Longitudinal Analysis of Tag Structure in Del.icio.us

Lijiang Guo  
Indiana University Bloomington  
1320 E. 10th St., LI 011  
Bloomington, IN 47405  
ljguo@indiana.edu

Elin Jacob  
Indiana University Bloomington  
1320 E. 10th St., LI 011  
Bloomington, IN 47405  
ejacob@indiana.edu

Nicolas George  
Indiana University Bloomington  
1320 E. 10th St., LI 011  
Bloomington, IN 47405  
ngeorge@indiana.edu

ABSTRACT
This paper describes a three-level structure of folksonomies that accounts for the aggregation of tags in a social bookmarking system and describes the results of a preliminary longitudinal analysis of user-assigned tags collected from del.icio.us.com for the period 2005-2007. Results of this analysis indicate that evolving community consensus on the meanings of tags can lead to the emergence of domain vocabularies that can be useful for retrieving domain resources.

Categories and Subject Descriptors
H.3.1 [Information Storage and Retrieval]: Content Analysis and Indexing -- indexing methods, thesauruses.

General Terms
Theory, Verification.

Keywords
Tags, Tagging, Folksonomies, Network folksonomies, System folksonomies, Social bookmarking systems, del.icio.us.com.

1. INTRODUCTION
Adding metadata to digital resources has become a common way of representing them for future retrieval. Metadata and metadata schemes are generally created by information professionals, but such methods encounter limitations in the environment of the World Wide Web (Web), where an enormous and dynamic repository of digital resources has made representation an important issue. Folksonomies are claimed to support an emergent classification and the use of natural language descriptors, tagging creates a feedback loop of asymmetric communication between users through the medium of the tags themselves (Mathes, 2004), allowing users to negotiate meaning and reach consensus about the referents of tags. On a social bookmarking site, there are at least two kinds of vocabulary: the user’s vocabulary and the system vocabulary. Each user has his own collection of tags and tag-URL assignments, which comprise that user’s unique vocabulary (a folksonomy). The vocabulary of the system (the system folksonomy) is the aggregation of all user vocabularies (folksonomies).

Despite the problems of reference that come with cognitive categorization and the use of natural language descriptors, tagging behavior in a network folksonomy appears to demonstrate patterns of stabilization and convergence. One possible reason is that tagging creates a feedback loop of asymmetric communication between users through the medium of the tags themselves (Mathes, 2004), allowing users to negotiate meaning and reach consensus about the referents of tags. On a social bookmarking site, there are at least two kinds of vocabulary: the user’s vocabulary and the system vocabulary. Each user has his own collection of tags and tag-URL assignments, which comprise that user’s unique vocabulary (a folksonomy). The vocabulary of the system (the system folksonomy) is the aggregation of all user vocabularies (folksonomies).

However, to make a user’s unique vocabulary communicative, agreement on vocabulary must be reached across users. Wittgenstein’s (1958) notion of a language game describes such a dynamic system. On the one hand, each user has a private language that is only known to the person speaking (or tagging); on the other hand, it must be possible, in principle, to align this with public standards and criteria for correctness. Therefore, the research question addressed here is whether such a dynamic can
contribute to stabilization and a shared domain vocabulary (folksonomy network).

The basic theoretical model of this research resides in the intersection of classical classification and categorization theory, human cognition, motivation for learning, and complex network systems. Adapted from the formal model of Hoito, Staab and Stumme (2006), a folksonomy network is defined as:

Definition 1. A folksonomy network is a tuple $F_n := (T, R, P)$ where,
- $F$ is a folksonomy that has $N$ folksonomy networks,
- $F_n (n \in N)$ is a folksonomy network of that folksonomy,
- $T$ and $R$ are subsets of tags and URLs in $F$,
- $f_R$ is the characteristic function a folksonomy network $F_n$.
- $P$ is a function of $T$ and $R$ in $F_n$, so that for any pair of $< t, r >$, $P = f_R(t, r)$.

3. METHOD
To draw a representative sample of tags, data was collected from del.icio.us.com, which is currently the largest social bookmarking website. Information about the dataset is shown in Table 1. To verify our model, we have used a complex network model to represent a folksonomy. We collapsed all tuples (tag, URL, user) to calculate cosine similarities between tags for all URLs and all users and decomposed the similarity matrix using eigen decomposition to extract dimensionality. Because each extracted dimension represents a part of the underlying structure of the dataset, only a certain portion of tags will have a strong correlation to a given dimension. From the scree plot, the top five dimensions were extracted for each year, assuming that the extracted dimensions are the most representative tag structure for that year since most variability is accounted for by these dimensions.

4. RESULT
A longitudinal comparison reveals that, although tagging as a whole is scale-free, consistent patterns of aggregated tagging dimensions (i.e., folksonomy networks) can be found in a system folksonomy. We observed that the system folksonomy consists of multiple folksonomy networks containing possible controlled vocabularies that are usable for searching in specific domains. In Figure 1, all biplots were generated from the same system folksonomy but result in different dimensions. Each dot represents a tag, whose position is given by its loadings on two dimensions of the five extracted. Figure 1 indicates that there are clear dimensionalities among tags according to user assignment patterns on URLs. Some tags are tightly clustered, suggesting a specific topical domain. The differences between two topical domains are maximized because there is minimum dependence between them, given that each component is orthogonal to all other components. Thus each dimension represents a folksonomy network.

<table>
<thead>
<tr>
<th>Year</th>
<th>All tags (step 1)</th>
<th>Tags (&gt;5 unique users) (step 2)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Tag</td>
<td>URL</td>
</tr>
<tr>
<td>2005</td>
<td>8,372</td>
<td>2,817</td>
</tr>
<tr>
<td>2006</td>
<td>10,244</td>
<td>3,503</td>
</tr>
<tr>
<td>2007</td>
<td>12,079</td>
<td>6,007</td>
</tr>
</tbody>
</table>

Figure 1 suggests that folksonomy networks are consistent structures evolving across time. By looking at one column at a time, we can compare structural change across years. The 2005 structure is more diverse due to its early stage in the development of the domain vocabulary. As the structure develops through 2006 and 2007, tags cluster more closely around their dimensions. From this comparison, it is possible to track the evolving structures of domain vocabulary development.

To verify that these patterns are non-random, we further randomized all users, tags, and URLs and applied the same procedure to the randomized data. The five principal components of this random network are shown in Figure 2. As Figure 2 demonstrates, the dataset is quite unidimensional, suggesting that no sub-dimensions were detected.

5. Conclusion
These results indicate that, while a system folksonomy is unlikely to result in a coherent global classification system, tagging of similar resources is highly aggregative and can point to domain dependent vocabularies that are useful for retrieving domain resources. It also supports the contention that the three-level structure of folksonomies provides an effective lens for interpreting the seemingly chaotic role of folksonomies for information retrieval in the Web environment.
6. REFERENCES


