

# Longitudinal Analysis of Tag Structure in Del.icio.us

Lijiang Guo  
Indiana University Bloomington  
1320 E. 10th St., LI 011  
Bloomington, IN 47405  
lijguo@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.idio.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 of Web resources, where semantic relations between tags and resources are assumed to be worked out by users in a collective and negotiated process. Few studies exist that investigate how this process is actually accomplished, what the commonalities of tags are, and how seemingly sporadic tagging by individuals can become useful metadata for information retrieval.

This research builds upon current discussions of the mechanisms behind collective tagging behavior and their theoretical roots to propose a three-level structure for the aggregation of tags: the individual folksonomy that aggregates tags assigned by a single user; the folksonomy network that aggregates user generated tags within a topical domain; and the system folksonomy that aggregates all tags assigned within a single social bookmarking site. It is proposed that a folksonomy network constitutes a rudimentary indexing language in that the aggregation of tags assigned within a topical domain can serve as a precursor to a controlled domain vocabulary. This proposal is supported by the

results of a social network analysis of tags assigned to bookmarks in the Delicious social bookmarking system over a period of three continuous years (2005 to 2007). A longitudinal comparison of the results from exploratory factor analysis reveals that, although tagging as a whole is scale-free, consistent patterns of aggregated tagging behavior can be found in folksonomy networks.

## 2. FOLKSONOMY

A folksonomy consists of user-generated metadata about digital resources. The word folksonomy was first used by Thomas Vander Wal (Vander Wal, 2007) as a fusion of the terms folk and taxonomy to describe the set of tags assigned to resources in an online information system by a single user. Recent studies of folksonomies have focused on the social bookmarking systems (Hammond et al., 2005) that can be found in many different environments and exemplify a wide range of purposes, including blogging (WordPress), photo sharing (Flickr), video sharing (YouTube), and social networking (Facebook). The advantage of social bookmarking is that it is a bottom-up categorization structure that generates an emerging indexing language for resources on the Web (Vander Wal, 2007). Contrary to traditional classification methods, where specially trained indexers generate a standard indexing language for all users, a folksonomy is the result of empowering users with absolute control over their own information repository. In other words, a folksonomy is an indexing language generated for the user by the user.

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 Hotho, 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  $F$ ,
- $f_n$  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  $\langle t, r \rangle, P = f_n(t, r)$ .

### 3. METHOD

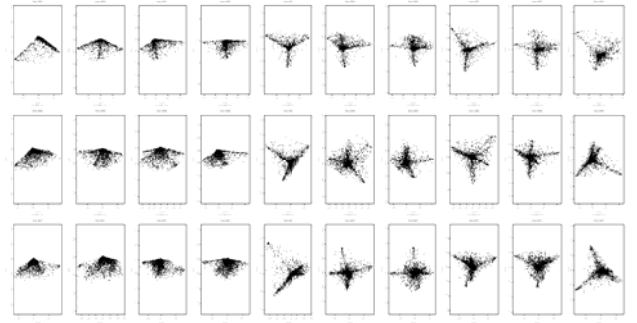
To draw a representative sample of tags, data was collected from del.icio.us.com, which is currently the largest social bookmarking website. Information about the dataset is shown in Table 1. To reify our model, we have used a complex network model to represent a folksonomy. We collapsed all tuples (tag, URL, user) to calculate cosine similarities between tags for all URLs and all users and decomposed the similarity matrix using eigen decomposition to extract dimensionalities. 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

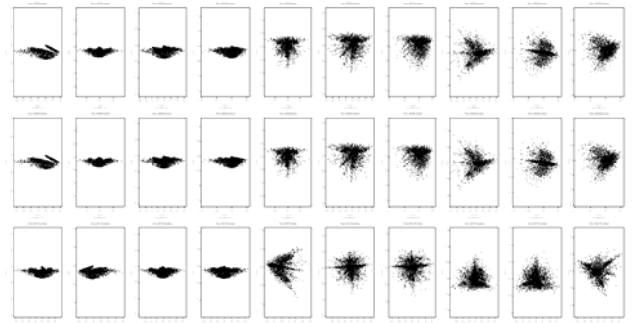
A longitudinal comparison reveals that, although tagging as a whole is scale-free, consistent patterns of aggregated tagging dimensions (i.e., folksonomy networks) can be found in a system folksonomy. We observed that the system folksonomy consists of multiple folksonomy networks containing possible 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 1. Data before and after noise reduction.**

Year	All tags (step 1)				Tags (>5 unique users) (step 2)			
	Tag	URL	User	Triple	Tag	URL	User	Triple
2005	7,372	2,817	32,794	296,998	1,586	2,802	32,267	284,281
2006	10,224	3,925	72,141	551,029	2,466	3,903	71,131	533,841
2007	12,079	6,007	104,696	759,203	3,195	5,985	103,181	737,547



**Figure 1. Biplot of component loadings (From top row 2005, 2006, and 2007; from left column D1xD2, D1xD3, D1xD4, D1xD5, D2xD3, D2xD4, D2xD5, D3xD4, D3xD5, and D4xD5, where “D” represents an extracted dimension)**



**Figure 2. Biplot of component loadings from randomized data (From top row 2005, 2006, and 2007; from left column D1xD3, D1xD4, D1xD5, D2xD3, D2xD4, D2xD5, D3xD4, D3xD5, and D4xD5 where “D” represents an extracted dimension)**

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 a 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

- [1] Hammond, T., Hannay, T., Lund, B., and Scott, J. 2005. Social bookmarking tools (I): A general review. *D-Lib Magazine*, 11, 4 (April, 2005). DOI=<http://www.dlib.org/dlib/april05/hammond/04hammond.html>
- [2] Hotho, A., Jäschke, R., Schmitz, C. and Stumme, G. 2006. In Information retrieval in folksonomies: Search and ranking. In *The semantic web: Research and application*. Springer, Berlin, Germany, 411-426. DOI=<http://www.springerlink.com/content/r8313654k80v7231/fulltext.pdf>
- [3] Mathes, A. 2004. Folksonomies: Cooperative classification and communication through shared metadata. DOI=<http://www.adammathes.com/academic/computer-mediated-communication/folksonomies.html>
- [4] Vander Wal, T. 2007. Folksonomy: Folksonomy coinage and definition. DOI=<http://vanderwal.net/folksonomy.html>
- [5] Wittgenstein, L. (1958). *Philosophical investigations*. Macmillan, New York, NY.

**Columns on Last Page Should Be Made As Close As Possible to Equal Length**