Tagging *tagging*.
A classification model for user keywords in scientific bibliography management systems

Markus Heckner (markus.heckner@paedagogik.uni-regensburg.de)
Susanne Mühlbacher (susanne1.muehlbacher@sprachlit.uni-regensburg.de)
Christian Wolff (christian.wolff@sprachlit.uni-regensburg.de)
University of Regensburg

1. Research Context and previous work

Recently, a growing amount of systems that allow personal content annotation (tagging) are being created, ranging from personal sites for organising bookmarks (*del.icio.us*), photos (*flickr.com*) or videos (*video.google.com, youtube.com*) to systems for managing bibliographies for scientific research projects (*citeulike.org*, *connotea.org*). Simultaneously, a debate on the pro and cons of allowing users to add personal keywords to digital content has arisen.

One recurrent point-of-discussion is whether tagging can solve the well-known vocabulary problem: In order to support successful retrieval in complex environments, it is necessary to index an object with a variety of aliases (cf. Furnas 1987). In this spirit, social tagging enhances the pool of rigid, traditional keywording by adding user-created retrieval vocabularies. Furthermore, tagging goes beyond simple personal content-based keywords by providing meta-keywords like *funny* or *interesting* that "identify qualities or characteristics" (Golder and Huberman 2006, Kipp and Campbell 2006, Kipp 2007, Feinberg 2006, Kroski 2005). Contrarily, tagging systems are claimed to lead to semantic difficulties that may hinder the precision and recall of tagging systems (e.g. the polysemy problem, cf. Marlow 2006, Lakoff 2005, Golder and Huberman 2006).

Empirical research on social tagging is still rare and mostly from a computer linguistics or librarian point-of-view (Voß 2007) which focus either on the automatic statistical analyses of large data sets, or intellectually inspect single cases of tag usage: Some scientists studied the evolution of tag vocabularies and tag distribution in specific systems (Golder and Huberman 2006, Hammond 2005). Others concentrate on tagging behaviour and tagger characteristics in collaborative systems. (Hammond 2005, Kipp and Campbell 2007, Feinberg 2006, Sen 2006). However, little research has been conducted on the functional and linguistic characteristics of tags.[1] An analysis of these patterns could show differences between user wording and conventional keywording. In order to provide a reasonable basis for comparison, a classification system for existing tags is needed.

Therefore our main research questions are as follows:

- Is it possible to discover regular patterns in tag usage and to establish a stable category model?
- Does a specific tagging language comparable to internet slang or chatspeak evolve?
- How do social tags differ from traditional (author / expert) keywords?
- To what degree are social tags taken from or findable in the full text of the tagged resource?
- Do tags in a research literature context go beyond simple content description (e.g. tags indicating time or task-related information, cf. Kipp et al. 2006)?

2. Goals and methodology – The Tag Category Model (TCM)

Our study was conducted in two steps using data from *connotea.org* :

(Step 1) *Explorative creation of a classification model.* A set of 2.000 bookmarks was extracted by random selection from *connotea*. The tags of these samples were distributed among information scientists with the instruction to derive classes from recurring patterns in tag usage. Finally, a category model was established in the course of an expert workshop.

(Step 2) *Explanatory case study: Applying and verifying the classification model.* A data set of 500 IT-related scientific articles was extracted by random selection from connotea. The expert group was instructed to assign the related 1300 user tags as well as the conventional keywords to the classification model in order to define and compare functional and linguistic characteristics.

[1] Some findings are presented in Kipp and Campbell 2006, Kipp 2007, Golder and Hubermann 2006.

3. Results

A stable category system including functional and linguistic characteristics of personal tags and expert and author keywords could be established. Author and expert keywords showed significant differences compared to personal tags.

In contrast to Kipp and Campbell (2006) who studied tagging strategies in *del.icio.us* bookmarks only few “time and task related tags” like “toread” and “cool” could be found in our dataset. This can possibly be ascribed to fundamental differences between scientific and standard language use (cf. Wüster 1991). First indicators for a language register specific to the tagging situation (compare e.g. Yew et al. 2006) are noted and are subject to review. User strategies for tag evasion could also be discovered. The tag category model as well as a detailed overview of our findings will be presented at the workshop.

References:

Courtial, J.P., Callon, M., & Sigogneau, M. (1984). Is indexing trustworthy? Classification of articles through co-word analysis. Journal of information science 9: 47-56.

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

Golder, S. & Huberman, B. A. (2006), 'The Structure of Collaborative Tagging Systems', Journal of Information Science 32, 198-208.

Hammond, T., Hannay, T., Lund, B. and Scott, J. Social Bookmarking Tools – A General Overview. D-Lib Magazine11, 4 (April 2005)

Kipp, Margaret E. I. and Campbell, D. Grant (2006) Patterns and Inconsistencies in Collaborative Tagging Systems : An Examination of Tagging Practices. In Proceedings Annual General Meeting of the American Society for Information Science and Technology, Austin, Texas (US).

Kipp, Margaret E. I. (2007). @toread and Cool: Tagging for Time, Task and Emotion. In: Proc. Information Architecture Summit 2007. Las Vegas. [Online: http://eprints.rclis.org/archive/00010445/]

Kroski, E. (2005). The hive mind: Folksonomies and user-based tagging. [Online: http://infotangle.blogsome.com/2005/12/07/the-hive-mind-folksonomies-anduser-based-tagging/]

Marlow, C.; Naaman, M.; Boyd, D. & Davis, M. (2006),HT06, tagging paper, taxonomy, Flickr, academic article, to read, in 'HYPERTEXT '06: Proceedings of the seventeenth conference on Hypertext and hypermedia', ACM Press, New York, NY, USA, pp. 31--40.

Salton, G. & McGill, M.J. (1983). Introduction to modern information retrieval. New York: McGraw-Hill.

Shirky, C. (2005b). Folksonomies + controlled vocabularies. [Online: http://many.corante.com/archives/2005/01/07/folksonomies_controlled_vocabularies.php]

Shirky, C. (2005). Ontology is overrated: categories, links and tags. [Online: http://www.shirky.com/writings/ontology_overrated.html]

Voß, J. (2007). Tagging, Folksonomy & Co - Renaissance of Manual Indexing? In: Osswald, A.; Stempfhuber, M.; Wolff, C. (Eds.): Open Innovation. Proc. 10th International Symposium for Information Science. Constance: UVK, 243-254.

Wüster, E. (1991). Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.Bonn: Romanistischer Verlag.

Yew, J., Faison, G., Teasley, S. (2007). Learning by tagging: group knowledge formation in a self-organizing learning community. ICLS '06: Proceedings of the 7th international conference on Learning sciences.