As mentioned by Guy et al. [4] the problem of folksonomy tags include typographical errors and spelling variations, and this was also evident in our inspected controlled domain.

Another observation is that in compound tags users tend to mix and match different tags forms (plural and singular), e.g. 'inspirationscss' (plural + singular), 'inspirationcss' (singular + singular), or tags where one of them containing spelling errors, e.g. 'tuotorialcss'. These acts make it very difficult to build the best normalization process. This also raises the issue of the need of more sophisticated natural language processing techniques.

The rights and language tags were two good sources of more information about the resource, however, their inconsistence appearance make it difficult to capture them in a general form.

Finally, the analysis of folksonomy tags uncovered an important finding in this work which states: not all tags can be considered useful for metadata creation, thus in our case study 63% of

folksonomy tags were facts about a web resource. This finding can be attributed to the variations in people vocabulary and their different background knowledge.

3. RELATED WORK

During the past couple of years many research were carried out to study and understand the behavior of folksonomy tags. Among these research was a study by Golder and Huberman [2], from HP Labs, who analyzed the structure of collaborative tagging to discover the regularities in user activity, tag frequencies, the kind of tags used and bursts of popularity in bookmarked URLs in the del.icio.us system. They developed a dynamic model that predicts the stable patterns in collaborative tagging and relates them to shared knowledge. Their results show that a significant amount of tagging is done for personal use rather than public benefit. However, even if the information is tagged for personal use other users can benefit from it. They also state that del.icio.us, for most users, functions as a recommendation system even without explicitly providing recommendation. Since their study focused on the user side; no statistical results were provided regarding the distribution of folksonomy tags.

Moreover, Sen et al. [1] have collapsed Golder and Huberman's classes into three general classes and used the modified classification metric to evaluate *The MovieLens* recommender system. They manually inspected 3,263 folksonomy tags and apparently their findings regarding the factual category do agree with our classification findings, i.e. they found that 63% of their inspected tags were factual. However, our findings differentiate to the findings of Sen et al. in the categories of personal and subjective, in Sen et al. 29% were subjective and 3% were personal. This differentiation might be due to two factors: (1) the methodology Sen et al. have carried out to classify the tags is different than ours, i.e. they used two people to categories tags while we relied on one, (2) the systems studied are completely different (The *MovieLens* system versus del.icio.us).

Finally, Kipp [3] has examined the differences and similarities between user keywords (folksonomies), the author and the intermediary (such as librarians) assigned keywords. She used a sample of journal articles tagged in the social bookmarking sites citeulike (http://www.citeulike.org/) and connotea (http://www.connotea.org/), which are specialized for academic articles. Her selection of articles was restricted to a set of journals known to include author assigned keywords and to the INSPEC database, so that each article selected would have three sets of keywords assigned by three different classes of metadata creators. Her methods of analyses were based on concept clustering via the INSPEC thesaurus, and descriptive statistics. She used these two methods to examine differences in context and term usage between the three classes of metadata creators. Kipp's findings showed that many users' terms were found to be related to the author and intermediary terms, but were not part of the formal thesauri used by the intermediaries; this was due to the use of broad terms which were not included in the thesaurus or to the use of newer terminology.

Kipp's work is different from ours in the following respect: (1) her classification scheme was based on thesaurus classification, and (2) her dataset was chosen from social bookmarking systems that are specialized for academic articles.

4. CONCLUSION AND FUTURE WORK

In previous work [1] we have compared folksonomy tags to keywords extracted using context based keyword extraction technique, and demonstrated the improved value of the folksonomy tags over context-based keywords. However, in this work we showed that folksonomy tags have the potential to be transformed into meaningful metadata by classifying them into semantic categories. This was proven by showing that a great shear of folksonomy tags were indeed factual. This successful classification of folksonomy tags into meaningful semantics can open the doors for future research in processing and using folksonomy tags in creating structured metadata that adhere to pre-defined ontologies, as we have done in [6]. Finally, to the best of our knowledge this is the first time an experiment with this magnitude was conducted manually to inspect the generic semantics of folksonomy tags (cf. [3]).

5. REFERENCES

[1] Sen, S., S. Lam, A. Rashid, D. Cosley, D. Frankowski, J. Osterhouse, M. Harper, and J. Riedl. Tagging, communities, vocabulary, evolution. In Proceedings of CSCW '06. 2006. ACM.

[2] Golder, S.A. and B.A. Huberman, The Structure of Collaborative Tagging Systems. Journal of Information Science, 2006. 32(2): p. pp. 198-208.

[3] Kipp, M.E. Exploring the context of user, creator and intermediate tagging. In IA Summit 2006. 2006. Canada.

[4] Guy, M. and E. Tonkin, Folksonomies: Tidying up Tags? D-Lib Magazine, 2006. V 12(1).

[5] Al-Khalifa, H.S. and Davis H. C., Exploring the Value of Folksonomies for Creating Semantic Metadata. International Journal on Semantic Web and Information Systems (IJSWIS), 2007. 3(1): p. pp. 13-39.

[6] Al-Khalifa, H.S. and Davis H. C., FolksAnnotation: A Semantic Metadata Tool for Annotating Learning Resources Using Folksonomies and Domain Ontologies. In Proceedings of the 2ed IEEE Conference on IIT. 2006. Dubai, UAE.