Supporting Resource-based Learning on the Web using automatically extracted Large-scale Taxonomies from multiple Wikipedia versions

Graph with links classified as "is-a"

Wikipedia Category Graph

Taxonomy

Dipl.-Inform. Renato Dominguez García
Dr.-Ing. Philipp Scholl
Dr.-Ing. Christoph Rensing

Prof. Dr.-Ing. Ralf Steinmetz
KOM - Multimedia Communications Lab
Overview

1 CROKODIL: Our Application Scenario
2 Our Approach
3 Evaluation
4 Summary and Conclusions
1 Our Application Scenario: CROKODIL

CROKODIL
- Support of self-directed, collaborative, resource-based learning
- Management of resources + Social network
- Based on Semantic Networks

Our goal
- Recommendation based on Taxonomies
Our Application Scenario: Requirements

Domain-independent, up-to-date and multilinguality

- Support of a large coverage of topics
- Support of current topics
- Support of different languages

→ Wikipedia fulfill these requirements

Why different languages?

- Users tag resources in different languages
- Socio-cultural specific knowledge
- E.g. “Castor-Transport”

Our Idea
2 Our Approach

Our Task
- Label pairs of categories (links) as is-a or not-is-a
- Using Machine Learning
  - Corpus (Feature vectors)
  - Classifier

We extract following features from a Link
- Syntactic Features
  - Position of Head
  - Co-occurrence of words
- Structural Features
  - Co-occurrence of articles
  - Number of super- and subcategories of $c_1$ and $c_2$
  - Incoming and Outgoing Wikilinks between $a_1$ and $a_2$
- Article features
  - Definition Sentence
  - Head in article content
  - Existence of $c_1$ and $c_2$ as articles
3 Evaluation

Our Corpus

<table>
<thead>
<tr>
<th>Language</th>
<th>English</th>
<th>Spanish</th>
<th>German</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of is-a links</td>
<td>1297 (29.9 %)</td>
<td>786 (36.1 %)</td>
<td>808 (33.6 %)</td>
<td>1135 (41.4 %)</td>
</tr>
<tr>
<td>Number of not-is-a links</td>
<td>3048 (70.1 %)</td>
<td>1388 (63.9 %)</td>
<td>1597 (66.4 %)</td>
<td>1604 (58.6 %)</td>
</tr>
</tbody>
</table>

Classification

- Decision trees (C 4.5)
- 10-fold cross-validation for results

Results

<table>
<thead>
<tr>
<th>Language</th>
<th>English</th>
<th>Spanish</th>
<th>German</th>
<th>Arabic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correctly classified links</td>
<td>3539 (81.4 %)</td>
<td>1841 (84.7 %)</td>
<td>1968 (81.8 %)</td>
<td>2255 (82.3 %)</td>
</tr>
<tr>
<td>Incorrectly classified links</td>
<td>806 (18.6 %)</td>
<td>333 (15.3 %)</td>
<td>437 (18.2 %)</td>
<td>484 (17.7 %)</td>
</tr>
<tr>
<td>Total number of links</td>
<td>4345</td>
<td>2174</td>
<td>2405</td>
<td>2739</td>
</tr>
</tbody>
</table>
4 Summary and Conclusions

Recommendations based on taxonomies can be used to support users in a collaborative learning scenario.

Extracting of taxonomies in multiple languages based on Wikipedia is feasible without relying on additional sources.

Next Steps

- Integration in CROKODIL
- Evaluation with real users
Questions & Contact

Dipl.-Inform. Renato Dominguez Garcia

Renato.Dominguez.Garcia@KOM.tu-darmstadt.de
Rundeturmstr. 10
64283 Darmstadt
Germany

Phone +49 (0) 6151/166139
Fax  +49 (0) 6151/166152
www.kom.tu-darmstadt.de

Source: http://www.crokodil.org