Tag-Based Filtering for Personalized Bookmark Recommendations

Pavan K. Vatturi*, Werner Geyer, Casey Dugan, Michael Muller, Beth Brownholtz

*School of EECS, Oregon State University
1148 Kelly Engineering Center
Corvallis, OR 97331, USA
+1 (541) 737 3617
vatturi@eecs.orst.edu

IBM T.J. Watson Research
One Rogers Street
Cambridge, MA 02116, USA
+1 (617) 693 4791
{werner.geyer, cadugan, michael_muller,
beth_brownholtz}@us.ibm.com

ABSTRACT
This paper investigates using social tags for the purpose of making personalized content recommendations. Our tag-based recommender creates a personalized bookmark recommendation model for each user based on “current” and “general interest” tags, defined by different time intervals.

Categories and Subject Descriptors
H.3.3 [Information Search and Retrieval]: Information filtering.

General Terms

Keywords
Recommendation, personalization, social tags, bookmarks.

1. INTRODUCTION
The goal of this paper is to study the utility of social tags as a new way to recommend personalized content to users. The basic assumption is simple: When users share content and associate tags with that content, it is likely they would be interested in additional content described by similar tags. As such, these tags can be seen as a fraction of a keyword-based user-interest profile. The more a user tags, the more complete the profile gets, and the more effectively it can be used for recommendation purposes. The advantage of using social tags is that they do not require users to create or update their profiles, or to provide explicit feedback [1]. The tag set can be used as a filter for incoming information and can be automatically updated over time by adding users’ recent tags and removing older ones. Hayes et al. [2] used tags for non-personalized blog recommendations. In our work, we focus on using tags from an enterprise social bookmarking system [3] to create personalized bookmark recommendations.

2. TAG-BASED RECOMMENDER
We create a personalized tag-based recommender for each user as described in Figure 1. Our recommender consists of two Naïve Bayes classifiers trained over different timeframes: One classifier predicts the user’s current interest; the other classifier predicts the user’s general interest in a bookmark. We aggregate both predictions to a final prediction in the following way: If either or both of the two classifiers predict a bookmark as interesting, we recommend the bookmark. If neither classifier predicts the bookmark as interesting, we do not recommend it.

The two classifiers are trained with a subset of the bookmarks created by a user. The tags of each bookmark, converted into a “bag of words”, are used as training features. The core idea is to consider recent bookmarks as good implicit user interest indicators. Previous research has shown that implicit indicators like bookmarks created by the user while browsing the web can be as predictive of interest levels as explicit ratings [1]. For both classifiers, more recent bookmarks are treated as positive training samples, i.e. interesting to the user, whereas older bookmarks are treated as negative training examples i.e. less interesting to the user. The general interest classifier uses bookmarks from a longer time interval as training samples in order to capture general interest topics. The current interest classifier is trained based on a shorter time interval in order to reflect current interests.

Copyright is held by the author/owner(s).
CIKM’08, October 26–30, 2008, Napa Valley, California, USA.

Figure 1. Tag-based Bookmark Recommender.
3. OFFLINE EXPERIMENTS

3.1 Methodology
The data for our experiments were obtained from Dogear, an enterprise social bookmarking site [3], similar to del.icio.us. For our analysis, we used 27 months of Dogear data from June 2005 to September 2007. We selected 60 users based on the following criteria: 30 of the users were required to have more than 500 tags. We call this group high taggers. The remaining 30 users were selected to have more than 200 but less than 500 tags. This group is called low taggers. We trained and tested classifiers for each user with different sets of bookmarks based on different time intervals for the positive and negative samples, using 75% of the bookmarks for training and 25% for testing. The percentage of test bookmarks correctly identified as positive and negative is used as the performance measure for evaluating the classifier.

3.2 Current and General Interest
We conducted an initial experiment in which we trained the current interest classifier with different time intervals for all users. The results were generally low and only marginally better than the random classifier, which can be considered as a baseline with 50% accuracy. A closer look at the training data revealed tags that were common to both positive and negative samples. These tags were regularly used by the user. As such they likely represent a more general interest in these topics as opposed to more current or immediate interests. The presence of these general interest tags in both positive and negative training samples prevented the classifier from differentiating more clearly between the two classes.

To get better prediction results for the current interest classifier, we removed tags that are used more frequently over longer periods of time from the negative and positive training data. In order to determine which tags to remove, we conducted an experiment in which we computed the average accuracy by removing tags with different monthly counts. Results showed that removing tags that are repetitively used in 6 or more months yields the highest accuracy for high taggers Due to the small tag data set for low taggers, removing frequently occurring tags did not yield better results. In order to find the best training interval for the current interest classifier, we conducted another experiment in which we varied the length of the training periods between 1 and 8 weeks. The best combined average accuracy for high and low taggers was achieved with a 4-week training period.

In order to accurately predict if a bookmark is interesting, we also need to take general interest tags into account. To find a good approximation for the training interval for the general interest classifier, we varied the interval between 3 and 9 months. The highest combined average accuracy for high and low taggers was achieved with a training interval of about 6 months.

3.3 Flexible Labeling
The training periods described above yield good results on average. However, we cannot assume that an interest lasts exactly some fixed number of days. To more accurately model the true duration of current and general interests, we used flexible labeling where we moved training samples to the positive or negative training set, respectively, based on which set they occurred more prominently in. Experimental results for the current interest classifier show that the total average accuracy can be improved by 4.2% for high taggers and 2.6% for low taggers.

3.4 Combined Predictions
Finally, the predictions made by the current and general interest classifiers are combined. If either or both of the classifiers predict a bookmark as interesting, then the bookmark is predicted to be interesting as shown in the decision matrix in Figure 1. For realistic testing, we used the test dataset from the current interest classifier including bookmarks tagged with general interest tags. The overall accuracy for our recommender was 80.2% for high taggers and 74.9% for low taggers.

4. USER STUDY
In addition to the offline experiments, we conducted a user study to assess how well our recommender system would work in a real world situation. We recommended bookmarks to a user that were created by other users.

40 high and low taggers participated in the survey which contained 30 bookmarks created by other Dogear users. The recommender predicted 15 of them as interesting and the rest as not interesting to the user. The users were unaware of the recommender predictions and were asked to rate the bookmarks as interesting or not interesting to them. The average accuracy of the user ratings was calculated by comparing the user ratings with the predictions of the tag-based recommender for those 30 bookmarks. We compared the user study results with the results of the offline experiments as shown in Table 1.

Table 1. Accuracy (%) of the tag-based recommender in a user study and offline experiments.

<table>
<thead>
<tr>
<th></th>
<th>User Study</th>
<th>Offline Experiment</th>
</tr>
</thead>
<tbody>
<tr>
<td>High Taggers</td>
<td>70.4% ± 3.5</td>
<td>80.2% ± 1.7</td>
</tr>
<tr>
<td>Low Taggers</td>
<td>64.4% ± 2.9</td>
<td>74.9% ± 2.1</td>
</tr>
</tbody>
</table>

The user study results show that the tag-based recommender performs well with real data. Our users mostly agreed with the decisions made, i.e. 70.4% of the high tagger bookmarks and 64.4% of the low tagger bookmarks were either properly selected as interesting or not interesting.

5. CONCLUSION
The analyses presented in this paper support that the tagging behaviour of a user can be leveraged to make effective personalized bookmark recommendations. Time is a critical component when building such a recommender system. We were able to achieve higher prediction accuracy by modelling different time frames for current and general interest tags.

6. REFERENCES