# Tagging patterns in a derived community of interests within a social bookmarking site

**Jung Sun Oh**
School of Information Sciences, University of Pittsburgh
135 N. Bellefield Ave., Pittsburgh, PA 15260
jsoh@pitt.edu

## INTRODUCTION

With the success and growth of social bookmarking/tagging sites, an unprecedented amount of user-generated metadata, in the form of tags, become available. While many discussions and propositions have been made on the potential utility of tags as a means for organization and/or retrieval of information, empirical research is still scarce to support the use of tag data for such an application.

This study is aimed to improve our understanding of the tagging phenomenon by adopting a new approach to collecting and analyzing data. We argue that both people's tagging behavior and the structure of tag data can be better understood within a clearly defined context that establishes the boundaries of data collection and analysis. The context used in this particular study is a community of interests derived from user activities within a social bookmarking site. Building upon a previous study (Oh, 2010), which identified communities of shared interests among active users of *delicious.com* using network analytic techniques, we collect and analyze tag data with reference to a specific community of shared interests.

## CONTEXT AND MOTIVATION

In the previous study, a large collection of bookmarking activities on *delicious.com* was analyzed to build a network of users based on their shared interests, and to identify communities within the network (see below for details). One of the motivations for identifying communities within a large social bookmarking site was to lay the groundwork for a further investigation of tagging behavior. If there exists a community consisting of users sharing a coherent set of interest, would the members of the community show different patterns of tagging? Would there be a difference in the set of tags being used and/or the emergent structure of tags (the pattern of connections among tags)?

Theories of categorization and previous empirical evidence suggest that background knowledge as well as individual experiences and expertise largely affect how people categorize objects (Heit, 1997; Chi et al., 1981; Medin et al., 1997). We can then speculate that, for instance, the categories that the members of a specialized community of practice would use to sort out information related to their practice may be different from the categories that others use. The question is whether we can apply the similar assumption to the social tagging environment.

We posit that people who care about a subject enough to build a large collection of bookmarks related to it would have a certain level of expertise in it. In addition, such users who show similar bookmarking patterns indicating a great deal of interests in the same subject may also adopt similar terms to categorize the information objects. In other words, while a community derived from bookmarking records does not constitute a real community of practice, it might as well share a relatively specialized vocabulary.

The main objective of the current study is to empirically examine the above proposition. In the following, a small pilot study conducted as a part of this study will be presented.

## METHODS

As described above, communities of shared interests within a network of active delicious.com users were identified in the previous study (Oh, 2010). More specifically, an affiliation network consisting of users and information objects that they had bookmarked was first constructed. The affiliation network was then transformed to create a network of users, such that two users who share a certain number of bookmarks would be connected. In order to discover communities of shared interests within the network, a technique called *m*-core (Scott, 2000) or *m*-slice (Nooy et al., 2005) was used.

An *m*-slice is a sub-network defined by the line multiplicity values. For a given *m* value, the *m*-slice consists of edges that have a value of *m* or higher and nodes that are incident on those edges. The basic procedure of *m*-slice analysis is similar to that of hierarchical clustering using a divisive method. Starting from the original network, edges and nodes are progressively removed as the value of *m* increases, and the original network is iteratively broken



*ASIST 2011,* October 9-13, 2011, New Orleans, LA, USA.

down into smaller sub-networks. It is, in effect, filtering out the weakest ties at each step so that areas with stronger connections are brought forth. Components that emerge at any point in this divisive clustering procedure can be regarded as subgroups or communities of varying cohesiveness. In other words, a component in an *m*-slice can be taken as a community within which nodes are connected by the minimum strength of *m*.

In the network of *delicious.com* users, each *m*-slice represents the sub-network where each connected pair of users has *m* or more bookmarks in common. In the initial network (the 1-slice), two users are connected if they have one or more shared bookmarks. When *m* increases to 2, the links between users who share only one bookmark will be removed and, therefore, each of the remaining pairs in the network has at least two shared bookmarks.

After 27 iterations, three communities, each of which representing coherent theme of interests, were emerged on the 28th slice. Note that the theme of the community was analyzed using the information objects that constituted the links (i.e., shared bookmarks) among the members of the community. The majority of the information objects that the members of the first and largest community had in common were related to the topic of web development and/or design, whereas the themes of the second and the third community were fan fiction and recipes, respectively.

In this pilot study, we take the second community for the analysis. The community is small in size but has a distinctive theme, making it amenable for an initial exploration. The information objects shared by its members are pieces of fiction, usually posted on a blog, by online amateur writers. The majority of them fall into the category of fan fiction. Fan fiction is creative writing "where fans create stories using characters, settings, and events from their favorite books, movies, or television shows." (Burns & Webber, 2009, p.27).

In order to compare the tagging patterns of community members to those of non-members, three steps were taken to extract tags from relevant bookmarking records. First, each link in the chosen community was traced back to the information objects. In the fan-fiction community, there were 1,306 links among the 248 members. When each link was traced back to the URLs that had contributed to make the link, we obtain 6,414 URLs with 55,635 occurrences. Note that since the communities at hand identified on the 28th slice, each and every pair of users had at least 28 common bookmarks. In other words, each link had at least 28 URLs.

Among those 6,414 URLs, 47 URLs were selected for further analysis, using the following two criteria:

- The item should be bookmarked by 30 or more users
- The proportion of the records created by the community members is between 20% and 80%

For each of the 47 URLs, all the instances of bookmark posting over the three-month period of February 2008 – April 2008 (the period used to build the original network) were examined to extract the tags assigned by the users. For each instance, whether the associated user is a member of the community was marked.

In total, there were 2,270 bookmark records associated with the 47 URLs, including 1,434 records by the members of the community and 836 by non-members.

The last step involved extracting all the tags from each bookmarking records. The resulting set of tags includes 2564 unique tags with 13,765 occurrences.

## PRELIMINARY FINDINGS

The analysis is still in progress. The first part of the analysis was to examine the distribution of tag frequency in the two groups: members and non-members. On average, the members assigned 6.33 tags per URL (the range is 0 to 34) while non-members had 5.61 tags per URL (the range is 0 to 27). In both groups, the tag frequency shows a clear signature of a long-tail distribution, with more than 50% of the cases (tags) having the frequency of 1. While the percentage of tags with a single occurrence is slightly higher in the non-member group, the shapes of distribution in the two groups are comparable. The lists of the top 20 most frequent tags from the two groups contain mostly the same terms while the ranks were different. By and large, there are no noticeable differences when simple frequencies were considered.

A more interesting observation was made when the proportion of instances per URL by members and those by non-members were compared. We compiled the list of all the tags with 20 or more instances and calculated, for each tag, the proportion of its assignment made by the members of the community and by non-members. The top tags that show proportionally higher occurrences in the member group are mostly 'compound tags,' tags consisting of two or more words concatenated by a special character (e.g. genre.drama, pairing:john/rodney), while the tags used more often by non-members are all single-term tags with a few exceptions. This suggests that the level of specificity of tagging is greater within the community.

We will continue the analysis of the tagging patterns in these two groups using both qualitative and quantitative methods.

## CONCLUSION

This paper presented a pilot study of a larger project that is in its early stage. The basic premise of the project is that people's tagging behavior and the utility of tags should be examined within a defined context, such as the context of shared interests in this study. Currently, tag data from a popular social bookmarking site are being examined along with a derived community of shared interests. While interesting observations were made, much remains to be

analyzed. The result of this study will have a practical implication for developing a better way to harness tag data as well as a theoretical implication for understanding people's tagging behavior.

## REFERENCES


Burns, E., & Webber, C. (2009). When Harry met Bella. School Library Journal, 55(8), 26-29.

Chi, M. T., Feltovich, P. J., & Glaser, R. (1981). Categorization and representation of physics problems by experts and novices. Cognitive Science, 5, 121-152.

Heit, E. (1997). Knowledge and concept learning. In K. Lamberts & D. Shanks (Eds.), Knowledge, concepts, and categories (pp. 7-41). London: Psychology Press.

Medin, D. L., Lynch, E. B., Coley, J. D., & Atran, S. (1997). Categorization and reasoning among tree experts: Do all roads lead to Rome? Cognitive Psychology, 32(1), 49-96.

Nooy, W. d., Mrvar, A., & Batagelj, V. (2005). *Exploratory Social Network Analysis with Pajek*. Cambridge ; New York: Cambridge University Press.

Oh, J. S. (2010). Network analysis of shared interests represented by social bookmarking behaviors. Unpublished Dissertation, University of North Carolina, Chapel Hill.

Scott, J. (2000). *Social network analysis : a handbook*. London ; Thousands Oaks, Calif.: SAGE Publications.


**The columns on the last page should be of approximately equal length.**