RC24407 (W0711-064) November 9, 2007
Computer Science

IBM Research Report

Anomalous Tagging Patterns Can Show Communities among Users

Michael Muller
IBM Research Division
One Rogers Street
Cambridge, MA 02142 USA

Research Division
Almaden - Austin - Beijing - Cambridge - Haifa - India - T. J. Watson - Tokyo - Zurich

Anomalous Tagging Patterns can show Communities among Users

Michael Muller
IBM Research
michael_muller@us.ibm.com

Abstract. This poster explores certain anomalous patterns that appear to be common in social tagging data, in which seemingly similar tags are used quite differently. We analyzed data from an enterprise social-tagging service, exploring the resources, tags, and users associated with tags that show these peculiar patterns of usage. We show that the anomalous patterns of these seemingly similar tags may be evidence of communities among users, with clear implications for design choices in tagging services.

Kipp and Campbell (2006) reported an anomalous pattern among social-tagging data from an internet service. Using multidimensional scaling (MDS) on tag co-occurrence data, they showed that similar tags did not always occupy adjacent positions in the two-dimensional MDS outcomes. Examples included the tags "tv" and "television", and a pair of tags that differed by a single punctuation character, "socialbookmarking" and "social_bookmarking". It is well-known that users tend to use a diversity of terms for the same concept (Furnas, 1987), and that at least a subset of these issues occurs in social tagging data (Golder and Huberman, 2006; Muller, 2007). Some have argued that only the author of a tag is likely to be able to make sense of that tag (for review, see Hammond et al., 2005). These anomalies would reduce the value of tagged data as a shared knowledge resource.

We analyzed data from Dogear, an enterprise social-tagging service used by thousands of IBM employees to share bookmarks on internet and intranet webpages and documents (Millen et al., 2006). MDS analyses revealed many sets of tags with an anomalous pattern similar to that of Kipp and Campbell. We focused our analyses on three pairs of tags that differed only in capitalization: Set1: “Java”/“java” (264 users); Set2: “RSS”/“rss” (183 users); and Set3: “Software”/“software” (250 users). We also examined a set of four tags with the same root word: Set4: “blog”/“blogs”/“blogging”/“Blogs” (342 users).

Within each Set, we examined bookmarks that contained these tags. As anticipated, there was a strong tendency for different resources to be associated with each tag-variant. There was a strong tendency for different subsets of users to be associated with each of the tags in these four sets. The mean pairwise overlaps of users were 37% for Sets1, 2, and 3, and 46% for Set 4. There were similar low overlaps among the resources described by the tags, and among the additional tags that co-occurred with the tag-variants under study. Thus, in all four Sets, each of the tag-variants was used differently, by different users, to tag different resources, in relation to different co-occurring tags. The anomalous data associated with apparently similar tags appears to be evidence of unsuspected communities of practice or other emergent user networks.

What kinds of communities? Because of the large number of users in the sample, we looked for formal, online data that could explain the distinctive choices of different tag variants. We looked at all the "official" data fields in the users' online directory entries, and we also obtained exhaustive lists of the self-subscriptions to IBM internal communities (827 communities with a total of 108,181 unique members, about a third of IBM's employees). Chi-square and discriminant analysis have so far not shown a relationship between formally-recorded employee attributes and the emergent communities that we found in these data. We continue to search for such distinctive, explanatory attributes.

How should tags be managed with regard to these user communities? Our results highlight the importance of tags as community-distinctive vocabularies. Contemporary tagging technology may affect the ability to form community tagging vocabularies. As is well-known, communities develop their own distinctive resources, including shared languages (e.g., Lave & Wenger, 1991). Sen et al. (2006) showed that tag-suggestions (i.e., type-ahead auto-completions when users are entering tags) can have powerful influences on a group's developing tag vocabulary. Thus, providers of tagging services have a choice: (a) Auto-suggest a single, common tagging vocabulary, which will enhance overall consistency but weaken the formation of community-specific vocabularies; vs. (b) Avoid auto-suggestion, thus enhancing the ability of communities to specialize their own vocabularies, but weakening overall consistency across all users. Our results can help organizations to make informed choices among these courses of action.

References

Furnas, G.W., Landauer, T.K., Gomez, L.M., & Dumais, S. T. (1987), 'The vocabulary problem in human-system communication,' Comm. ACM 30 (11), 964-971.

Golder, S.A., & Huberman, B.A. (2006): 'Structure of collaborative tagging systems,' J. Info. Sci. 32(2), April, 2006.

Hammond, T., Hannay, T., Lund, B., & Scott, J. (2005): 'Social bookmarking tools (I): A general review,' D-Lib Magazine 11(4), April 2005, http:// www.dlib.org/dlib/ april05/hammond/04hammond.html (verified 17 May 2007).

Kipp, M.E.I. & Campbell, D.G. (2006): 'Patterns and Inconsistencies in Collaborative Tagging Systems : An Examination of Tagging Practices,' Proc. Am. Soc. for Info. Sci. & Tech., Austin, Texas (US), 2006.

Lave, J., & Wenger, E. (1991): *Situated Learning: Legitimate Peripheral Participation*, Cambridge: Cambridge University Press.

Millen, D.R., Feinberg, J., & Kerr, B. (2006): 'Dogear: Social bookmarking in the enterprise,' Proc CHI 2006.

Muller, M.J. (2007): 'Patterns of tag usage across four diverse enterprise tagging services,' Paper at HCIC 2007, Winter Park, CO, USA, February 2007.

Sen, S., Lam, S.K., Rashid, A.M., Cosley, D., Frankowski, D., Osterhouse, J., Harper, F.M., & Riedl, J. (2006): 'Tagging, communities, vocabulary, evolution,' Proc CSCW 2006.