Uncertainty over Intensional Data

Antoine Amarilli

Télécom ParisTech, Paris, France

January 27, 2014
Background

- Lots of raw information on the Web.
- Extract structure from it.
- Integrate various sources.
- Leverage them for complex queries.

⇒ Is there a pizza place open near ENS now?
⇒ Find an affordable place to rent near ENS with $\geq 20 \ m^2$?
⇒ Find a fountain with drinking water near me?
We cannot collect all information:

- Storage space
- Bandwidth
- Access restrictions

Need to access remote data sparingly.

Data management becomes much harder.

- Web crawling
- Web APIs
- Crowdsourcing
- Deep Web
- Expensive processing
- Rule consequences
Structure

- Need to leverage existing structure.
- Structure can be heterogeneous.

⇒ XML/JSON
⇒ Views
⇒ Web graph
⇒ RDF triples
⇒ Relational DBs
⇒ Parse trees
Uncertainty

- Data is imprecise.
- Data is wrong.
- Represent priors on remote data.
- Processing induces uncertainty.

⇒ Fuzzy rules  ⇒ Crowdsourcing  ⇒ Data integration
⇒ NLP  ⇒ Annotations  ⇒ Information extraction
Goals

⇒ To support the **heretogeneous** structure of information.
⇒ To manage **intensional sources** efficiently.
⇒ To maintain **uncertainty** along all steps.
⇒ To scale to **large quantities** of data.
⇒ To decide **relevance** of accesses.
⇒ To answer **expressive queries** through this framework.
⇒ To choose **execution plans** for queries.
Ontology alignment

- Find links between semantic Web sources.
- Iterative alignment (like PageRank)
- Challenges:
  - Support approximate string matching.
  - Align more complex patterns.
  - Improve scalability, parallelize.
  - Better theoretical understanding.

⇒ Marilena Oita (Télécom, IMR), A.A., Pierre Senellart (advisor)
  *Cross-Fertilizing Deep Web Analysis and Ontology Enrichment*
  Very Large Data Search, 2012, Istanbul.
⇒ Coll. Pierre Senellart and Fabian Suchanek (MPI, Télécom).
Probabilistic models

- Uncertain representations for relational databases.
- Uncertain representations for XML trees.

⇒ Are there connections between both models?

<p>| | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>ENS</td>
<td>x</td>
</tr>
<tr>
<td>John</td>
<td>Centrale</td>
<td>¬x</td>
</tr>
<tr>
<td>Jane</td>
<td>ENS</td>
<td>¬x ∧ y</td>
</tr>
<tr>
<td>Jane</td>
<td>Centrale</td>
<td>x ∧ y</td>
</tr>
</tbody>
</table>

⇒ Antoine Amarilli, Pierre Senellart

*Connections btw. Relational and XML Probabilistic Data Models*

Possibility problem for probabilistic XML

- Tree with **probabilistic** nodes:
  - Local choices.
  - Global events.
- Compute **probability** of a tree.
  ⇒ **Complexity** of this problem?

⇒ Antoine Amarilli

*The Possibility Problem for Probabilistic XML*

Query pricing

- Define a **price** for a data collection.
- Sell **partial** subsets with discount.

⇒ How to price user **queries**?
⇒ Does the price **leak** information?
⇒ How to avoid **arbitrage**?
⇒ How to **sample** a priced subset?

⇒ Coll. Pierre Senellart, Ruiming Tang (Nat. Univ. of Singapore)
Crowd data mining

- **Data mining**: find patterns in databases.
- **Frequent itemsets**: common item sets.
- Mine patterns from the crowd.
- **Taxonomy** over the items.
- Find the next question to ask.

⇒ What is the complexity of this problem?

⇒ A.A., Yael Amsterdamer, Tova Milo (Tel Aviv University)

*Complexity of Mining Itemsets from the Crowd Using Taxonomies*

Open-world query answering

- Database of facts.
- Deduction rules.
- Is a query certain?

⇒ When is this decidable?
⇒ Finite or infinite completions?

{Killer(John, Jack)}
Killed(p, q)
⇒ ∃r, Killed(q, r)
Killed(p, x) ∧ Killed(q, x)
⇒ p = q
q : ∃x, Killed(x, John)

⇒ Antoine Amarilli (supervised by Michael Benedikt, Oxford)
Open-World Query Answering Under Number Restrictions
Query answering under uncertain rules

- Database of facts.
- Uncertain deduction rules.
- Reasoning using facts and rules.
- Answer queries with probabilities.

⇒ Learn general tendencies.
⇒ Extrapolate from them.

⇒ Coll. Pierre Bourhis (Oxford, Lille), Pierre Senellart
Provenance for order-aware queries

- Keep link between original database and query results.
- Used for access control, view updates, etc.
- Nice algebraic framework (semirings).

⇒ What about databases with order?

<table>
<thead>
<tr>
<th>John</th>
<th>ENS</th>
<th>$t_1$</th>
<th>ENS</th>
<th>Paris</th>
<th>$s_1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>Mines</td>
<td>$t_2$</td>
<td>Mines</td>
<td>Paris</td>
<td>$s_2$</td>
</tr>
<tr>
<td>John</td>
<td>X</td>
<td>$t_3$</td>
<td>X</td>
<td>Saclay</td>
<td>$s_3$</td>
</tr>
</tbody>
</table>

$q$: “Is John in Paris?”

$t_1 \cdot s_1 \oplus t_2 \cdot s_2$

⇒ A.A., Lamine Ba (Télécom), Daniel Deutch (Tel Aviv), P.S.

Provenance for Nondeterministic Order-Aware Queries

Conclusion

- Funding: *Allocation spécifique* and various grants.
- **DBWeb team**, Télécom ParisTech, 46 rue Barrault.
- Supervised by Prof. **Pierre Senellart**.
- Graduate school **EDITE**.
- **Teaching duties:**
  - *Technologies du Web*, COMASIC master.
  - *Théorie des langages*, Télécom first year.
  - *Entraînement aux concours de programmation*.
- **Collaborations:** Lille, Tel Aviv, Oxford, Singapore.
Conclusion

- **Funding:** Allocation spécifique and various grants.
- **DBWeb team**, Télécom ParisTech, 46 rue Barrault.
- Supervised by Prof. Pierre Senellart.
- Graduate school EDITE.
- **Teaching duties:**
  - *Technologies du Web*, COMASIC master.
  - *Théorie des langages*, Télécom first year.
  - *Entraînement aux concours de programmation*.
- **Collaborations:** Lille, Tel Aviv, Oxford, Singapore.

Thanks for your attention!