Exploiting Session-like Behaviors in Tag Prediction

Dawei Yin Liangjie Hong Brian D. Davison
Department of Computer Science & Engineering, Lehigh University
Bethlehem, PA, USA
{day207, lih307, davison}@cse.lehigh.edu

ABSTRACT

In social bookmarking systems, existing methods in tag prediction have shown that the performance of prediction can be significantly improved by modeling users' preferences. However, these preferences are usually treated as constant over time, neglecting the temporal factor within users' behaviors. In this paper, we study the problem of session-like behavior in social tagging systems and demonstrate that the predictive performance can be improved by considering sessions. Experiments, conducted on three public datasets, show that our session-based method can outperform baselines and two state-of-the-art algorithms significantly.

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

General Terms: Algorithms, Experimentation, Performance

Keywords: social tagging, tag recommendation, personalized tag prediction

1. INTRODUCTION

Collaborative tagging systems have become widely used for sharing and organizing resources. While collaborative tagging systems could allow their users to freely choose nearly all possible terms to tag resources, the usability of these systems might be greatly affected and even degraded when vocabularies are totally uncontrolled. Thus, a system that can provide suggestions and recommendations when users are about to assign tags to new resources can improve human-computer interactions and the organization of a collective knowledge base as well.

In web search, it is well-understood that users will create and refine queries that reflect an underlying information need, providing some level of temporal stability in topics of interest as reflected in the sequence of queries used in a search session. In other forms of collaborative filtering and recommendation systems, performance of recommenders has been greatly improved by incorporating temporal factors into the models [1]. However, in tagging systems, temporal behavior analysis is often neglected in existing methods. In the long term, users' high-level interests may be relatively stable and drift slowly over time. In the short-term, however, users may hold a narrow interest within a short time period and so we call it session-like behavior. In this paper, we first verify the existence of session-like behaviors in tagging systems and based on this property, we propose a tag prediction method which can outperform the state-of-the-art.

WWW 2011, March 28–April 1, 2011, Hyderabad, India.
ACM 978-1-4503-0637-9/11/03.

2. SESSION-LIKE BEHAVIORS

Let U be the set of users u, I be the set of resources i, T be the set of tags t and M be the set of timestamps τ. Additionally, S is the set of all records s, representing the relations among the four types of objects, $S \subseteq U \times I \times T \times M$. Each record $(u, i, t, \tau) \in S$ means that user u has tagged an item i with the tag t at time τ. Here, we also define P_s as all the distinct user-item-time combinations: $P_s = \{(u, i, \tau) | \exists t \in T : (u, i, t, \tau) \in S\}$. We use three public datasets, the Bibsonomy dataset of the ECML PKDD 09 Challenge Workshop[1] with content, a nd the Delicious and Flickr datasets crawled by Gorlitz[2] without content. We arbitrarily choose the Jaccard coefficient to measure the similarity of the set of tags between two posts p and p' as $J_{p,p'} = \frac{|T_{p'} \cap T_p|}{|T_{p'} \cup T_p|}$. To verify the existence of session-like behavior, the similarity of posts within a short time interval is calculated. For a post p of user u, we compute its similarity to neighboring posts defined by a time interval τ. $J_u(p, \tau) = \frac{1}{|P'|-1} \sum_{p' \in P', p \neq p'} J_{p,p'}$, where $P' = P_{\tau_p - \tau/2 < \tau < \tau_p + \tau/2, u}$ means the set of posts which are temporally around post p within $\tau/2$. Then we can generate the mean for the whole data set to examine how typical similarity might change as the time interval varies

$$J(t) = \frac{1}{|U|} \sum_u \left[\frac{1}{|P_u|} \sum_{p \in P_u} J_u(p, l) \right]$$

Figure 1 shows the results. The first two points on the time-axis is $\tau = 10$ minutes and $\tau = 30$ minutes and the remaining points are fractions of a day. The three datasets display similar trends. We can see that within 10 minutes, the users' behaviors tend to be the same and for all three datasets, we get highest values on similarity measurement. From 10 minutes to 0.2 days, a dramatic drop occurs and when $\tau > 0.2$ day, the mean similarity decreases slowly. That means, on average, users' interest sessions may hold for up to about five hours, during which users tend to use similar tags. Beyond five hours, users tend to switch to other topics.

3. TAG PREDICTION

From the above analysis, it is clear that users may keep their interest in some topic for a while. The results only show the average proximity on the whole dataset. We believe that different users may have different session lengths. Intuitively, if two consecutive posts do not share any tags, it is highly likely that there is a topic switch. Here, we again use Jaccard's coefficient to define the topic switches. For a user, let p_{i-1},p_i be two consecutive

[1]http://www.kde.cs.uni-kassel.de/ws/dc09/
[2]https://www.uni-koblenz.de/FB4/Institutes/IFI/AGStaab/Research/DataSets/PINTSExperimentsDataSets/

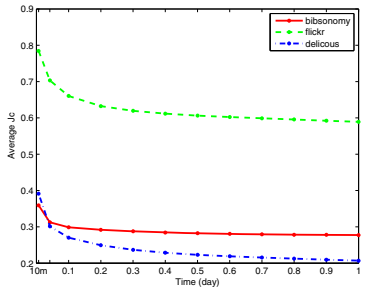

Figure 1: Jaccard's coefficient as a function of time

posts, whose timestamps are $\tau_{i-1} \leq \tau_i$ and tag sets are T_{i-1} and T_i. Use J_{p_{i-1},p_i} as the measurement of the topic switch at post p_i. The personalized session lengths for each user are controlled by a global parameter κ. If $J_{p_{i-1},p_i} < \kappa$, the post J_{p_i} is considered as a topic switch. If we make an assumption that the current test post is not a topic switch post, our method will find an earlier post where the most recent topic switch happened, and predict tags based only on the latest session. Although κ is a shared parameter among all users, it generates personalized topic lifetime for users.

We model the short term interests within a session by assuming that $P(t|u)$—the probability of all tags occurring—follows a multinomial distribution, from which the MLE gives us a simple representation of session-interests:

$$P_{\tau_p}(t|u) = \frac{\sum_{p' \in P'_u} c(t,p'|u)}{\sum_{t'} \sum_{p' \in P'_u} c(t',p'|u)}$$

where $c(t',p'|u)$ is the number of times that tag t' occurs on post p', and usually users use a tag only once. P'_u is the set of posts which belong to the latest interest session for user u. We call it session model.

In the above session model, we made an assumption that the current test post is not a topic switch post; however, in fact, the current post may be the start of a new session. We believe that the time interval from the current test post to the most recent post can help predict such a case. Intuitively, the longer the interval is, the higher probability of the new session starts. To measure whether the current post p_c is the start of a new session, we propose a function $J_{p_c} = f(\tau_c), \mathbb{R} \rightarrow \mathbb{R}$ where J_{p_c} is the predicted tag similarity between the current test post p_c and the last post based on the time interval. For the current test post p_c of user u, we have all past posts of user u—P_u. For every two consecutive posts p_{i-1},p_i, we have a time interval $\tau_i = \tau_{p_{i,u}} - \tau_{p_{i-1}}$ and their similarity value $J_i = J_{p_{i-1},p_i}$. Then we have a set of samples $(\tau_1, J_1), (\tau_2, J_2), ..., (\tau_n, J_n)$, from which we need to learn the function $J_{p_c} = f(\tau_c)$. While there are many regression methods, we use a non-parametric technique—the nearest neighbor method. Compared to kernel methods, the nearest neighbor method defines points local to τ_c not through the fixed kernel bandwidth, but instead on a set of points closest to τ_c, measured by the distance $d_{i,c} = |\tau_i - \tau_c|$. Then the regression at τ_c is calculated as $J_{p_c} = \frac{\sum_i w_i \cdot J_i}{\sum_i w_i}$. w_i is a tri-cube weight function

$$w_i = \begin{cases} (1 - (\frac{d_{i,c}}{d_{k,c}})^3)^3 & d_{i,c} \leq d_{k,c} \\ 0 & d_{i,c} > d_{k,c} \end{cases}$$

where only k of n points closest to τ_c are considered as the neighborhood and $d_{k,c}$ is the distance of the furthest τ_c. Following the previous definition: if $J_{p_c} \geq \kappa$, the current test post will still stay in the current session and the session-based prediction method will be employed while if $J_{p_c} < \kappa$, we will treat this test post as the start of a new session and so at this moment, we will employ content-based methods to predict. In our experiments, the leading content-only method—Lipzcak's method is employed [2].We call it the combination model.

Table 1: F-measure results for tag prediction

Method	Bibsonomy	Delicious	Flickr
Long-term model	0.244	0.162	0.369
Session model	0.333	**0.272**	**0.670**
Combination model	**0.357**	N/A	N/A
LHKM	0.136	N/A	N/A
YXHD	0.309	N/A	N/A

4. EXPERIMENTAL RESULTS

For the Bibsonomy dataset, we use the same test dataset as in [3]. There are 668 test posts which are randomly sampled along the timeline and the remaining posts constitute the training dataset S_{train}. In Delicious and Flickr, we randomly choose 1000 test posts. We use the online evaluation model, which is suggested in [3]. F-Measure is measured at the break-even point. κ is tuned and set as $\kappa = 0.1$.

In all three datasets, we compare this session model with the baseline—long term interests—which simply uses the most frequent tags in the past. The results are shown in the Table 1. As expected, the results of session-based interests model is much better than long term interests. In Bibsonomy data where item content are included, we combined the session-based method with the content-based methods—LHKM [2]. We also compare with the state-of-the-art method [3], which we term YXHD. It is surprising that the results of the long-term interest model can outperform the champion non-personalized content method, LHKM [2]. We find that the combination model achieves the best performance—F-measure of .357. We also notice that even the single session model can outperform the state-of-the-art. This implies a new direction—temporal analysis—for the tag prediction problem.

5. CONCLUSIONS AND FUTURE WORK

In this paper, we verify the existence of session-like behaviors in tagging systems and find that our session-based methods can outperform the baseline methods and two state-of-the-art algorithms. These results also suggest that temporal analysis is an important factor in tagging systems.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation under award IIS-0545875.

6. REFERENCES

[1] Y. Koren. Collaborative filtering with temporal dynamics. In *Proc. of ACM SIGKDD*, pages 447–456, 2009.

[2] M. Lipczak, Y. Hu, Y. Kollet, and E. Milios. Tag sources for recommendation in collaborative tagging systems. In *ECML/PKDD Discovery Challenge Workshop (DC09)*, 2009.

[3] D. Yin, Z. Xue, L. Hong, and B. D. Davison. A probabilistic model for personalized tag prediction. In *Proc. of ACM SIGKDD*, pages 959–968, 2010.