RC24409 (W0711-067) November 9, 2007
Computer Science

IBM Research Report

Divergent Vocabularies in Four Enterprise Tagging Services

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

Divergent Vocabularies in Four Enterprise Social-Tagging Services

Michael Muller
IBM Research
michael_muller@us.ibm.com

Abstract. We examined the tagging vocabularies that developed in four enterprise social-tagging services. Despite the fact that many of the same users were involved with the different services, and despite the fact that those users were doing real work with those services, there was surprisingly small overlaps (or re-use) of tagging vocabularies across the services. We discuss strategies to improve the consistency of social-tagging vocabularies across services that can be applied during tag-entry, tag-storage, and tag-based search.

Social-tagging services have become popular on the internet (Hammond et al., 2005). Several projects have explored how these services can work inside enterprises (Damianos, 2006; Farrell et al., 2007; John, 2006; Millen et al., 2006; Muller et al., 2007). The formal differences are that enterprise tagging services permit full authentication of each user, and can support social-tagging of both public and company-confidential resources. Other phenomena emerge through use. This poster provides a first public examination of one of those emergent phenomena.

We examined the consistency in the use of tags across four tag-based services that are used by hundreds of employees within IBM as part of their daily jobs.[1] Dogear is an internal service that is similar to Delicious (del.icio.us) (Millen et al., 2006). Bluepages+1 is an enhanced directory service that allows one user to write tags directly onto the directory description of another user (Farrell et al., 2007). BlogCentral is the corporate internal blogging site, which allows blog authors to write tags onto their blogs and their individual blog postings. Activities is a product prototype that allows users to share diverse media in structured collections, which can be tagged (Moore et al., 2006). Over two years, 4987 IBM internal users have contributed over 120,000 bookmarks and similar references to these four systems, involving 28460 unique tags across those services.

We reduced the bookmarks from each service to lists of unique tags within each service, and then tested for consistency using the Overlap Coefficient[2] (i.e., taking the ratio of the intersection of unique tags between services, divided by the number of unique tags in the smaller of the two services). The average overlap of

---

[1] Previous research has examined tagging vocabularies *within individual services* (Golder & Huberman, 2006; Sen et al., 2006). We think this is the first study of tagging vocabularies *across services*.

[2] For review and comparison with other measures of corpus similarity, see Chapman (n.d.)

tags was only 36%, despite the fact that these several thousand users were doing real work with these systems. Surprised by this low figure, we examined each person's individual tagging vocabulary, and found an average overlap rate of only 3% of tags across services per user. Other analyses, based on methods for comparing corpora (Kilgarriff, 1997) corroborated the initial findings of low vocabulary overlap across services.

Because of these inconsistencies in tagging vocabularies, tag-based searches across services may fail to return many relevant bookmarks and resources – a new instance of the "vocabulary problem" initially described by Furnas et al. (1987). Conventional systems attempt to remedy this kind of problem, on a within-system basis, at the point of tag-*entry*. However, there is evidence that seemingly trivial variations in tag spelling, capitalization, and punctuation may be highly significant to users (Muller, 2007), and normalization during tag-entry would erase these distinctions Therefore, this poster will explore alternative points of intervention, namely at tag-*storage* and at tag-based-*search*. Trade-offs and relative advantages will be presented.

References

Chapman, S. (n.d.): 'Sam's String Metrics," http://www.dcs.shef.ac.uk/~sam/stringmetrics.html (verified 17 May 2007).

Damianos, L., Griffith, J., & Cuomo, D. (2006): 'Onomi: Social Bookmarking on a Corporate Intranet,' Position paper in WWW 2006 Tagging Workshop.

Farrell, S., Lau, T., Wilcox, E., & Muller, M.J. (2007): 'Socially Augmenting Employee Profiles with People-Tagging,' to appear in *Proc UIST 2007*, Newport, RI, USA, October 2007.

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.

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).

John, A., & Seligmann, D. (2006): 'Collaborative tagging and expertise in the enterprise,' in *Proc WWW 2006*.

Kilgarriff, A. (1997): 'Using word frequency lists to measure corpus homogeneity and similarity between corpora,' Proc 5th ACL SIGDAT Workshop on Very Large Corpora, 231-245, Beijing and Hong Kong, 1997.

Muller, M.J. (2007): 'Anomalous Tagging Patterns can show Communities among Users,' poster at ECSCW 2007, Limerick, Ireland, September 2007.

Millen, D.R., Feinberg, J., & Kerr, B. (2006): 'Dogear: Social bookmarking in the enterprise', *Proc CHI 2006*, Montreal, ACM, pp. 111-120.

Muller, M.J., Geyer, W., Brownholtz, B., Dugan, C., Millen, D.R., &Wilcox, E. (2007): 'Tag-Based Metonymic Search in an Activity-Centric Aggregation Service,' *Proc ECSCW 2007*, Limerick, Ireland, September 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*.