# Information-Theoretic Models of Tagging

Harry Halpin

Institute for Communicating and Collaborative Systems, School of Informatics, University of Edinburgh, 2 Buccleuch Place, Edinburgh, EH8 9LW, Scotland, UK

**Abstract.** In earlier work, we showed using Kulback-Leibler (KL) divergence that tags form a power law distribution very quickly. Yet there is one major observed deviation from the ideal power law distribution for the top 25 tags, a large "bump" in increased frequency for the top 7-10 tags. We originally hypothesized that the "bump" in the data could be caused by a preferential attachment mechanism. However, an experiment that tested both feedback and no-feedback conditions over tagging (200+ subjects) shows that the power law distribution arises regardless of any feedback effect. We hypothesize that an information-theoretic analysis of tags lead to a power law without feedback.

In Halpin, Robu, and Shepard, we showed using Kulback-Leibler (KL) divergence that tags form a power law distribution very quickly. This can be demonstrated by taking the KL divergence between every two consecutive points in time of the distribution, so stabilization occurs when the KL divergence goes to zero. An alternative method is to take the KL divergence with regards to an "ideal" power law, checking with each iteration if the KL divergence decreases. We demonstrated both these measures converge quickly using 500 randomly selected tags from del.icio.us both from "popular" (heavily tagged) and "recent" (randomly chosen) tags by inspecting their tagging histories [2].

Yet there is one major observed deviation from the ideal power law distribution for the top 25 tags, a large "bump" in increased frequency for the top 7-10 tags. We originally hypothesized that the "bump" in the data could be caused by a preferential attachment mechanism. However, in a recent experiment that tested both feedback and no-feedback conditions over tagging (200+ subjects) shows that the power law distribution arises regardless of any feedback effect [1]. There is some increased variance in tags without feedback and reinforced tags move up the power law distribution quicker with feedback. Yet, even without feedback the fundamental power law distribution arises.

Can an information-theoretic analysis of tags lead to a power law without feedback? In a classical information retrieval paradigm, each group of tags would have an entropy assigned to it depending on what URIs they retrieve. For example, a tag applied to every relevant resource would retrieve every document, and so have an entropy of 0 while a tag that selects a single relevant resource would have an entropy of 1. Obviously, tags should not retrieve a single resource, but some small constant number, such as seven. If users selected an non-ideal-yet-approximate and useful encoding with every choice, a power law could result.

Dagstuhl Seminar Proceedings 08391
Social Web Communities
http://drops.dagstuhl.de/opus/volltexte/2008/1787

## References

1. Dirk Bollen and Harry Halpin. The role of feedback in folksonomies: Is social tagging really social?, 2008. Under preparation.
2. Harry Halpin, Valentin Robu, and Hana Shepherd. The complex dynamics of collaborative tagging. In *Proc. of the 16th International World Wide Web Conference (WWW'07)*, pages 211–220. ACM Press, 2007.