An Ontology-driven Document Retrieval Strategy for Organizational Knowledge Management Systems

Carlos Manuel Toledo
Mariel Alejandra Ale
Omar Chiotti
María Rosa Galli

Institute of Development and Design
National Council of Scientific and Technological Research
Avellaneda 3657, Santa Fe, Argentina
Contents

- Organizational knowledge
- Knowledge management
- KM Architecture
- A document annotation/retrieval strategy
- The ontology-driven KM system
- Experimental results
- Conclusion and future work
Organizational knowledge

- Knowledge is currently recognized as a valuable asset for organizations.

- Such significance implies the necessity to ensure its protection by means of:
  - Safeguarding the right access.
  - Persistence over time.
  - Adequate retrieval.
Knowledge management

- Several KM approaches based on **document annotation** and **retrieval strategies** have been implemented in organizations.

- These efforts often fail to manage the natural **heterogeneity** of organisational knowledge sources, and present no adaptation capabilities to add **new knowledge sources**.
A set of social, technological, cultural, political, and economical requirements that a KM model should fulfill.

An organisational memory model.

An ontology-driven KM architecture called Onto-DOM.

- Knowledge **autonomy** principle.
- Knowledge **coordination** principle.
Knowledge Management Architecture

- Complex concepts composed of **adjectives** or more than one **noun**, nor **named entities** (name of cities, persons, localizations, etc.).

  back formal dinning table  
  **vs**  
  mid-century danish table
We propose a new strategy that take in account the **modifiers of nouns** (e.g. adjectives) in order to keep all the semantics of a document.
A document annotation/retrieval strategy

Back formal dinning table

mid-century

Danish

dining

Table

Arnodo

Newstile

Back formal dinning table
A document annotation/retrieval strategy

- **Linguistic analysis**
  - Tokenization
  - Lemmatization
  - POSTagging

- **Concept filtering**
  - Noun\(^+\)
  - Adjective\(^+\) Noun\(^+\)

- **Instance matching**
  - Occurrences between DCs and instances of concepts

- **Concept matching**
  - Occurrences between DCs and the ontological concepts

- **Semantic expansion**
  - Occurrences between DCs and synonyms/hyperonyms of ontological concepts

---

*Mid-century danish* table

*Mid-century danish* desk/furniture

WordNet
A document annotation/retrieval strategy

Ontological Engineer
What hotel has a panoramic view?
The ontology-driven knowledge management system

KReprM: Knowledge representation module
KReTrM: Knowledge retrieval module
OEM: Ontology evolution module
IDCM: Inter-domain communication module
NLPTK: Natural language processing toolkit

Ontological Engineer (OE)

KReprM
- Descriptor selection
- Concept identification
- Concept pattern filter
- Semantic linkage
- Semantic expansion
- Linguistic processing

KReTrM
- Concept identification
- Semantic expansion
- Concept pattern filter
- Linguistic processing

 Protégé API

OWL Domain ontology

WordNet

Unstructured or semi-structured data resources
Stanford NLP toolkit
WordNet Similarity Measures

BDI-Agent
- The agent's goals
- Propose add concepts
- Document processing
- Verify missing nouns

Inter-domain KM agent

Knowledge base

Other domain organizational memories

NLQ

Received documents

Knowledge Consumers

Carlos Manuel Toledo
cmtoledo@santafe-conicet.gov.ar
Experimental results

- A tourism enterprise that includes a domain: **Africa travel**
- An ontology formed by 267 concepts.
- A corpus composed by 125 documents arbitrarily selected from the Internet.
- After that processing the documents for each document the following averages were obtained:
  - 917 words,
  - 269 nouns, and
  - 27 descriptors
Experimental results

Q1: What city offers Unesco world heritage excursions?
Q2: What beach can I take a golf course?
Q3: What hotel has a panoramic view?
Q4: What city has a nightclub?
Q5: What island offers water sport activities?
Q6: Where can I play tennis?
Q7: Where can I do yoga?
Q8: What restaurant can I eat?
Q9: Where can I go on a safari?
Q10: What city has famous gardens?
Q11: Where can I visit Kunene River?
Q12: Where can I see elephants?
Q13: What animals are there in the national park in Africa?
Q14: What rivers are there in Africa?
...

Carlos Manuel Toledo
cmtoledo@santafe-conicet.gov.ar
Experimental results

Evaluation measures:
Mean Average Precision (MAP):

\[ MAP(Q) = \frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} Precision(R_{jk}) \]
### Experimental results

**Precision** ($R_{jk}$)

<table>
<thead>
<tr>
<th>$Q(q_j)$</th>
<th>$d_1$</th>
<th>$d_2$</th>
<th>$d_3$</th>
<th>$d_4$</th>
<th>$d_5$</th>
<th>$d_6$</th>
<th>$d_7$</th>
<th>$d_8$</th>
<th>$d_9$</th>
<th>$d_{10}$</th>
<th>$d_{11}$</th>
<th>$d_{12}$</th>
<th>$d_{13}$</th>
<th>$d_{14}$</th>
<th>$\sum P(R_{jk})$</th>
<th>$\frac{1}{m_j} \sum P(R_{jk})$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$q_1$</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.75</td>
<td>0.8</td>
<td>0.83</td>
<td>0.86</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.91</td>
<td>0.84</td>
</tr>
<tr>
<td>$q_2$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>12</td>
<td>1</td>
</tr>
<tr>
<td>$q_3$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0.86</td>
<td>0.75</td>
<td>0.78</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>8.38</td>
<td>0.93</td>
</tr>
<tr>
<td>$q_4$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$q_5$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$q_6$</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>$q_7$</td>
<td>1</td>
<td>0.5</td>
<td>0.33</td>
<td>0.25</td>
<td>0</td>
<td>0.2</td>
<td>0.33</td>
<td>0.43</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.05</td>
<td>0.44</td>
</tr>
<tr>
<td>$q_8$</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.75</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3.42</td>
<td>0.85</td>
</tr>
<tr>
<td>$q_9$</td>
<td>1</td>
<td>0.5</td>
<td>0.67</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2.17</td>
<td>0.72</td>
</tr>
<tr>
<td>$q_{10}$</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>$q_{11}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$q_{12}$</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.75</td>
<td>0.8</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4.22</td>
<td>0.84</td>
</tr>
<tr>
<td>$q_{13}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>10</td>
<td>1</td>
</tr>
<tr>
<td>$q_{14}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>14</td>
<td>1</td>
</tr>
<tr>
<td>$q_{15}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>9</td>
<td>1</td>
</tr>
<tr>
<td>$q_{16}$</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.75</td>
<td>0.5</td>
<td>0.6</td>
<td>0.5</td>
<td>0.57</td>
<td>0.63</td>
<td>0.56</td>
<td>0.6</td>
<td>0.64</td>
<td>0.58</td>
<td>0.54</td>
<td>0.57</td>
<td>8.95</td>
</tr>
<tr>
<td>$q_{17}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>$q_{18}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
</tr>
<tr>
<td>$q_{19}$</td>
<td>1</td>
<td>1</td>
<td>0.67</td>
<td>0.5</td>
<td>0.4</td>
<td>0.33</td>
<td>0.43</td>
<td>0.5</td>
<td>0.56</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5.38</td>
<td>0.6</td>
</tr>
<tr>
<td>$q_{20}$</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
<td>1</td>
</tr>
</tbody>
</table>

**MAP** ($Q$) = $\frac{1}{|Q|} \sum_{j=1}^{|Q|} \frac{1}{m_j} \sum_{k=1}^{m_j} \text{Precision}(R_{jk})$

$q_j$: query number $j$; $d_k$: document retrieved in position $k$

**MAP** ($Q$) = 0.895
Conclusions

- A straightforward domain-independent approach for ontology-driven document annotation and knowledge retrieval.
- It does not require additional linguistic analysis methods.
- An integration of information retrieval technologies, domain ontologies, and agents into an organizational KM model.
- The result of the test have been highly satisfactory both quantitative results and for the complexity of queries answered.

Future work

- We will develop an extention of an ontology evolution strategy that take into account complex concepts.
Any question?
Thank You!

Carlos Manuel Toledo
Mariel Alejandra Ale
Omar Chiotti
María Rosa Galli

INGAR (CONICET-UTN)
Institute of Development and Design
National Council of Scientific and Technological Research

Santa Fe Research and Technological Center
Argentina

Twitter: cm_toledo
E-mail: cmtoledo@santafe-conicet.gov.ar
A flexible knowledge management framework

Framework class diagram

- Property
- Concept
- Individual
- Ontology
- Jena
- Protégé
- GenericPOSTagger
- GatePOSTagger
- StanfordPOSTagger
- Thesauri
- WordNet
- Descriptor
- Document
- Word
- KMSSystem
- Domain
- KRStrategy
- OntoDOM_KRStrategy
- Another_KRStrategy
- OLStrategy
- SimilarityStrategy
- KSStrategy
- OntoDOM_KSStrategy
- Another_KSStrategy