Feeding a Hybrid Recommendation Framework with Linked Open Data and Graph-based Features

Background

Non-trivial and useful features available in RDF format

Linked Open Data cloud

Topological features can be calculated by mining the tripartite graph-based representation connecting users, items and resources in the LOD cloud

150 billions triples and 10k datasets

Research Questions

“Is it possible to encode this information to enrich item representation in a hybrid recommendation framework?”

“How do LOD-based features and graph-based features impact on the overall performance of the recommender system?”

Methodology

Basic Features

Popularity features 
#ratings, ratio of positive ratings

Collaborative features 
We encoded a column of the users/items matrix

Content-based features 
Text was tokenized and stemmed through Lucene and Snowball

Extended Features

LOD-based features 
The most relevant features are extracted from Dbpedia by mapping item names to URIs

Graph-based features 
Degree Centrality, Average Neighbor Degree, PageRank score, Node Redundancy and Cluster Coefficient, calculated with Jung library

Recommendation Framework

Item Representation 
Items represented through different combinations of basic and extended features

Recommendation

Unseen items are labeled as relevant or not relevant by a classification algorithm

Results

Experiment 1 - Performance of basic features

F1@5

<table>
<thead>
<tr>
<th></th>
<th>MovieLens</th>
<th>DBbook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Popular (P)</td>
<td>53.38</td>
<td>55.67</td>
</tr>
<tr>
<td>Collaborative (C)</td>
<td>56.18</td>
<td>56.42</td>
</tr>
<tr>
<td>Content (T)</td>
<td>56.35</td>
<td>56.78</td>
</tr>
<tr>
<td>P+C</td>
<td>56.32</td>
<td>56.59</td>
</tr>
<tr>
<td>P+T</td>
<td>55.49</td>
<td>55.83</td>
</tr>
<tr>
<td>C+T</td>
<td>55.83</td>
<td>56.67</td>
</tr>
<tr>
<td>P+C+T</td>
<td>56.78</td>
<td></td>
</tr>
</tbody>
</table>

Datasets: MovieLens (1M ratings, 96.4% sparsity) and DBbook (72k ratings, 99.8% sparsity). Random Forests as classification algorithm. Metric: F1@5

Experiment 2 - Impact of LOD-based and graph-based features

F1@5

<table>
<thead>
<tr>
<th></th>
<th>MovieLens</th>
<th>DBbook</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>56.35</td>
<td>56.07</td>
</tr>
<tr>
<td>Baseline+Graph</td>
<td>56.21</td>
<td>56.78</td>
</tr>
<tr>
<td>Baseline+LOD</td>
<td>56.27</td>
<td>56.59</td>
</tr>
<tr>
<td>Baseline+LOD+Graph</td>
<td>56.67</td>
<td></td>
</tr>
</tbody>
</table>

Outcome: LOD + Graph-based features led to the best results

AI*IA 2017 – 16th International Conference of the Italian Association for Artificial Intelligence

Bari, Italy - November 15, 2017 - Poster Session