A Probabilistic Model for Personalized Tag Prediction

Dawei Yin, Zhenzhen Xue, Liangjie Hong
and Brian D. Davison
Lehigh University
The Tag Prediction Problem
The Tag Prediction Problem
Why Tag Prediction?

- Improve web search
- Query expansion
- Personalized search
- Capture users' interests
- Automated web resource classification
- Improve efficiency of user interface
Related Work

Content-based Methods

- encode users' preferences from textual information
- can predict tags for new users and new items
- Lipczak et al.'s method [ECML PKDD 2009]

Graph-based Methods

- usually have stronger assumptions than content-based ones
- can provide better performance
- Rendle et al.'s methods [KDD 2009]
Tag Prediction Principles

- The ego-centric effect
- Environmental effects
- Item content
Tag Prediction Principles
Tag Prediction Principles
Tag Prediction Principles

Users’ preference and tag language model
Tag Prediction Principles

Users’ preference and tag language model
Tag Prediction Principles

Users' preference and tag language model

Ego-effects and environmental effects
Tag Prediction Principles

**Prediction Model**

\[ P(t|i, u) \propto \prod_{w \in W_i} \sum_{u_k} P(w|t, u_k)P(u_k|u) \times \sum_{u_k} P(t|u_k)P(u_k|u) \]
Data Set

The bookmark dataset of the ECML PKDD 09 Challenge Workshop

2,679 users, 263,004 items, 56,424 tags, 262,336 posts and 1,401,104 records.

We randomly chose 668 posts as test data
Traditional Evaluation

Randomly Choose Test Posts
Traditional Evaluation

Randomly Choose Test Posts
Traditional Evaluation

Randomly Choose Test Posts

Training posts

Test posts
Traditional Evaluation

Temporally Generate Test Posts
Traditional Evaluation

Temporally Generate Test Posts

Time line

Split time point
Traditional Evaluation

Temporally Generate Test Posts

-Time line-
Traditional Evaluation

Temporally Generate Test Posts

Training posts

Test posts
Online Evaluation
Online Evaluation

Time line
Online Evaluation

Time line
Online Evaluation

Time line

Training posts

Tuesday, July 27, 2010
Online Evaluation

Time line

Training posts

Test post
Online Evaluation

Time line

Training posts
Online Evaluation

Time line

Training posts

Test post
Online Evaluation

Time line

Training posts

Test post
Online Evaluation

Time line

Training posts
Online Evaluation

Time line

Training posts

Tuesday, July 27, 2010
Online Evaluation

Time line

Training posts

Test post

Tuesday, July 27, 2010
Time Sensitive Sampling

From 262,336 posts, randomly choose 668 posts, total 2307 records

Statistics under online evaluation mode

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>New</th>
<th>Old</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>668</td>
<td>41</td>
<td>627</td>
</tr>
<tr>
<td>Items</td>
<td>668</td>
<td>602</td>
<td>66</td>
</tr>
<tr>
<td>Tags</td>
<td>2,307</td>
<td>321</td>
<td>1,986</td>
</tr>
</tbody>
</table>

Tuesday, July 27, 2010
Time Sensitive Sampling

From 262,336 posts, randomly choose 668 posts, total 2307 records

Statistics under online evaluation mode

In the real world, the problem is dominated by the need to predict tags for existing users when they tag new items.

<table>
<thead>
<tr>
<th></th>
<th>Total</th>
<th>New</th>
<th>Old</th>
</tr>
</thead>
<tbody>
<tr>
<td>Users</td>
<td>668</td>
<td>41</td>
<td>627</td>
</tr>
<tr>
<td>Items</td>
<td>668</td>
<td>602</td>
<td>66</td>
</tr>
<tr>
<td>Tags</td>
<td>2,307</td>
<td>321</td>
<td>1,986</td>
</tr>
</tbody>
</table>
Experimental Results

Compare with Lipczak et al.’s methods

- Winner of content-based method ECML PKDD Challenge Workshop
- LHKM-C Content-based method and LHKM-G Graph-based method
Conclusions

Time-sensitive sampling shows that predict tags for existing users when they tag new items.

Online evaluation mode,

more realistic

Our method,

can improve F-measure by over 30%

find that ego-centric effect is often high
Thank you!

Our poster will be in **Poster Session II & Demo Session**

*Date:* Tuesday, July 27, 2010  
*Time:* 5:45pm - 8:00pm  
*Location:* Independence Center B, floor 1