An Analysis of Tag-Recommender Evaluation Procedures

Stephan Doerfel & Robert Jäschke
University of Kassel & L3S Research Center
doerfel@cs.uni-kassel.de & jaeschke@l3s.de

Social Bookmarking & Folksonomies

Social bookmarking systems: users can collect and manage resources like bookmarks, publications, images, videos, etc.

Folksonomy: underlying data structure that models the process of users creating posts by annotating resources with freely chosen keywords – so-called tags

Through tagging, users collaboratively generate a corpus of publicly visible, annotated resources. Resources can be retrieved using the tags and can be shared with other users.

Figure 1: A folksonomy tag example with three users A, B, C, and five tags 1, 2, 3, 4, 5 (○). The circle represents the presence of a user for a tag.

Figure 2: The graph-core at level 2.

Figure 3: The post-core at level 2.

Graph-Core

Cores

• Folksonomies are modeled as graphs, where users, resources, and tags form the node set. User u is connected to resource r and tag t by a hyperedge, if u assigned t to r.

• Cores reduce the dataset by iteratively removing nodes (and all connected edges) until all remaining nodes satisfy some specific property.

• Sedlmeier and later Batagelj and Zaversnik developed the theory on cores to analyze graphs.

For tag recommendations they are used commonly to yield denser subgraphs of a folksonomy.

Does the choice of core influence the evaluation?

Cleansing

Before the experiments we conducted appropriate preprocessing to remove automatic imports by:

• eliminating posts created at exactly the same time by the same user

• ignoring tag assignments with the tags imported, public, system:imported, no system:unfilled

Additionally, all tags were converted to lower case, and all characters which were neither numbers nor letters were removed.

Cores

Evaluation

Evaluation Methodology

LeavePostOut: For each user u conduct the following experiment: Select one post of u at random, remove it from the data and recommend tags for its resource for that user.

Repeat LeavePostOut for each user 5 times for statistical validity.

Evaluation: Compare the recommended tags to the actual tags of the left-out posts. Use precision@5, recall@5, and MAP to evaluate the recommender quality.

Evaluate influence of cores: Repeat the experiments on cores at different levels with five well-established tag recommendation algorithms: most popular tags, most popular tags by resource, most popular tags by user, adapted PageRank, and FolkRank.

Recommender Ranking Correlation

Table 1: The mean pairwise Pearson’s r, the number of discontent pairs (i, j) in the algorithm rankings on graph-cores and post-cores of different levels. Each ranking is a list of the five recommendation algorithms, ordered by their recommendation quality according to one of the measures MAP, prec@5 or rec@5.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Metric</th>
<th>Level</th>
<th>Prec@5</th>
<th>Rec@5</th>
<th>MAP</th>
</tr>
</thead>
<tbody>
<tr>
<td>public</td>
<td>0.999</td>
<td>0.005</td>
<td>2.491</td>
<td>1.005</td>
<td></td>
</tr>
<tr>
<td>deli</td>
<td>0.986</td>
<td>0.101</td>
<td>1.636</td>
<td>1.007</td>
<td></td>
</tr>
<tr>
<td>book</td>
<td>0.964</td>
<td>0.009</td>
<td>1.564</td>
<td>1.014</td>
<td></td>
</tr>
<tr>
<td>book</td>
<td>0.999</td>
<td>0.002</td>
<td>1.566</td>
<td>1.009</td>
<td></td>
</tr>
<tr>
<td>deli</td>
<td>0.870</td>
<td>0.156</td>
<td>1.564</td>
<td>1.151</td>
<td></td>
</tr>
<tr>
<td>deli</td>
<td>0.902</td>
<td>0.091</td>
<td>1.455</td>
<td>1.066</td>
<td></td>
</tr>
<tr>
<td>deli</td>
<td>0.998</td>
<td>0.011</td>
<td>0.545</td>
<td>0.503</td>
<td></td>
</tr>
<tr>
<td>deli</td>
<td>0.998</td>
<td>0.011</td>
<td>0.545</td>
<td>0.503</td>
<td></td>
</tr>
</tbody>
</table>

Results:

• High correlations among the rankings of algorithms

• On the smaller datasets (public, book) on average one or two recommenders switch their positions in their ranking

• Higher consistency on the larger dataset deli

Recommendations

• Test your recommender within a real system!

• Framework for tag and item recommendations

• Contact: doerfel@cs.uni-kassel.de

Figure 4: The performance of different tag recommendation algorithms using the graph-core as dataset and samples from the graph-core (●) or the post-core (○) for LeavePostOut.

For the latter we use the property that the post-core is always a subset of the graph-core.

Conclusion

• Recomenders perform differently in different core setups of the same dataset.

• Evaluating recommender performance on another core type or at another core level might look like changes in the results

• Focusing on one particular core can produce non-stable results.

• No guarantee that the best recommender in one setup is also the best in another setup (even on the same dataset).

• But: Even cores at higher levels yield correlate results to those of the raw-data.

Evaluation should always be performed either directly on the raw data or on several core types and levels.

© 2013 Stephan Doerfel & Robert Jäschke