The Recurrence Dynamics of Social Tagging

Dell Zhang
SCSIS
Birkbeck, University of London
Malet Street
London WC1E 7HX, UK
dell.z@ieee.org

Robert Mao
Microsoft Corp.
EPDC5/2352
South County Business Park
Leopardstown, Dublin, Ireland
robmao@microsoft.com

Wei Li
Google Inc.
1600 Amphitheater Parkway
Mountain View
CA 94043, USA
alexweili@gmail.com

ABSTRACT

How often do tags recur? How hard is predicting tag recurrence? What tags are likely to recur? We try to answer these questions by analysing the RSDC08 dataset, in both individual and collective settings. Our findings provide useful insights for the development of tag suggestion techniques etc.

Categories and Subject Descriptors

H.2.8 [**Database Management**]: Database Applications—*data mining*; H.3.1 [**Information Storage and Retrieval**]: Content Analysis and Indexing

General Terms

Measurement

Keywords

Web 2.0, Social Tagging, Folksonomy.

1. INTRODUCTION

Social tagging (aka folksonomy) is popular in Web 2.0 applications, such as Delicious, CiteULike, Last.fm, Flickr and YouTube.

In this paper, we study the recurrence dynamics of social tagging, i.e., how tags recur in folksonomies. Specifically, we try to answer the following questions. How often do tags recur? How hard is predicting tag recurrence? What tags are likely to recur?

These questions are crucial to understanding and improving social tagging. Although there exists some investigation on the dynamics of tag co-occurrence [1], the dynamics of tag recurrence is not well-studied.

2. DATA

We model the social tagging data as a time-ordered series of posts. Each *post* is described by a tuple (*user*, *item*, *tag-set*, *time*) which means the *user* annotated the *item* with the tags in *tag-set* at the *time*. Here we assume the system time is represented as an integer which is incremented by one on each post.

WWW 2009, April 20–24, 2009, Madrid, Spain.
ACM 978-1-60558-487-4/09/04.

We use the RSDC08 dataset[1] from the ECML/PKDD-2008 Discovery Challenge. It is provided by Bibsonomy[2], a Web 2.0 service for bookmarking and sharing bibliographic references. The dataset, after cleaning, consists of 152,171 posts from 2,475 users.

3. ANALYSIS

We consider two settings of tag recurrence: (i) **individual**: a tag in the given post from user u is regarded as a recurrence if it has been used by u herself before; (ii) **collective**: a tag in the given post from user u is regarded as a recurrence if it has been used by any user (either u herself or another user) before.

3.1 How often do tags recur?

Figure 1 shows the growth of tag recurrence number. In both individual and collective settings, the tag recurrence number increases at a fairly *steady* rate. On average, among the 3.7 tags used per post, 2.8 of them are individual recurrences and 3.4 of them are collective recurrences. That is to say, about 75% of tags have been used before individually and about 90% of tags have been used before collectively.

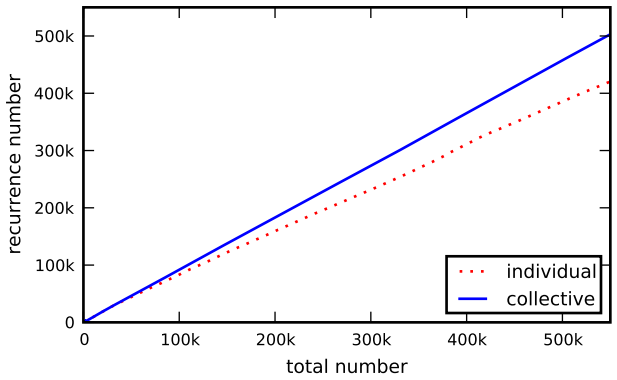

Figure 1: The rate of tag recurrence.

3.2 How hard is predicting tag recurrence?

We use the *entropy* of tags in the tagging history to characterize the uncertainty about tag recurrence: the larger the entropy, the harder it is to predict the next tag to be

[1]http://www.kde.cs.uni-kassel.de/ws/rsdc08/
[2]http://www.bibsonomy.org/

reused. In the collective setting, we calculate the tag entropy as $H(tag) = -\sum_{tag} p(tag) \log p(tag)$. In the individual setting, we calculate the conditional tag entropy given the user as $H(tag|user) = H(tag, user) - H(user)$.

Figure 2 shows the entropy of tags over time. Although the tag vocabulary grows with a nearly constant speed (0.3 new tags per post), the collective entropy increases very slowly to 12 bits, while the individual entropy remains almost stable at less than 8 bits. The collective entropy of 12 bits corresponds to a perplexity of 4,096 tags, far less than the number of (distinctive) tags. The individual entropy of 8 bits corresponds to a perplexity of 256 tags, which implies that *personalisation* has the potential to reduce the tag search space by more than 16 times.

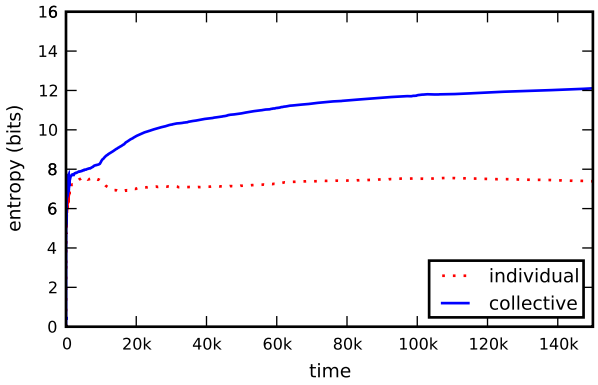

Figure 2: The uncertainty of tag recurrence.

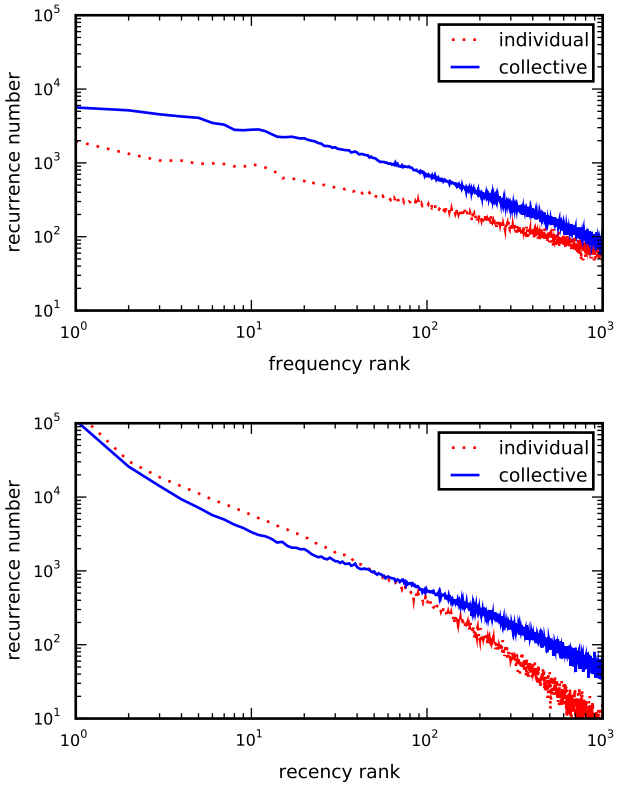

Figure 3: The likelihood of tag recurrence.

3.3 What tags are likely to recur?

Figure 3 shows two *log-log* plots of tag recurrence number versus tag frequency-rank and tag recency-rank respectively. We see that (1) more *frequently* used tags are more likely to recur; and (2) more *recently* used tags are more likely to recur. These observations are consistent with our intuition. Moreover, the nearly-straight lines in the above log-log plots suggest that the relationship between tag recurrence number n and frequency or recency rank r roughly follows the *power law*: $n \propto 1/r^{\alpha}$, where the scaling-exponent α is around 0.5 on tag frequency and around 1.0 on tag recency.

4. CONCLUSIONS

Our findings provide useful insights about the development of *tag suggestion* [2, 3] techniques etc.

Making tag suggestions based on the tagging history is feasible, because most of the tags to be used have actually been used before. Since tags recur more in the collective tagging history than in the individual tagging history, collective tag suggestion should be able to achieve better *recall* than individual tag suggestion.

Making tag suggestions based on the tagging history can continue to be effective when more and more tags are added into the system, because the entropy of tags will stay relatively small. Since the entropy of tags is higher in the collective tagging history than in the individual tagging history, individual tag suggestion should be able to achieve better *precision* than collective tag suggestion.

More frequently or recently used tags should be favoured for tag suggestion, because they have higher probabilities to be reused. Since the power law distribution scaling-exponents on tag recency are larger than those on tag frequency, tag recency is probably more useful than tag frequency for tag suggestion on our dataset. The strong correlation between tag recurrence and tag recency also suggests that a simple but up-to-date predictive model will probably work better than a comprehensive but out-of-date model for tag suggestion.

5. ACKNOWLEDGEMENTS

We would like to thank Prof. Wee Sun Lee (NUS), Dr. Yee Whye Teh (UCL) and Dr. Jun Wang (UCL) for stimulating discussions.

6. REFERENCES

[1] H. Halpin, V. Robu, and H. Shepherd. The complex dynamics of collaborative tagging. In *Proceedings of the 16th International Conference on World Wide Web (WWW)*, pages 211–220, Banff, Alberta, Canada, 2007.

[2] B. Sigurbjornsson and R. van Zwol. Flickr tag recommendation based on collective knowledge. In *Proceedings of the 17th International World Wide Web Conference (WWW)*, pages 327–336, Beijing, China, 2008.

[3] Y. Song, Z. Zhuang, H. Li, Q. Zhao, J. Li, W.-C. Lee, and C. L. Giles. Real-time automatic tag recommendation. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 515–522, Singapore, 2008.