Master Thesis Presentation

Automated Ontology Mapping of Tagged Data in a Pandisciplinary Repository for Research Projects

M. Politze
Thesis Committee:
N. Roos, F. C. Schadd, B. Decker
Outline

- **Introduction**
  - Setting and Problem Statement
  - Metadata and Description Logic
  - Ontologies and Semantic Web

- **Semantic Repository**
  - Semantic Extensions to SharePoint
  - Evaluation

- **Ontology Matching**
  - Structural Measures
  - Evaluation

- **Conclusion and Future Research**
Setting

- Part of the research project: Projekt Repository
- Project participants: CCC + several institutes of RWTH Aachen University
- Funded by German Research Foundation
- Goal: Build a central repository infrastructure based on SharePoint to store and retrieve research data
Goal

Lay theoretical basis for an extension to SharePoint that...

- allows storing, retrieving and updating ontologies.
- uses the structure provided in the ontology to structure the repository.
- offers retrieval techniques based on that structure.
- provides an interface that allows multiple repositories to exchange data.
- offers long term retrievevability of the data.
Ontologies

- A philosophical discipline about
  - things that exist
  - their categories of being
  - their relations

- In computer science
  - Formal semantics readable by a computer
  - Link real-world terminologies and things to computer processable content
  - [Based on description logic (OWL)]

- Why do we need ontologies?
Metadata and Description Logic

Metadata
Structured information about data:

→ Type of image: microscopy
→ Date taken: 14.03.2012
→ Region of the body: left breast
→ Diagnosis: 4 (Suspicious abnormality)

Description Logic
Language to formally model classes (concepts), individuals and their relationships:

→ (image, microscopy): imageType
→ (image, 14.03.2012): dateTaken
→ (image, left breast): bodyRegion
→ (image, 4): diagnosis

Definition of rules such as

MicroscopeImage = ∃ imageType. microscopy
MicroscopeImage ⊑ Image
Web Ontology Language & Semantic Web

- Web Ontology Language (OWL) is a language to define an ontology
- Describes the terminology as well as the assertions
- Offers XML Syntax
- OWL is recommended by the W3C for exchange of ontologies via the internet
- Semantic Web allows data to be
  - shared across application, enterprise, and community boundaries
  - reused
Research Questions

- Which structures of Microsoft SharePoint can be used to represent ontologies?

- How can these structures be used to tag data saved in the repository?

- How can different retrieval techniques be used to retrieve data from the repository?

- Can different ontologies be matched using only their structure?

- Can structural measures describe the elements of an ontology?
What is SharePoint?

- Enterprise-scale process support
- Extendible using .NET programming languages
- Out-of-the-box
  - Users, rights and roles
  - File storage
  - Semantic features?
SharePoint Lists

Document library based on ontology: dterms.rdf

<table>
<thead>
<tr>
<th>Typ</th>
<th>Name</th>
<th>Geändert</th>
<th>Geändert von</th>
<th>License</th>
<th>rdf:type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ApacheLicense</td>
<td>14.06.2012 09:27</td>
<td>Marius Politze</td>
<td>-</td>
<td>License Document</td>
</tr>
<tr>
<td></td>
<td>CreativeCommonsLicense</td>
<td>14.06.2012 16:09</td>
<td>Marius Politze</td>
<td>-</td>
<td>License Document</td>
</tr>
<tr>
<td></td>
<td>image</td>
<td>14.06.2012 16:12</td>
<td>Marius Politze</td>
<td>-</td>
<td>License Document</td>
</tr>
<tr>
<td></td>
<td>sp_site_structure</td>
<td>14.06.2012 09:30</td>
<td>Marius Politze</td>
<td>-</td>
<td>License Document</td>
</tr>
</tbody>
</table>

Dokument hinzufügen
Example

- Based on the Dublin Core Metadata Terms

<table>
<thead>
<tr>
<th>Title</th>
<th>Creator</th>
<th>Subject</th>
</tr>
</thead>
<tbody>
<tr>
<td>Description</td>
<td>Publisher</td>
<td>Contributor</td>
</tr>
<tr>
<td>Creation Date</td>
<td>Type</td>
<td>Format</td>
</tr>
<tr>
<td>Identifier</td>
<td>Source</td>
<td>Language</td>
</tr>
<tr>
<td>Relation</td>
<td>Coverage</td>
<td>Rights</td>
</tr>
</tbody>
</table>

- apacheLicense.txt: RightsStatement
- (apacheLicense.txt, 2004 01 01): createdAt
- (apacheLicense.txt, Apache License v2): titleOf
- (file.pdf, 2012 06 13): createdAt
- (file.pdf, Testfile): titleOf
- (file.pdf, apacheLicense.txt): rightsDocumentOf
Example
### Semantic Repository: Ontology Instantiation

**dcterms.rdf** > ontology::dcterms.rdf

Based on ontology::dcterms.rdf

<table>
<thead>
<tr>
<th>Typ</th>
<th>Name</th>
<th>Date Created</th>
<th>rdf:type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>ApacheLicense</td>
<td>01.01.2004</td>
<td>License Document</td>
</tr>
</tbody>
</table>

**dcterms.rdf2** > ontology::dcterms.rdf

Based on ontology::dcterms.rdf

<table>
<thead>
<tr>
<th>Typ</th>
<th>Name</th>
<th>Geändert</th>
<th>Geändert von</th>
<th>License</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>sp_site_structure</td>
<td>13.06.2012 00:30</td>
<td>Marius Politze</td>
<td>1 - ApacheLicense.txt(Microsoft.SharePoint.Taxonomy.LabelCollection)</td>
</tr>
</tbody>
</table>
Semantic Repository: “Type of” Relation

<table>
<thead>
<tr>
<th>Coverage</th>
<th>Creator</th>
<th>Audience Education Level</th>
<th>Extent</th>
<th>Format</th>
<th>Instructional Method</th>
<th>Language</th>
<th>License</th>
<th>Mediator</th>
<th>Medium</th>
<th>Provenance</th>
<th>Publisher</th>
<th>Rights</th>
<th>Rights Holder</th>
<th>Spatial Coverage</th>
<th>Temporal Coverage</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

![Diagram of RDF types]

- rdf:type
  - Agent
  - Agent Class
  - Bibliographic Resource
  - Class
  - Frequency
  - Linguistic System
  - Location, Period, or Jurisdiction
  - Media Type or Extent
  - Method of Accrual
  - Method of Instruction
  - Physical Resource
  - Policy
  - Provenance Statement
  - Rights Statement
  - License Document
  - Standard

---

Master Thesis Presentation
Marius Politze | Center for Computing and Communication
Semantic Repository: Inter Class Relations
Semantic Repository: Search & Retrieval

- **Search by Keyword**
  - Strict, only exact matches
  - Widely supported by SharePoint

- **Faceted Search**
  - Uses classes from ontologies as facets
  - Guides the user

- **OWL Export**
  - Allows round trip engineering
  - Import of information from one repository by an other
  - Information for arbitrary agents
Semantic Repository: OWL Export

<owl:Class rdf:about="&dcterms;LicenseDocument">
  <rdfs:label>Licenсе Document</rdfs:label>
  <rdfs:subClassOf rdf:resource="&dcterms;RightsStatement"/>
</owl:Class>

<owl:ObjectProperty rdf:about="&dcterms;license">
  <rdfs:label>Licenсе</rdfs:label>
  <rdfs:range rdf:resource="&dcterms;LicenseDocument"/>
</owl:ObjectProperty>

<owl:NamedIndividual rdf:about="file.pdf">
  <dcterms:title>Testfile</dcterms:title>
  <dcterms:license rdf:resource="ApacheLicense.txt"/>
</owl:NamedIndividual>

<owl:NamedIndividual rdf:about="ApacheLicense.txt">
  <dcterms:title>Apache License v2</dcterms:title>
  <dcterms:created>01.01.2004</dcterms:created>
</owl:NamedIndividual>
Semantic Repository: Use Case Evaluation

1. [ontology not available in OWL]
2. Create OWL ontology based on Structure given by researcher
3. Upload of ontology into SharePoint Repository
4. Create one or more Tables with the classs instantiation property
5. Load ontology into Protégé to review
6. Create one or more Tables with other properties from the ontology
7. upload several files to the repository and edit the generated properties
8. [ontology not available in OWL]
Ontology Matching: Task Description

- Long term retrievability needs to deal with changing ontologies!
- The entities from the old and the new ontology need to be matched to each other
- Find a mapping $A$ from one ontology to an other ontology and define a mapping function:

$$f(e_1) = \begin{cases} e_2, & \text{if } A \text{ maps } e_1 \text{ to } e_2 \\ \bot, & \text{otherwise} \end{cases}$$
Ontology Matching: Task Description

- Special requirement: Perform mapping on structural basis only!

  - Breast Cancer diagnosis: no labels, names are widely given based on structure

<table>
<thead>
<tr>
<th>Type</th>
<th>Operatres on</th>
</tr>
</thead>
<tbody>
<tr>
<td>Terminological</td>
<td>strings, labels, comments</td>
</tr>
<tr>
<td>Extensional</td>
<td>(common) instances</td>
</tr>
<tr>
<td>Semantic</td>
<td>logical structure, inference</td>
</tr>
<tr>
<td>Structural</td>
<td>relations, hierarchies</td>
</tr>
</tbody>
</table>
Structural Alignment

- Use of structural graph measures to align ontologies
  - Direct Neighbors
  - Extended Neighbors
  - Depth
  - Centrality
  - Modularity

- Find a mapping for classes that have similar measures
  - Best-First-Search
  - Tabu Search
  - Simulated Annealing
Structural Measures: Example
Experiments

- **Ontology Alignment Evaluation Initiative: Benchmark Dataset**

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>101</td>
<td>Reference ontology. All other ontologies will be aligned against this one.</td>
</tr>
<tr>
<td>201</td>
<td>Same structure as reference alignment but all names are replaced by random strings. The order of class definition is shuffled.</td>
</tr>
<tr>
<td>202</td>
<td>Same as 201 but comments were removed. Again shuffled differently.</td>
</tr>
<tr>
<td>221</td>
<td>Hierarchies are completely removed.</td>
</tr>
<tr>
<td>222</td>
<td>Hierarchy is flattened, some intermediate levels were removed.</td>
</tr>
<tr>
<td>223</td>
<td>Hierarchy is extended, additional levels are introduced.</td>
</tr>
</tbody>
</table>

- **Ontology Alignment Evaluation Initiative: Anatomy Dataset**

<table>
<thead>
<tr>
<th></th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mouse</td>
<td>Structural anatomic description of the mouse</td>
</tr>
<tr>
<td>Human</td>
<td>Structural anatomic description of the human</td>
</tr>
</tbody>
</table>
Evaluation: Benchmark Dataset, Self alignment

- **Single measures**

- **Combined Measures**
Evaluation: Benchmark Dataset, Scrambled Ontologies

- Scrambled Ontologies (201, 202)
Evaluation: Benchmark Dataset, Changed Structures

- Flattened Hierarchies (222)
Evaluation: Benchmark Dataset, Changed Structures

- Extended Hierarchies (223)
Evaluation: Benchmark Dataset, Ordering

- Ordering by centrality can greatly increase performance of the algorithm!
Evaluation: Benchmark Dataset, Tabu Search

- Very few iterations using local search already significantly improve the results
Evaluation: Benchmark Dataset, Simulated Annealing
Evaluation: Anatomy Dataset, Distance Measures

- On Benchmark Dataset: precision and recall quite feasible

- Anatomy Dataset is much larger 3000 classes (Benchmark: 73)

- Matching algorithm performs unsatisfactory

<table>
<thead>
<tr>
<th>Measure</th>
<th>Min</th>
<th>Max</th>
<th>(\mu)</th>
<th>(\sigma)</th>
<th>Min</th>
<th>Max</th>
<th>(\mu)</th>
<th>(\sigma)</th>
<th>(\Sigma)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth</td>
<td>0</td>
<td>15</td>
<td>7.8</td>
<td>4.0</td>
<td>0</td>
<td>21</td>
<td>11.1</td>
<td>5.6</td>
<td>105</td>
</tr>
<tr>
<td>Children</td>
<td>1</td>
<td>15</td>
<td>7.3</td>
<td>3.8</td>
<td>1</td>
<td>21</td>
<td>10.6</td>
<td>5.4</td>
<td>56</td>
</tr>
<tr>
<td>Sinks</td>
<td>3</td>
<td>13</td>
<td>6.3</td>
<td>3.1</td>
<td>3</td>
<td>19</td>
<td>9.9</td>
<td>4.4</td>
<td>40</td>
</tr>
<tr>
<td>Pagerank</td>
<td>0</td>
<td>20</td>
<td>7.8</td>
<td>4.0</td>
<td>0</td>
<td>23</td>
<td>11.0</td>
<td>5.6</td>
<td>105</td>
</tr>
<tr>
<td>Modularity</td>
<td>3</td>
<td>17</td>
<td>8.9</td>
<td>3.9</td>
<td>3</td>
<td>24</td>
<td>12.3</td>
<td>5.0</td>
<td>98</td>
</tr>
<tr>
<td>Combined</td>
<td>0</td>
<td>18</td>
<td>7.1</td>
<td>3.2</td>
<td>0</td>
<td>25</td>
<td>11.2</td>
<td>5.7</td>
<td>290</td>
</tr>
</tbody>
</table>
Evaluation: Anatomy Dataset, Consistency
Conclusion

A Prototype of the Semantic Repository was presented to the researchers and tested with them:

- SharePoint can profit from ontologies using their structure
- Lists can be populated from an ontology definition
- Ontologies are a basis for faceted search
- Ontologies can make long term retrievability possible even by foreign agents
- Workflow was widely accepted for daily use in projects

A structural ontology matching algorithm was proposed

- Based on graph measures
- Results by itself not 100% satisfactory
Future Research

- **Semantic Repository**
  - Better integration with UI and high level techniques offered by SharePoint
  - Gather more testing scenarios and real world experience from researchers

- **Ontology Matching**
  - Extend to other measures from graph theory
  - Combine with other matching algorithms
  - Use structural approach to partition ontology
Final Remarks

- Semantic Web is only slowly evolving, not comparable to WWW
- Semantic Repository can bring the Semantic Web closer to the user
- If researchers make their data retrievable, future generations would be able access it
- Increasing amount of data, information and services raises the need of explicit semantic information
Thanks for your attention! 😊
Questions