Comparing Vocabulary Term Recommendations using Association Rules and Learning To Rank

A User Study

Johann Schaible, Pedro Szekely, and Ansgar Scherp

at ESWC 2016
Problem Statement: 
Reuse Vocabulary Terms!

- When modeling LOD, it is accustomed to reuse vocabulary terms (→ classes and properties)
- However, it is a challenging task

Vocabulary Term Recommendations
Term Recommendations based on...

- Popularity of a candidate (i.e. vocabulary term)
  - Number of LOD sources using a candidate
  - Number of LOD sources using candidate’s vocabulary
  - Number of total occurrences of a candidate
- Candidate from an already used vocabulary
- Collaborative filtering:
  - How did others use a recommendation candidate

Example of Schema-Level Patterns (SLPs):

\[ slp = (\{\textit{swrc:Publication}\}, \{\textit{dc:creator}\}, \{\textit{foaf:Person}\}) \]

Resources of type \textit{swrc:Publication} are connected to resources of type \textit{foaf:Person} via the property \textit{dc:creator}.
Utilized Approaches (1/2): Association Rules (AR)

- Rules calculated on the set of frequent SLPs:

\[ \text{SLP}_{LOD} = \text{SPLs computed from datasets on the LOD cloud} \]

\[ \text{SLP}_{LOD} = \{slp_1, slp_2, ..., slp_n\} \]

\[ slp_i = (\{\text{swrc:Publication}\}, \{\text{dc:creator}\}, \{\text{foaf:Person}\}) \]

\[ s_i = (\{\text{swrc:Publication}\}, \{\}, \{\text{foaf:Person}\}) \]

When using a set of given vocabulary terms, which further classes and properties did others also use?

Recommendation:

\[ s_i \rightarrow (slp_i - s_i) := \text{dc:creator} \]
Utilized Approaches (2/2): Learning To Rank (L2R)

- Supervised machine learning algorithm based on data with relevance annotations
- State of the art in IR to compute a generalized ranking model over a given set of features
- Ranking model is derived by observing correlations between feature values and candidate relevance

Features:
- (i) number of datasets using a vocabulary term,
- (ii) number of total occurrences of a vocabulary term
- Term from an already used vocabulary
- SLP-feature
Why a User Study?

- In offline evaluations
  - There is no gold standard data
  - No observations of users and their behavior
- In A/B-Tests (online evaluation)
  - No full functioning system yet
  - Not enough users to make meaningful results

Study in a controlled lab environment with invited participants
Latin-square within subject design study
- Each participant asked to model three different datasets as LOD (max. 6 minutes each) with
  (a) Learning To Rank based recommendations,
  (b) Association Rule based recommendations,
  (c) No recommendations

Participants first train on example data
- Avoids carry-over effects
- Participants get used to the system
Task: Finish the model for the data from Music, Museum, and Product Offers domain with Karma\(^1\)

- Replace \texttt{owl:Thing} and \texttt{rdfs:label} with \textit{better} fitting classes, properties respectively
- Define object properties specifying that
  - a musician is a member of a band
  - a musician recorded an album
  - a musician has a Wikipedia page

1) \url{http://usc-isi-i2.github.io/karma/}
User Study - Measurements

- The participants’ effort
  - Task Completion time (max. 6 min. to avoid tiredness)
  - Recommendation acceptance rate (number of terms chosen from recommendations)

- The quality of the resulting data
  - Number of vocabulary terms that were also used by five different data modeling experts

- Level of satisfaction with both recommenders
  - 5-point Likert scale rating AR and L2R
  - Ranking of L2R, AR, and using no recommendations
Results (1/3) - Participants

- 20 participants (5 female)
  - 18 in academia, 2 in both academia and industry
  - 2 master students, 14 research associates, 3 post docs, 1 professor
  - 8 recruited from USC, 12 recruited from GESIS

- Knowledge and experience
  - Karma: 7 high to expert knowledge, 13 none at all
  - LOD: average experience of 3 years
  - Self-rated experience (5-point Likert): $M = 2.8$, $SD = 1.6$
  - Task knowledge (5-point Likert): $M = 2.1$, $SD = 1.1$
Results (2/3)

### Time Completion

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>Music domain</th>
<th>Museum domain</th>
<th>Product Offer domain</th>
<th>Average in total</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2R</td>
<td>5:31 (0:41) - 6</td>
<td>5:48 (0:24) - 7</td>
<td>5:42 (0:25) - 7</td>
<td>5:41 (0:30) - 20</td>
</tr>
<tr>
<td>AR</td>
<td>4:50 (1:10) - 6</td>
<td>4:37 (1:02) - 7</td>
<td>3:16 (1:02) - 7</td>
<td><strong>4:13 (1:03) - 20</strong></td>
</tr>
<tr>
<td>No Rec</td>
<td>5:22 (0:59) - 8</td>
<td>5:28 (0:37) - 6</td>
<td>5:34 (0:42) - 6</td>
<td>5:28 (0:47) - 20</td>
</tr>
</tbody>
</table>

### Recommendation Acceptance

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>Music domain</th>
<th>Museum domain</th>
<th>Product Offer domain</th>
<th>Average in total</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2R</td>
<td>2.00/7 (.89)</td>
<td>2.37/7 (.91)</td>
<td>1.67/6 (.81)</td>
<td><strong>2.05/6.67 (.88)</strong></td>
</tr>
<tr>
<td>AR</td>
<td>4.33/7 (.51)</td>
<td>4.85/7 (.69)</td>
<td>5.28/6 (.95)</td>
<td><strong>4.85/6.67 (.95)</strong></td>
</tr>
</tbody>
</table>

### Data Quality

<table>
<thead>
<tr>
<th>Recommendations</th>
<th>Music domain</th>
<th>Museum domain</th>
<th>Product Offers domain</th>
<th>Average in total</th>
</tr>
</thead>
<tbody>
<tr>
<td>L2R</td>
<td>4.02/7 (0.64)</td>
<td>3.75/7 (1.05)</td>
<td>3.87/6 (0.83)</td>
<td><strong>3.88/6.67 (.14)</strong></td>
</tr>
<tr>
<td>AR</td>
<td>4.87/7 (0.88)</td>
<td>4.75/7 (1.03)</td>
<td>5.25/6 (0.64)</td>
<td><strong>4.95/6.67 (.26)</strong></td>
</tr>
<tr>
<td>No Rec</td>
<td>3.84/7 (0.83)</td>
<td>3.37/7 (0.91)</td>
<td>3.52/6 (0.70)</td>
<td>3.57/6.67 (.24)</td>
</tr>
</tbody>
</table>
Results (3/3)

- General level of satisfaction (5-point Likert scale)
  - Learning To Rank: $M = 3.00, SD = 1.1$
  - Association Rules: $M = 4.23, SD = 0.7$

- Comparing AR to L2R directly
  - Rating: $M = 4.56, SD = 0.4$
    - AR much worse
    - AR much better
  - Ranking: All participants ranked AR higher than L2R
Discussion of Results

- AR filters out inappropriate terms, L2R ranks them at a lower position
- Additional features let L2R rank popular but inappropriate terms higher
- With L2R based recommendations, it was observed that participants
  - overlooked relevant recommendation in the top-10 list
  - felt uncertain, such that they searched longer and often used string based search
Conclusion

- AR based recommendations, i.e., collaborative filtering, performs better in
  - Time, effort, quality, general satisfaction
- With L2R unsure, with AR more sure the recommendations are correct and commonly used

Using AR-based recommendations, participants with little LOD and domain expertise were able to produce high quality LOD close to the experts

⇒ Easier vocabulary reuse to decrease heterogeneity in data representation
Thank You!

Acknowledgements:
Generating the gold standard:
Laura Hollink, Benjamin Zapilko, Ruben Verborgh, Jérôme Euzenat, and Oscar Corcho

Thanks to the participants of the user study

For the interested reader:
1. Copy of the questionnaire and raw results of the user study:
   http://dx.doi.org/10.7802/1206
2. Accompanying material and modeling results:
   https://github.com/WanjaSchaible/termpicker_karmaeval_material