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?*

Intensionality

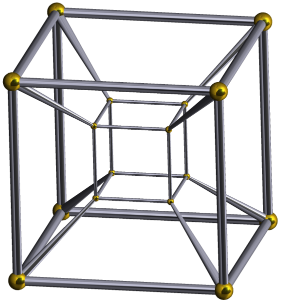
- We cannot collect all information:
  - ⇒ Storage space
  - ⇒ Bandwidth
  - ⇒ Access restrictions
- Need to access remote data sparingly.
- Data management becomes much harder.

⇒ *Web crawling* ⇒ *Crowdsourcing* ⇒ *Expensive processing*

⇒ *Web APIs* ⇒ *Deep Web* ⇒ *Rule consequences*

Structure

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



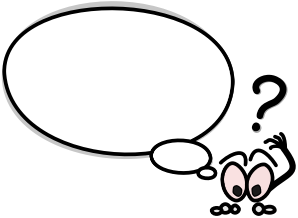
⇒ *XML/JSON*
⇒ *Views*

⇒ *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*

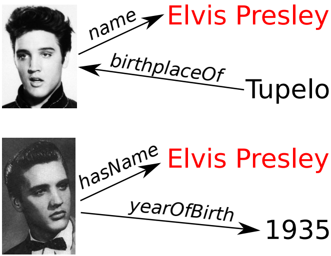
⇒ *NLP* ⇒ *Annotations* ⇒ *Information extraction*

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.

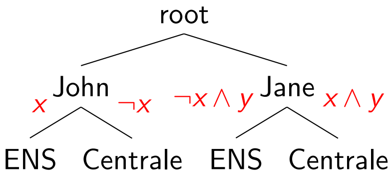
⇒ Coll. Pierre Senellart and Fabian Suchanek (MPI, Télécom).

Probabilistic models

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

$\Rightarrow$ Are there connections between both models?

John	ENS	x
John	Centrale	$\neg x$
Jane	ENS	$\neg x \land y$
Jane	Centrale	$x \land y$

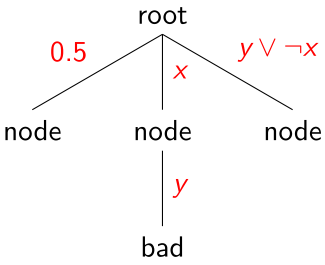


$\Rightarrow$ Antoine Amarilli, Pierre Senellart
Connections btw. Relational and XML Probabilistic Data Models
British National Conference on Databases, 2013, Oxford.

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

Subm. Alberto Mendelzon Intl. Workshop, 2014, Colombia.

Query pricing

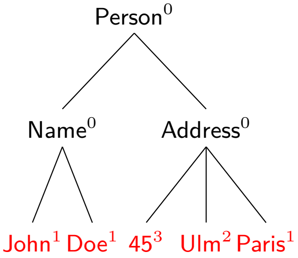
- Define a price for a data collection.
- Sell partial subsets with discount.

⇒ How to price user queries?

⇒ Does the price leak information?

⇒ How to avoid arbitrage?

⇒ How to sample a priced subset?

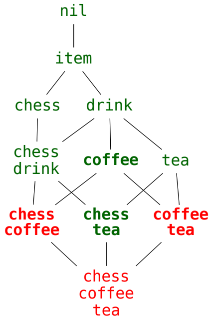


⇒ Coll. Pierre Senellart, Ruiming Tang (Nat. Univ. of Singapore)

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
International Conference on Database Theory, 2014, Athens.

Open-world query answering

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

⇒ When is this decidable?

⇒ Finite or infinite completions?

$\{\text{Killer}(\text{John}, \text{Jack})\}$

$\text{Killed}(p, q) \Rightarrow \exists r, \text{Killed}(q, r)$

$\text{Killed}(p, x) \wedge \text{Killed}(q, x) \Rightarrow p = q$

$q : \exists x, \text{Killed}(x, \text{John})$

⇒ Antoine Amarilli (supervised by Michael Benedikt, Oxford)
Open-World Query Answering Under Number Restrictions
Subm. Principles of Database Systems, 2014, Snowbird, Utah.

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.

$$\{\mathrm{Norm}(\mathrm{John})\}$$

$$\mathrm{Norm}(p) \Rightarrow^{40\%} \mathrm{PhD}(p, \mathrm{ENS})$$

$$\mathrm{Norm}(p) \Rightarrow^{20\%} \mathrm{PhD}(p, \mathrm{X})$$

$$\mathrm{PhD}(p, x) \wedge \mathrm{PhD}(p, y) \Rightarrow^{80\%} x = y$$

$$q(x) : \mathrm{PhD}(x, \mathrm{ENS})$$

⇒ 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?

John	ENS	t_1
John	Mines	t_2
John	X	t_3

ENS	Paris	s_1
Mines	Paris	s_2
X	Saclay	s_3

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
Subm. Principles of Database Systems, 2014, Snowbird, Utah.

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!

Slide 4 image credits: Robert Webb's Stella software, `http://www.software3d.com/Stella.php`