NJVR: The NanJing Vocabulary Repository

Gong Cheng, Min Liu, and Yuzhong Qu
Motivation

Vocabulary-oriented research topics

A large and freely-accessible collection of real-world vocabularies
State of the art

**Top-down efforts**

- **DAML Ontology Library**
- **Schemapedia**
  - RDF schema compendium

**Bottom-up efforts**

- **Swoogle**
  - Semantic web search 2007
- **Watson**
  - Exploring the semantic web

**Size:** small (hundreds) → **Access:** directly (via browsing) → **Size:** large (thousands) → **Access:** indirectly (via searching) →
State of the art

**Top-down efforts**

- **BioPortal**
  - DAML Ontology Library
- **Schemapedia**
  - RDF schema compendium

**Bottom-up efforts**

- **Swoogle**
  - semantic web search

**Our goal**

- **Size**: small (hundreds)
- **Access**: directly (via browsing)
- **Size**: large (thousands)
- **Access**: indirectly (via searching)
Contribution

• NJVR: A large and freely-accessible vocabulary repository
  – **Source**: An index of **4.1B** RDF triples distributed in **15.9M** RDF documents crawled from **5.8K** pay-level domains (PLDs)
  – **Constitution**:
    • RDF descriptions of **2,996** dereferenceable vocabularies crawled from **261** PLDs
    • Document-level statistical data on their instantiations (e.g. term frequency)
  – **Accessibility**: Publicly downloadable
Construction of NJVR

1. Crawling
2. Vocabulary identification
3. Vocabulary instantiation
Crawling (2007—May 2011)

1. Initialization (of the URI pool)
   - Other freely-accessible repositories, e.g. pingthesemanticweb.com
   - LOD cloud
   - Search results, e.g. Swoogle, Google

2. URI Dereference and document parsing
   - java.net package
   - Jena

3. Pool expansion
   - URIs in parsed documents
   - Submissions from the users of Falcons
Vocabulary identification

• Bottom-up strategy
  1. **Term**: URI that identifies a class/property in its dereference document
  2. **Vocabulary**: Terms in a common namespace
Results

• 455,718 terms
  – 396,023 classes, 59,868 properties, 173 both

• 2,996 vocabularies
  – From 261 PLDs

Fig. 1. Log-log plot of the distribution of the number of vocabularies crawled from a PLD.  
Fig. 2. Log-log plot of the distribution of the number of constituent terms of a vocabulary.
Vocabulary instantiation

- Instantiation found for:
  - 115,707 classes (29.2%)
  - 25,963 properties (43.4%)
  - 1,874 vocabularies (62.6%)

Fig. 3. Log-log plot of the distribution of the number of PLDs offering instantiations of a class.

Fig. 4. Log-log plot of the distribution of the number of PLDs offering instantiations of a property.
Applications of NJVR

- Vocabulary matching
- Vocabulary ranking
- ...
NJVR for vocabulary matching

• Using NJVR as a source of test cases for vocabulary matching
  – Abundant matchable vocabularies?

**Fig. 5.** Log-linear plot of the distribution of the number of pairs of matchable vocabularies under different settings.
NJVR for vocabulary ranking

- Using NJVR as a test case for vocabulary ranking

Table 1. Correlation between rankings generated by different measures of centrality

<table>
<thead>
<tr>
<th></th>
<th>Indegree</th>
<th>Closeness</th>
<th>Betweenness</th>
<th>Eigenvector</th>
<th>PageRank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Closeness</td>
<td></td>
<td>0.075</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Betweenness</td>
<td>0.709</td>
<td></td>
<td>0.187</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Eigenvector</td>
<td>0.995</td>
<td>0.080</td>
<td>0.698</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PageRank</td>
<td>0.996</td>
<td>0.079</td>
<td>0.700</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>HITS authority</td>
<td>0.996</td>
<td>0.077</td>
<td>0.693</td>
<td>0.991</td>
<td>0.992</td>
</tr>
</tbody>
</table>
Future work

• Removal of low-quality vocabularies from NJVR
• Comparative analysis of NJVR and other repositories
• ...

Just use it!

ws.nju.edu.cn/njvr