Tag Prediction for Effective Social Media Retrieval

Hsin-Hsi Chen
Department of Computer Science and Information Engineering
National Taiwan University
Taipei, Taiwan
From Taxonomy To Folksonomy

Collaborative Tagging:
Metadata of a resource is generated by both the creator and its readers collaboratively.
del.icio.us

url: http://www.foodsubs.com/
description: The Cook's Thesaurus
notes:
tags: food reference

recommended tags:
export food guide search

your network:
for:joshua for:jwhiting

popular tags:
cooking food reference recipes thesaurus Dictionary cook
puppets ha noi

- vietnam
- hanoi
- Ha
- Noi
- puppet
- play
- show
- water
- puppetry
- doll
Golden Lion at the Venice Film Festival

真的很棒
金吉利，你的影評真是太棒，我很多朋友都有看過，我觀察他們之間的互動，其實跟電影裡面的都完全一樣，所以這個世界沒什麼太大改變，我是一個生意人，平常很忙，沒時間也沒興趣去看電影，但是昨天我多年來第一次進電影院，電影真的很悶，張愛玲的小說也很悶，我很喜歡，因爲很共鳴，男人在追求金錢與權力的道路上，非常辛苦，但是那是一種價值觀的社會肯定他的一種方式，你的影評對李安的用心很深刻，很精彩，我有時間時，會再上來看您其他的作品，完全給你肯定啦！

2007-10-06 14:48 | 小祿

2007-10-01 18:12 | JJ

我不喜歡這部電影，劇中的角色沒有一個讓我感動或是讓我受到衝擊。我感受不到易先生的殺戮感。《红楼梦》的主觀思想，我感受不到女主角的內心衝擊。全片的主題也模模糊糊的。我失望。

2007-09-30 13:56 | joe

看電影多年，第一次，難忘。有人拍手，看電影多年，第一次，細看完了。過了那麼多年還在想劇情。

至於其它的話就不想說了，只想給參予戲所有人道聲謝。

2007-10-06 22:56 | shihyunyin

這是什麼呀？？？
不如建議香港那些黑社會古惑仔電影導演全都殺頭好了＠＠拍外遇的強姦的更是都要宮刑...

2007-09-29 20:50 | MAY
Social Annotation

• Multimedia
  – Text (news, forum, blog, etc)
  – Image
  – Video
• Explicit Tagging vs. Implicit Tagging
• Search, Recommendation, and so on
Social Annotation Enable Searching
From Search Point of View

- Content of resources (by indexing)
- Structures of resources (by link analysis)
- Linguistic resources (by concept expansion)
- User log (by concept recommendation)
- Social Annotation (by ???)
Motivation

• Annotation may facilitate recommendation and effective retrieval.

• Not all resources can gain the benefits from that.
  – Ill-tagged period of URLs prevents them from being retrieved.
  – The retrieval performance for new-coming URLs degrades inevitably.

• How to predict a quality tagging set for a resource is an important issue.
Predicting Social Annotation

Indexing + Annotation Histories = Prediction
Framework of Tag Prediction Algorithm

- Text Collection of Training URLs
- Annotations for Training URLs
- Content Text of Testing URL
- Early Annotations for Testing URL
- Content Tag Selection
- Spreading Activation
- Tag Correlation Computation
- Tag Correlation Graph
- Predicted Tags
Tag Prediction Model (1/5)

• Content Tag Selection
  – Follow the statistical translation model in information retrieval
  – Estimate the probability that a query would be generated as a translation of a document
  – Rank terms in a URL as candidate tags

\[
\text{CTScr} \left( t_i \mid d \right) = \sum_{term_j \in d} \log \left( N \times \max \left( P_c(t_i \mid term_j), \frac{1}{N} \right) \right)
\]

• \( N \): the number of documents in the training data
• \( P_c(t_i \mid term_j) \): the probability of \( t_i \) as a stabilized tag when \( t_i \) and \( term_j \) co-occur in the same document.
• \( D_j \): the set of documents where \( term_j \) occurs
• \( D_{j,i} \): the set of documents in \( D_j \) with \( t_i \) as a stabilized tag.
• \( P_c(t_i \mid term_j) = \frac{|D_{j,i}|}{|D_j|} \).
• All terms in the document are ranked according to their scores.
Tag Prediction Model  (2/5)

• Tag-correlation graph
  – node: tag
  – edge between two nodes: their correlation

• Correlation: the strength of the edge from $t_j$ to $t_i$
  – asymmetric

\[
P_T(t_i \mid t_j) = \frac{\# \text{ of stabilized tag sets containing } t_i \text{ and } t_j}{\# \text{ of stabilized tag sets containing } t_j}
\]
• Spreading Activation
• Spreading Activation
  – At each iteration, node $t_j$ propagates a portion ($\lambda$) of its energy to its neighbors, and gains some energy from its neighbors.

  \[
  E(t_i) = \lambda \times E(t_i) + (1 - \lambda) \times \sum_{t_j \in \text{Neighbor}(t_i)} (E(t_j) \times W(t_j, t_i))
  \]

  \[
  W(t_j, t_i) = \frac{P_T(t_i \mid t_j)}{\sum_{t_k \in \text{Neighbor}(t_j)} P_T(t_k \mid t_j)}
  \]

  – Spreading activation is performed for a fixed number of iterations and eventually tags of the highest energies are proposed as the predicted stabilized tags.
Spreading Activation

- A tag is called an *activation origin* if it is a selected candidate content tag or if it is an early user-annotated tag.
- We initialize the energy for each tag to 0 except those activation origins.
- For each candidate content tag $t_i$ selected from the document $d$, its initial energy $E(t_i)$ is its content selection score $\text{CTScr}(t_i|d)$ normalized by the highest content selection score of terms co-occurring with $t_i$ in $d$.
- For each of the early user-annotated tags, the initial energy is its total tagging by early users.
- For a content tag also annotated by early users, its initial energy is the sum of these two values.
# Experimental Data

<table>
<thead>
<tr>
<th>Dataset</th>
<th># URLs</th>
<th>Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Training</td>
<td>59,090</td>
<td>• 13,934 sufficiently-annotated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• 45,156 annotated by 20+ users</td>
</tr>
<tr>
<td>Testing</td>
<td>2,000</td>
<td>• sufficiently-annotated</td>
</tr>
<tr>
<td></td>
<td></td>
<td>• top 25 stabilized tags</td>
</tr>
</tbody>
</table>
### Experimental Results

<table>
<thead>
<tr>
<th>Metric Method</th>
<th>R-Precision</th>
<th>Recall</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>CTScr</td>
<td>0.2687</td>
<td>0.3569</td>
<td>0.1768</td>
</tr>
<tr>
<td>User-5</td>
<td>0.2359</td>
<td>0.2343</td>
<td>0.2077</td>
</tr>
<tr>
<td>CTScr + User-5</td>
<td>0.3972</td>
<td>0.4733</td>
<td>0.3439</td>
</tr>
<tr>
<td>CTScr+User-5+SA (λ=0.5, iteration=2)</td>
<td>0.4493</td>
<td>0.6072</td>
<td>0.3879</td>
</tr>
<tr>
<td>User-10</td>
<td>0.3445</td>
<td>0.3436</td>
<td>0.3025</td>
</tr>
<tr>
<td>CTScr + User-10</td>
<td>0.4591</td>
<td>0.5360</td>
<td>0.4154</td>
</tr>
</tbody>
</table>
Tag Categorization

• Categorize the tags into three types according to the relations between tags and target URL.
  – Topic-description Tags
    • stabilized tags describe the implicit topic of the target URL
    • highly correlated to the terms in the URL content, and correlated to some other tags of the same type
  – Function-related Tags
    • describe or conceptually related to the function of the target URL
    • some of them are highly correlated to other tags but most of them usually have no evident correlations with terms in the text content
  – Personal-use Tags
    • occur in the stabilized tag set arbitrarily
    • E.g., read and reference
## Tag Normalization

<table>
<thead>
<tr>
<th>Canonical Form</th>
<th>Group of Tag Strings</th>
</tr>
</thead>
<tbody>
<tr>
<td>common lisp</td>
<td>Commonlisp, clisp, common_lisp</td>
</tr>
<tr>
<td>drag and drop</td>
<td>drag-and-drop, draganddrop, dragndrop</td>
</tr>
<tr>
<td>online game</td>
<td>Onlinegames, online_game, olg</td>
</tr>
<tr>
<td>web 2.0</td>
<td>web2.0, web_2.0, web2, web20, web_20</td>
</tr>
</tbody>
</table>
Relevant Publications


Conclusion

• The experiment results show that our tag prediction model is able to predict a considerably large portion (~51%) of the stabilized tag set with only 5 user annotations.

• That facilitates recommendation and effective retrieval of new-coming resources.
Thank You!