A proposal for an Interactive Ontology Design Process based on Formal Concept Analysis

Rokia Bendaoud    Amedeo Napoli    Yannick Toussaint

LORIA - CNRS - INRIA - Universités de Nancy
Nancy – France
The KDD process: from heterogeneous textual resources to knowledge systems

Heterogeneous textual resources (thesaurus, documents, DBs)
  ↓ Text processing
    →↓ Preparation / Formatting
    ↑ Prepared data
    ↑ ↓ Data mining step
    ↑ ↓ Symbolic KDD methods: FCA & RCA
    ↑ Discovered patterns
    ↑ ↓ Post-processing of discovered patterns
    ↑ Ontology schema
    ↑ ↓ Interpretation / Evaluation
    ←↓ customizing resources
  Question answering and problem-solving

Bendaoud, Napoli & Toussaint
FCA for an Interactive Ontology Design Process
The objective of the present talk

- **Building and/or refining** an ontology schema, i.e. a set of hierarchically organized concepts, using **text processing** and **text mining** methods applied to a set of heterogeneous textual resources.

- **Representing** the ontology schema within a knowledge representation formalism based on description logics (DLs) to obtain an operational ontology.

- **Interacting** with the experts to customize the ontology schema using a set of simple operations over the resources.
Heterogeneous and available textual resources

1. A thesaurus is composed of a set of classes organized within a hierarchy; upper and partial ontologies may be available too.

2. A set of textual documents includes the description of domain objects based on sets of attributes and relations.

3. A set of databases includes the description of domain objects with set of attributes and relations.
Real-world experiments: Astronomy

Analysis of a set of textual heterogeneous resources holding on “the description of celestial objects” for extracting an ontology schema. The considered resources are:

- **Thesaurus**: the thesaurus of astronomy is a taxonomy of celestial objects (about 3,000 entities)
- **Set of texts**: a collection of texts is composed of 11,591 abstracts selected by domain experts (A&A Journal).
Real-world experiments: Microbiology

Analysis of a set of textual heterogeneous resources holding on “resistance of bacteria to antibiotics” for extracting an ontology schema. The considered resources are:

- **Thesaurus**: the NCBI taxonomy includes 13,380 classes of bacteria.
- **Ontology**: a domain ontology based on expert knowledge represents a set of antibiotics.
- **Set of texts**: a collection of texts is composed of 1244 abstracts selected by domain experts.
Identifying objects

Extracting attributes

- We report the discovery of strong flaring of the object HR 2517.
- NGC 1818 contains as many Be Stars as the slightly younger SMC cluster NGC 330.
- We investigate to what extent non adiabatic temperature variation at the surface of slowly rotating and pulsating beta Cephei and slowly pulsating Be stars affect silicon line profile variations.

Stanford parser

Extraction of pairs (object, property)

- (HR 2517, is_flaring)
- (NGC 1818, is_containing)
- (Be star, is_contained)
- (beta Cephei, is_rotating)
- (beta Cephei, is_pulsating)
- (Be star, is_pulsating)
The relations between objects are extracted using NLP tools

The locus nfxD, which contributes to high-level quinolone resistance in *Escherichia coli*, is only expressed in the presence of gyrA mutation, and maps to the region of the parC and parE genes, was outcrossed into strain KF130, creating strain DH161. DNA sequence analysis of DH161 revealed no changes in the topoisomerase IV parC quinolone resistance-determining region but did identify a single T-to-A mutation in parE at codon 445, leading to a change from Leu to His.

GATE: A General Architecture for Text Engineering

Annotated text

Bendaoud, Napoli & Toussaint  FCA for an Interactive Ontology Design Process
Dealing with attributes describing objects

- The first step consists in building concept hierarchies using FCA: for bacteria and antibiotics.
- FCA is aimed at building a concept lattice from a formal context $K = (G, M, I)$ where $G$ is a set of domain objects, $M$ is a set of attributes, and $I$ is a binary relation defined on $G \times M$.
- Two concept lattices are built:
  - The lattice of antibiotics based on expert knowledge.
  - The lattice of bacteria based on the NCBI taxonomy.
### The lattice of Antibiotics

<table>
<thead>
<tr>
<th></th>
<th>FRB1</th>
<th>FRB3</th>
<th>ARB1</th>
<th>ARB2</th>
<th>HBA5</th>
<th>HBA10</th>
</tr>
</thead>
<tbody>
<tr>
<td>Clarithromycin</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Ciprofloxacin</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cefotaxim</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macrolide</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>
The primary Bacteria concept lattice (first point of view)

Context $K_1 = (G, M_1, I_1)$:
- $G$ is the set of bacteria.
- $M_1$ is the set of bacteria classes in the NCBI taxonomy.
- $I_1$ assigns to an object its class and superclasses in the NCBI taxonomy.

<table>
<thead>
<tr>
<th></th>
<th>Proteobacteria</th>
<th>γ-proteobacteria</th>
<th>Actinobacteria</th>
<th>Bacilli</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helicobacter P.</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klebsiella P.</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycobacterium S.</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Streptococcus P.</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Klebsiella O.</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
The primary Bacteria lattice extracted from NCBI taxonomy
The secondary Bacteria concept lattice (second point of view)

Building a secondary lattice (second point of view) from pairs (object, attribute) given by expert knowledge.

- The formal context $K_2 = (G, M_2, I_2)$ is defined as follows:
  - $G$ is the set of bacteria.
  - $M_2$ is a set of attributes characterizing bacteria.
  - $I_2$ assigns an attribute to an object.

<table>
<thead>
<tr>
<th></th>
<th>spherical</th>
<th>sticks</th>
<th>negative gram</th>
<th>neutral gram</th>
<th>positive gram</th>
<th>aerobic</th>
<th>anaerobic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helicobacter P.</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klebsiella P.</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Mycobacterium S.</td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Streptococcus P.</td>
<td></td>
<td>x</td>
<td></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klebsiella O.</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
The secondary Bacteria concept lattice (continued)
Refining the representation of bacteria

- One set of objects (Bacteria)
- Several kinds of resources
- Two lattices
- Refining and enriching the representation of Bacteria with apposition (lattice extension).

Let $K_1 = (G_1, M_1, I_1)$ and $K_2 = (G_2, M_2, I_2)$ be two contexts with $G_1 = G_2$ and $M_1 \cap M_2 = \emptyset$, then:

$K = K_1|K_2 = (G, M_1 \cup M_2, I_1 \cup I_2)$ is the apposition of $K_1$ and $K_2$. 

### The apposition of $K_1$ and $K_2$

<table>
<thead>
<tr>
<th>Bacteria</th>
<th>attributes</th>
<th>classes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>spherical</td>
<td>Proteobacteria</td>
</tr>
<tr>
<td></td>
<td>sticks</td>
<td>Delta epsilonSubproteobacteria</td>
</tr>
<tr>
<td></td>
<td>negative gram</td>
<td>proteobacteria</td>
</tr>
<tr>
<td></td>
<td>neutral gram</td>
<td>Enterobacteriales</td>
</tr>
<tr>
<td></td>
<td>positive gram</td>
<td>Actinobacteria</td>
</tr>
<tr>
<td></td>
<td>aerobic</td>
<td>Bacilli</td>
</tr>
<tr>
<td>Helicobacter P.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Klebsiella P.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Mycobacterium S.</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>Streptococcus P.</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Klebsiella O.</td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
The lattice associated to the apposition of $K_1$ and $K_2$
Refining the apposition lattice

- Refine the representation of Bacteria (and extending the concept lattice resulting from apposition):
  - by taking into account relations between bacteria and antibiotics (extracted from resources).
  - using the Relational Concept Analysis RCA formalism.
- The RCA process builds an ordered sequence of concept lattices that converges toward a final lattice, where the intension of concepts is composed of attributes and relational attributes.
Relational Concept Analysis (RCA) is an extension of the FCA allowing to design concept lattices taking into account relations between objects.

- The RCA process is based on a pair \((K, R)\) where:
- \(K\) is a set of contexts of the form \(K_i = (G_i, M_i, I_i)\) (a context for each object type)
- \(R\) is a set of relations \(r_k \subseteq G_i \times G_j\), where \(G_i\) is the domain and \(G_j\) the range of \(r_k\).
- Example: \(K = \{\text{Bacteria}, \text{Antibiotics}\}\), \(R = \{\text{Resist}\}\)
The relation Resist between Bacteria and Antibiotics

<table>
<thead>
<tr>
<th>Resist</th>
<th>Clarithromycin</th>
<th>Ciprofloxacin</th>
<th>Cefotaxim</th>
<th>Macrolide</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helicobacter P.</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klebsiella P.</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycobacterium S.</td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Streptococcus P.</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td>Klebsiella O.</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
Motivations and objectives
The FCA and RCA frameworks
Customizing resources
Conclusion and future work

Building concept lattices with FCA
The apposition of two contexts
Introducing RCA
Representation of concepts and reasoning within a KR formalism

The final RCA lattice
An example of concept in the final RCA lattice:

C₁₂ is a concept whose description is:

**Intent:** Bacilli, aerobic, positive gram, sticks,
**Resist:** A₂
**Extent:** Streptococcus P.

C₁₂ is in relation through Resist with concept A₂ whose description is:

**Intent:** FRB1, ARB2, HBA5
**Extent:** Ciprofloxacin
An example of concept definition:

- Intent of $C_{12}$ in the RCA concept lattice:
  \[ \{ \text{Bacilli, aerobic, positive gram, sticks, Resist.A}_2 \} \]
  Extent: \{ Streptococcus P. \}

- $FLE$ concept:
  \[ C_{12} \equiv \text{Bacilli} \sqcap \exists \text{aerobic.} T \sqcap \exists \text{posGram.} T \sqcap \exists \text{sticks.} T \sqcap \exists \text{Resist.A}_2 \]
  \[ C_{12}(\text{Streptococcus P}) \]
The real version of the final bacteria lattice

Ontology schema: 232 concepts
Interacting with experts

The ontology schema contains all the possible concepts that could be built using the resources.

- Experts identify points in the ontology where there is no agreement on the classification of objects:
  - There may be noise in the resources or in the extraction process
  - The expert is not satisfied with the ontology and wants it to be more in accordance with his needs.

- Some classes should be refined, some other should be merged, some other should be created.
Interacting with experts

Problem with concept C205 which intent is:
{sticks, neutral gram, heterotrophic, immobile, Aerobic, Resist.A0, ...}
and its extent is:
{Mycobacterium S., Mycobacterium T.}.

“There is no means for a Neutral Gram attribute. In biology, it is Positive Gram.”

“Could you merge the two attributes?”
Interacting with experts

Motivations and objectives
The FCA and RCA frameworks
Customizing resources
Conclusion and future work

Bendaoud, Napoli & Toussaint
FCA for an Interactive Ontology Design Process
Interacting with experts

positive gram, immobile, spheric, Resist.A2 ...

sticks, neutral gram, heterotrophic, immobile, aerobic, Resist.A0, ...

Bendaoud, Napoli & Toussaint
FCA for an Interactive Ontology Design Process
### Merging Neutral and positive Gram

<table>
<thead>
<tr>
<th></th>
<th>spherical</th>
<th>sticks</th>
<th>negative gram</th>
<th>neutral gram</th>
<th>positive gram</th>
<th>aerobic</th>
<th>anaerobic</th>
</tr>
</thead>
<tbody>
<tr>
<td><em>Helicobacter P.</em></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Klebsiella P.</em></td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
<tr>
<td><em>Mycobacterium S.</em></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Streptococcus P.</em></td>
<td>x</td>
<td></td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><em>Klebsiella O.</em></td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
</tr>
</tbody>
</table>
Merging Neutral and positive Gram

<table>
<thead>
<tr>
<th></th>
<th>spherical</th>
<th>sticks</th>
<th>negative gram</th>
<th>neutral gram</th>
<th>positive gram</th>
<th>aerobic</th>
<th>anaerobic</th>
</tr>
</thead>
<tbody>
<tr>
<td>Helicobacter P.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Klebsiella P.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mycobacterium S.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Streptococcus P.</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
</tr>
<tr>
<td>Klebsiella O.</td>
<td>x</td>
<td>x</td>
<td></td>
<td></td>
<td></td>
<td>x</td>
<td></td>
</tr>
</tbody>
</table>
Back to resources, Neutral and Positive Gram are merged:

- Previous concepts are preserved
- 9 new concepts (new associations of individuals, of intents)
Expert modify resources using simple operations on object descriptions. Operations depend on the type of the resource.

<table>
<thead>
<tr>
<th>Hierarchical description (thesaurus)</th>
<th>Attributes</th>
<th>Relations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Adding a new class</td>
<td>Adding an attribute to an object or to all objects</td>
<td>Adding relations between objects</td>
</tr>
<tr>
<td>Deleting a class</td>
<td>Removing an attribute to an object or to all objects</td>
<td>Deleting a relation between two objects</td>
</tr>
</tbody>
</table>
The set of antibiotics \{\text{Macrolide, Rifampicin, Tetracycline}\} kill bacteria by destroying the DNA but genes.

\{\text{gyrA, gyrB, inhA, parE, rrs}\} have the property of binding DNA.

Thus, the set of bacteria \{\text{Mycobacterium smegmatis, Mycobacterium tuberculosis, Neisseria gonorrhoeae}\} resists the set of antibiotics \{\text{Macrolide, Rifampicin, Tetracycline}\}.
A way of semi-automating the extraction of attributes and relations from heterogeneous and textual Web resources.

The use of Formal Concept Analysis and Relational Concept Analysis to automatically build an ontology schema formally encoded in Description Logic from heterogeneous resources.

Customizing the ontology schema using a set of simple operations over the resources
Future work

Extending the present work:

- **Scaling** could be used to make the ontology more in accordance with experts.
- Keeping a **trace** of all simple operations so that these operations can be performed again on new resources when updating an ontology.