Personalized Tag Recommendations via Tagging and Content-based Similarity Metrics

Andrew Byde
HP Labs
Filton Road, Stoke Gifford
Bristol, UK
andrew.byde@hp.com

Hui Wan
State University New York
at Stony Brook
Stony Brook, New York 11790
hwan@cs.sunysb.edu

Steve Cayzer
HP Labs
Filton Road, Stoke Gifford
Bristol, UK
steve.cayzer@hp.com

Abstract

This short paper describes a novel technique for generating personalized tag recommendations for users of social bookmarking sites such as del.icio.us. Existing techniques recommend tags on the basis of their popularity among the group of all users; on the basis of recent use; or on the basis of simple heuristics to extract keywords from the url being tagged. Our method is designed to complement these approaches, and is based on recommending tags from urls that are similar to the one in question, according to two distinct similarity metrics, whose principal utility covers complementary cases.

Keywords

Tagging, Bookmarking, Classification

1. Introduction

This paper addresses tag recommendation in social bookmarking sites. We address two problems, namely paucity of information for tag recommendation in the case of too few other users having tagged a url; and personalization of tag recommendations.

The first of these issues is especially important in the emerging field of enterprise-scale bookmarking and social networking sites: the manifest knowledge management value that such sites provide has not gone un-noticed in the world of business, but a key problem there is the lack of scale. While on the web, most pages I might choose to tag will already have been tagged by someone – whose recommendation can assist my choice and aid term-convergence – this is not the case within an enterprise, where there are not enough users for the system to rely on recommendations from peers. The second issue operates at any scale, and comes down to the observation that the laudable bias towards term-convergence provided by using other users' existing tags as recommendations discourages the easy development of personalized semantics.

In this paper we will develop tag recommendations based on two different page similarity metrics (a "tagging"- and "content"-based method, see Section 2). Our method recommends terms that the user has already used, selected according to analysis of the url in question. We envisage the terms recommended by these methods being presented in parallel with the "common" (frequent) terms used by other users, and claim the following benefits of using them:

1. The content-based method is capable of recommending terms even for urls that have not been previously tagged by anyone. This is of clear benefit for document collections that are at present sparsely tagged – such as is typical within the enterprise.
2. For a url with a very large number of tags, a users' preferred tags will likely be diluted by other tags relevant to different areas of interest, different specific vocabularies, and different languages or character sets than the user is interested in. Our method recommends from within the users' own field of interest, and thus is more pertinent and useful.

2. Method

The problem we face is, given a collection of urls u, tagged with a set of tags $T_p(u)$ by users p, to provide a particular user with a list of N recommended terms for a particular url. We will evaluate such a recommendation method with respect to del.icio.us data scraped from the web. For each of a collection of users, we take each url that they have tagged in turn. We fetch the "common" tags for that url, and crop the list to the top N.

Our method for the task of recommending tags to user p for url u is as follows: For each url u' that the user has already tagged, we calculate a similarity $sim(u, u') \in [0, 1]$ to the task url, to be described shortly. Given these similarities for each url, we summed similarities to give a user-specific weight for each tag:

$$w_{p,u}(t) = \sum_{i:t \in T_p(u_i)} sim(u, u_i). \quad (1)$$

The weight depends on the user in that we only sum similarities to u over other urls u_i that the user has tagged. The weights on each tag provided a ranking of tags, and we selected the top N ranked tags as the recommendation.

Note that this method scales with the number of urls tagged by the user, not the total number of all urls tagged by every user. As we shall see, the median number of urls tagged is 150, and only a tiny minority tag more than 1000, meaning that our method is scalable to cases of practical interest.

2.1 Similarity Metrics

Our similarity metrics are both variants on the cosine similarity familiar from text mining and information retrieval [1]:

$$sim(u, u') = \frac{u \cdot u'}{\sqrt{(u \cdot u)(u' \cdot u')}}, \quad (2)$$

Method	coverage	overall better	some gained	some lost	overall worse	uncommon recommended
Tagging	63.1% ± 6.9%	18.5% ± 5.6%	23.6% ± 5.8%	23.8% ± 6.1%	19.4% ± 5.1%	47.5% ± 7.2%
Content	93.1% ± 1.3%	36.1% ± 6.4%	41.8% ± 6.4%	30.7% ± 6.6%	25.8% ± 5.9%	46.4% ± 7.2%

Table 1: *Average across users of various performance metrics: the proportion of urls for which the recommended tags had, in comparison to the common tags, a larger total number of true tags; some additional true tags; some true tags lost; and a smaller total number of tags. "Uncommon recommended" refers to the average proportion of uncommon true tags that were recommended. Confidence intervals are 95%.*

where a url is represented as a vector, $u = (v_1, v_2, \ldots, v_M)$, and we define $u \cdot u' = \sum_{i=1}^{M} v_i\, v'_i$. All that remains is to specify how a particular url is given a vector space representation of the form required.

- **Tagging-based Similarity**. For this metric we take $(v_1, v_2, \ldots, v_M)$ in (2) to be the vector of common tag frequencies scraped from del.icio.us.
- **Content-based Similarity**. For this metric we take $(v_1, v_2, \ldots, v_M)$ in (2) to be a vector of word frequencies found in the contents of the url itself.

3. Results

We found from a sample of 200 users based on recent tagging activity that the median number of urls tagged is roughly 150, and 75% have tagged less than 400; only 6% have more than 1000, although the largest number found was 5188. Given this, we scraped common tag sets for 6180 urls tagged by 36 users of del.icio.us, selected for having close to the median number of urls – a larger study was prevented by IP blocking that limited our ability to scrape data.

In order for the content-based similarity metric to be calculable, the content pointed to by a url must be *reachable*, and *readable* in the sense that it contains at least some text words. 268 urls, or approximately 4.5% were unreachable, and a further 163, or approximately 2.5% had no words in common with any other user url, and so were deemed to be unreadable[1].

In terms of number of common tags per url, the maximum number of common tags that del.icio.us reports is 25, and roughly 25% of urls have enough tagging data to reach the maximum number of common tags. The most frequent number of common tags is unfortunately not 25 but 0: about 35% of the urls we sampled had no common tags at all; these urls have tagging-based similarity zero to all other urls tagged by the user, which means that a tagging-based recommendation is impossible.

3.1 Performance

The principal measures of performance we used were **coverage** – the number of urls for which it was possible to make a recommendation – and the proportion of users' urls for which the top N recommended tags had larger intersection with the set of true tags than the set of N common tags. The results are shown in Table 1, which reveals that the content-based method is clearly significantly superior to the tagging-based method.

From the point of view of performance, the content-based method was clearly superior to the tagging-based method. The tagging-based method did generate new good tags in nearly a quarter of cases, but the content-based method generated new tags in 40% of cases, and was overall better in a larger proportion of cases than the content-based method.

More promising still, although for the tagging-based method the proportions of winners and losers were statistically indistinguishable, for the content method the winners significantly out numbered the losers, indicating that the content-based method gives a better overall recommendation than common tags. This is surprising because user selection of tags is biased towards the common tags that they see at the time of tagging. One way of interpreting the result is to say that, not surprisingly, a user's tagging behaviour is even more strongly biased towards their own previously used tags. In this context it would be valuable to evaluate our methods against the list of recently used tagging data, which is unfortunately not publicly available.

Surprisingly, given the superior performance of the content-based method in general, in terms of the proportion of uncommon true tags found its performance was indistinguishable from that of the tagging-based method. This apparent conundrum is resolved by observing that the average is taken only over those urls for which each method was able to give a recommendation at all. The tagging-based method fails in all cases where there are no common tags – for which every true tag is therefore uncommon, and on which it is therefore correspondingly difficult for the content-based method to get a good score. Thus the urls covered by content but not tagging-based method are biased with respect to difficulty in recommending a high proportion of uncommon true tags, and this explains why the content method does not do as well as expected.

4. Conclusion

In this paper we have presented a novel method for recommending semantic tags on the basis of similarity metrics derived either from tagging data, or from content analysis. Our method gives personalized recommendations that provide a promising addition to existing tag recommendations based on commonly used tags.

Of the two methods, the tagging-based method is far more lightweight to implement, since it does not require a separate index for the content of the urls itself, but it is less effective at finding good recommendations than the content-based method, and has much lower coverage (although only slightly lower than the coverage of common tags themselves).

References

[1] R. A. Baeza-Yates and B. A. Ribeiro-Neto. *Modern Information Retrieval*. ACM Press / Addison-Wesley, 1999.

[1] Another possibility is that the page to which the url points is there solely for the purpose of redirection, and is otherwise empty.