Understanding the Hidden Web

Pierre Senellart

Athens University of Economics & Business, 28 July 2008
Simple problem

Contact all co-authors of Serge Abiteboul.

It’s easy! You just need to:

- Find all co-authors.
- For each of them, find their current email address.
### Advanced Scholar Search

**Find articles**
- with all of the words
- with the **exact phrase**
- with at least one of the words
- **without** the words
- **where my words occur** anywhere in the article

**Author**
Return articles written by
- abiteboul
- e.g., "PJ Hayes" or McCarthy

**Publication**
Return articles published in
- e.g., *J Biol Chem* or *Nature*

**Date**
Return articles published between
- e.g., 1996

**Subject Areas**
- Return articles in all subject areas.
- Return only articles in the following subject areas:
  - Biology, Life Sciences, and Environmental Science
  - Business, Administration, Finance, and Economics
  - Chemistry and Materials Science
  - Engineering, Computer Science, and Mathematics
  - Medicine, Pharmacology, and Veterinary Science
  - Physics, Astronomy, and Planetary Science
  - Social Sciences, Arts, and Humanities

©2007 Google
Different Modules

Conclusion

Introduction

Generic Framework

J Widom
R Hull
J McHugh
V Vianu
S Abiteboul

Abstract: A logic-based language for manipulating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...

Cited by 150 - Related Articles - Web Search - SUDOC Catalogue

Querying documents in object databases - all 3 versions »

S Abiteboul, S Cluet, V Christophides, T Milo, G ... - International Journal on Digital Libraries, 1997 - Springer

Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...

Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity - all 2 versions »


Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...

Cited by 156 - Related Articles - Web Search

[PDF] Change-Centric Management of Versions in an XML Warehouse - all 10 versions »


Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...

Cited by 154 - Related Articles - View as HTML - Web Search

[BOOK] COL: A Logic-Based Language for Complex Objects - all 6 versions »

S Abiteboul, S Grumbach - Springer

Abstract: A logic-based language for manipulating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...

Cited by 150 - Related Articles - Web Search - SUDOC Catalogue
Introduction

General Framework

Different Modules

Conclusion

Advanced Scholar Search

Scholar Preferences

Scholar Help

Scholar

All articles - Recent articles

Results 21 - 30 of about 950 for author:abiteboul. (0.13 seconds)

Querying documents in object databases - all 3 versions »
S Abiteboul, S Cluet, V Christophides, T Milo, G ... - International Journal on Digital Libraries, 1997 - Springer
Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...

Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity - all 2 versions »
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...

Cited by 156 - Related Articles - Web Search

[PDF] Change-Centric Management of Versions in an XML Warehouse - all 10 versions »
Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...

Cited by 154 - Related Articles - View as HTML - Web Search

[BOOK] COL: A Logic-Based Language for Complex Objects - all 6 versions »
S Abiteboul, S Grumbach - Springer
Abstract: A logic-based language for manipulating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...

Cited by 150 - Related Articles - Web Search - SUDOC Catalogue
Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommutateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *

Poster vos questions et commentaires, sauf les demandes de mises à jour qui doivent être effectuées auprès du service informatique de votre site.
Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommutateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *

Poster vos questions et commentaires,
sauf les demandes de mises à jour qui doivent être effectuées auprès du service informatique de votre site.
Annuaire de l’Inria

1 personne(s) trouvée(s).

<table>
<thead>
<tr>
<th>Prénom Nom</th>
<th>Courrier électronique</th>
<th>Téléphone</th>
<th>Site</th>
<th>Projet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sophie Cluet</td>
<td><a href="mailto:Sophie.Cluet@inria.fr">Sophie.Cluet@inria.fr</a></td>
<td></td>
<td>futurs</td>
<td>GEMO</td>
</tr>
</tbody>
</table>

Interroger à nouveau l’annuaire Inria Poster vos questions et commentaires.
Annuaire de l'Inria

1 personne(s) trouvée(s).

<table>
<thead>
<tr>
<th>Prénom Nom</th>
<th>Courrier électronique</th>
<th>Téléphone</th>
<th>Site</th>
<th>Projet</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sophie Cluet</td>
<td><a href="mailto:Sophie.Cluet@inria.fr">Sophie.Cluet@inria.fr</a></td>
<td>futurs</td>
<td>GEMO</td>
<td></td>
</tr>
</tbody>
</table>

Interroger à nouveau l'annuaire Inria Poster vos questions et commentaires.
Introduction

General Framework

Different Modules

Conclusion

abiteboul
S Abiteboul
V Vianu
J McHugh
R Hull
J Widom

Querying documents in object databases  - all 3 versions »
S Abiteboul, S Cluet, V Christophides, T Milo, G ... - International Journal on Digital Libraries, 1997 - Springer
Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...
Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity  - all 2 versions »
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...
Cited by 156 - Related Articles - Web Search

PDF Change-Centric Management of Versions in an XML Warehouse  - all 10 versions »
Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...
Cited by 154 - Related Articles - View as HTML - Web Search

BOOK COL: A Logic-Based Language for Complex Objects  - all 6 versions »
S Abiteboul, S Grumbach - Springer
Abstract: A logic-based language for manipulating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...
Cited by 150 - Related Articles - Web Search - SUDOC Catalogue
Introduction

General Framework

Different Modules

Conclusion

Scholar All articles - Recent articles

Results 21 - 30 of about 950 for author:abiteboul. (0.13 seconds)

Querying documents in object databases - all 3 versions »
S Abiteboul, S Cluet, V Christophides, T Milo, G ... - International Journal on Digital Libraries, 1997 - Springer
Abstract. We consider the problem of storing and accessing documents (SGML and HTML, in particular) using database technology. To specify the database image of documents, we use structuring schemas that consist in grammars annotated ...
Cited by 159 - Related Articles - Web Search

Generic Computation and its complexity - all 2 versions »
Permission to copy without fee all or part of this material is granted provided that the copies are not made or distributed for direct commercial advantage, the ACM copyright notice and the title of the publication and its date appear, ...
Cited by 156 - Related Articles - Web Search

[PDF] Change-Centric Management of Versions in an XML Warehouse - all 10 versions »
Abstract: We present a change-centric method to manage versions in a Web WareHouse of XML data. The starting points is a sequence of snapshots of XML documents we obtain from the web. By running a diff algorithm, we compute ...
Cited by 154 - Related Articles - View as HTML - Web Search

[BOOK] COL: A Logic-Based Language for Complex Objects - all 6 versions »
S Abiteboul, S Grumbach - Springer
Abstract: A logic-based language for manipulating complex objects constructed using set and tuple constructors is introduced. A key feature of the language is the use of base and derived data functions. Under some stratification res- ...
Cited by 150 - Related Articles - Web Search - SUDOC Catalogue
Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommутateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *
Annuaire de l'Inria.

Annuaire (prototype) des personnes travaillant à l'Inria. Les données de cet annuaire sont issues des bases des autocommutateurs de l'Inria.

On peut saisir un ou plusieurs champs. Pour élargir votre requête, préfixez la par une *

Poster vos questions et commentaires,
sauf les demandes de mises à jour qui doivent être effectuées auprès du service informatique de votre site.
Annuaire de l'Inria

0 personne(s) trouvée(s).

Prénom Nom Courrier électronique Téléphone Site Projet

Interroger à nouveau l'annuaire Inria Poster vos questions et commentaires.
Lessons learned

Tedious!

- Identify all relevant services.
- Understand the way to query them.
- Query them one by one.
- Merge the results.

Goal

Have all this work done by the computer, in a fully automatic way.
The hidden Web

Définition (Hidden Web, Deep Web, Invisible Web)

All contents of the Web that are not accessible through the hyperlinked structure of the World Wide Web. Generally speaking: HTML forms, Web services.

Size estimate (2001): 500 more data than on the surface Web.

How to understand and benefit from this content?
Introduction

Understanding the hidden Web

Goal

- **Intensional** indexing of the hidden Web.
- **High-level** queries.
- ⇒ a **semantic** search engine for the hidden Web.

In a fully automatic and unsupervised way!

- **Difficult** and **broad** problem.
- Use of **domain knowledge** (ontology, instances).
- Example with the domain of research publications.
Understanding the hidden Web

Goal

- **Intensional** indexing of the hidden Web.
- **High-level** queries.
- ⇒ a **semantic** search engine for the hidden Web.

In a fully automatic and unsupervised way!

- **Difficult** and **broad** problem.
- Use of **domain knowledge** (ontology, instances).
- Example with the domain of research publications.
1 Introduction

2 General Framework
   - General Process
   - Imprecision
   - Probabilistic XML Data Model
   - Semantic Model

3 Different Modules

4 Conclusion
Process for the semantic interpretation of the hidden Web
Process for the semantic interpretation of the hidden Web

WWW \textit{discovery} \rightarrow \textit{discovery} \rightarrow \text{HTML form}

\text{Web service} \rightarrow \text{discovery} \rightarrow \text{WWW}
Process for the semantic interpretation of the hidden Web

- **WWW**
  - discovery to **HTML form**
  - probbing to **Analyzed form + result pages**

- **Web service**
  - discovery to **HTML form**
Process for the semantic interpretation of the hidden Web

WWW → discovery → HTML form → probing

Web service → discovery → Analyzed form + result pages → induction
Process for the semantic interpretation of the hidden Web

1. Discovery of the WWW
2. Discovery of web service
3. Wrapper induction
4. Semantic analysis
5. Probing of analyzed form + result pages
6. Analyzed web service
Process for the semantic interpretation of the hidden Web

WWW

discovery

Web service

discovery

wrapper induction

analyzed web service

semantic analysis

HTML form

probing

analyzed form + result pages

wrapper induction

service index

indexing

query results
Process for the semantic interpretation of the hidden Web

WWW

Web service

Analyzed web service

HTML form

Analyzed form + result pages

Service index

User query

results

discovery

probing

wrapper

induction

semantic analysis

indexing
Imprecise data and tasks

Observations

- Many needed tasks generate *imprecise* data, with some *confidence* value.
- Need to manage this imprecision, to work with it throughout a complex process.
A probabilistic XML warehouse

Module 1

Module 2

Module 3

Update operation + confidence

Query results + confidence

Update interface

Query interface

Probabilistic XML warehouse
A probabilistic XML warehouse (hidden Web)

Focused crawler → Form analyzer → Info. extractor

Update operation + confidence

Query

Results + confidence

Update interface

Query interface

Probabilistic XML warehouse
A probabilistic XML warehouse (hidden Web)

- Focused crawler
- Form analyzer
- Info. extractor

Crawled URLs + confidence

Query

Results + confidence

Update interface

Query interface

Probabilistic XML warehouse
A probabilistic XML warehouse (hidden Web)

Focused crawler  Form analyzer  Info. extractor

Update operation + confidence  Form URLs?  Results + confidence

Update interface  Query interface

Probabilistic XML warehouse
A probabilistic XML warehouse (hidden Web)
A probabilistic XML warehouse (hidden Web)

- **Focused crawler**
- **Form analyzer**
- **Info. extractor**

**Update interface**

**Query interface**

**Probabilistic XML warehouse**

**Update operation** + confidence

**Query**

Results + confidence

**Analyzed form** + confidence
A probabilistic XML warehouse (hidden Web)

Focused crawler

Form analyzer

Info. extractor

Update interface

Query interface

Probabilistic XML warehouse
A probabilistic XML warehouse (hidden Web)

- Focused crawler
- Form analyzer
- Info. extractor

Update operation + confidence

Query + confidence

Update interface

Query interface

Probabilistic XML warehouse
A probabilistic XML warehouse (hidden Web)

Focused crawler

Form analyzer

Info. extractor

Person → ISBN + confidence

Query

Results + confidence

Update interface

Query interface

Probabilistic XML warehouse
Probabilistic trees

Framework

- **Unordered** data trees.
- Simplifications (without loss of generality): no attributes, no mixed content...

Universe: Set of all such data trees.

Probabilistic tree: Syntactic representation of a discrete probability distribution in this universe.
Probabilistic trees

Framework

- **Unordered** data trees.
- Simplifications (without loss of generality): no attributes, no mixed content...

Universe: Set of all such data trees.

Probabilistic tree: Syntactic representation of a discrete probability distribution in this universe.
Probabilistic trees

**Framework**

- **Unordered** data trees.
- Simplifications (without loss of generality): no attributes, no mixed content...

Universe: Set of all such data trees.

Probabilistic tree: Syntactic representation of a discrete probability distribution in this universe.

(multiset semantics)
Probabilistic tree model

- Data trees with event conditions (conjunction of probabilistic events or their negations) assigned to each node.
- Probabilistic events: boolean random variables, supposed independent, with their own probability distribution.

<table>
<thead>
<tr>
<th>Event</th>
<th>Prob.</th>
</tr>
</thead>
<tbody>
<tr>
<td>$w_1$</td>
<td>0.8</td>
</tr>
<tr>
<td>$w_2$</td>
<td>0.7</td>
</tr>
</tbody>
</table>

\[ p_1 = 0.06 \quad p_2 = 0.70 \quad p_3 = 0.24 \]
Probabilistic tree model

- Data trees with **event conditions** (conjunction of probabilistic events or their negations) **assigned to each node**.

- Probabilistic events: **boolean random variables**, supposed independent, with their own probability distribution.

\[
\begin{array}{c|c}
\text{Event} & \text{Prob.} \\
\hline
\neg w_1 & 0.8 \\
\neg w_2 & 0.7 \\
\end{array}
\]

\[
\begin{array}{c|c}
p_1 & 0.06 \\
p_2 & 0.70 \\
p_3 & 0.24 \\
\end{array}
\]
Features of the probabilistic tree model

- Well founded possible world semantics.
- Complete expressive power, reasonable concision.
- Possibility of applying queries and updates directly on probabilistic trees, in an efficient way.
- Implementation available.

Complexity study

- Tree-pattern queries are PTIME.
- PTIME insertions, but exponential-time deletions.
Conceptual model

- IsA ontology of concepts (simple acyclic graph)

- Typed $n$-ary roles
  - AuthorOf(Publication, Person)
  - HasName(Person, Name)
A service is described by:

- A tuple of **typed** input parameters.
- A **nested** type for its output.
- Semantic relations between inputs and outputs (**Datalog-like description**).
Services and queries

Exemple

Service giving authors from a publication title:

\[ \langle A \rangle \leftarrow \text{AuthorOf}(A,P), \text{HasTitle}(P,T), \text{Input}(T) \]

Exemple

Query:

\[ \langle A, T* \rangle \leftarrow \text{AuthorOf}(A,P), \text{Article}(P), \text{HasTitle}(P,T), \text{KeywordOf}(\text{xml},P) \]
1 Introduction

2 General Framework

3 Different Modules
   - Form analysis
   - Result Page Analysis
   - Semantic Analysis of Services

4 Conclusion
1 Introduction

2 General Framework

3 Different Modules
   - Form analysis
   - Result Page Analysis
   - Semantic Analysis of Services

4 Conclusion
HTML form analysis

Analyze the **structure** of HTML forms.

**Problem**

Map relevant **domain concepts** to each form field.
First step: structural analysis

1. Build a context for each field:
   - label element;
   - id and name attributes;
   - text immediately before the field.

2. Remove stop words, stem.

3. Match this context with concept names, extended with WordNet.

4. Obtain thus candidate annotations.
Second step: confirmation with probing

For each field annotated with concept $c$:

1. Probe the field with a nonsense word to produce an error page.
2. Probe the field with instances of $c$ (chosen representatively of the frequency distribution of $c$).
3. Compare obtained pages with the error page (using a clustering according to the tree structure), to distinguish between error pages and result pages.
4. Confirm the annotation if enough result pages are produced.

Practically speaking, very good precision and good recall; but some limitations on the type of forms that can be dealt with.
Second step: confirmation with probing

For each field annotated with concept $c$:

1. Probe the field with a nonsense word to produce an error page.
2. Probe the field with instances of $c$ (chosen representatively of the frequency distribution of $c$).
3. Compare obtained pages with the error page (using a clustering according to the tree structure), to distinguish between error pages and result pages.
4. Confirm the annotation if enough result pages are produced.

Practically speaking, very good precision and good recall; but some limitations on the type of forms that can be dealt with.
Introduction

General Framework

Different Modules
- Form analysis
- Result Page Analysis
- Semantic Analysis of Services

Conclusion
Query result pages

**Extract** data from Web pages resulting from a query.

<table>
<thead>
<tr>
<th>Showing results 1 through 25 (of 94 total) for all:xml</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. cs.LO/0601085 [abs, ps, pdf, other] :</td>
</tr>
<tr>
<td>Title: A Formal Foundation for ODRL</td>
</tr>
<tr>
<td>Authors: Riccardo Pucella, Vicky Weissman</td>
</tr>
<tr>
<td>Subj-class: Logic in Computer Science, Cryptography and Security</td>
</tr>
<tr>
<td>ACM-class: H.2.7; K.4.4</td>
</tr>
<tr>
<td>2. astro-ph/0512493 [abs, pdf] :</td>
</tr>
<tr>
<td>Title: VOFilter, Bridging Virtual Observatory and Industrial Office Applications</td>
</tr>
<tr>
<td>Authors: Chen-Zhou Cui (1), Markus Dolensky (2), Peter Quin (2), Yong-heng Zhao (1), Françoise Genova (3) (1)NAO China, (2) ESO, (3) CDS</td>
</tr>
<tr>
<td>Comments: Accepted for publication in ChJIA (9 pages, 2 figures, 185KB)</td>
</tr>
<tr>
<td>3. cs.DS/0512061 [abs, ps, pdf, other] :</td>
</tr>
<tr>
<td>Title: Matching Subsequences in Trees</td>
</tr>
<tr>
<td>Authors: Philip Bilke, Inge Li Goertz</td>
</tr>
<tr>
<td>Subj-class: Data Structures and Algorithms</td>
</tr>
<tr>
<td>4. cs.IR/0510025 [abs, ps, pdf, other] :</td>
</tr>
<tr>
<td>Title: Practical Semantic Analysis of Web Sites and Documents</td>
</tr>
<tr>
<td>Authors: Thierry Despeyroux (INRIA Rocquencourt / INRIA Sophia Antipolis)</td>
</tr>
<tr>
<td>Subj-class: Information Retrieval</td>
</tr>
<tr>
<td>5. cs.CR/0510013 [abs, pdf] :</td>
</tr>
<tr>
<td>Title: Safe Data Sharing and Data Dissemination on Smart Devices</td>
</tr>
<tr>
<td>Authors: Luc Bouganim (INRIA Rocquencourt), Cézanne Cremarenc (INRIA Rocquencourt), François Dang Ngoc (INRIA Rocquencourt, PRISM - UVSQ), Nicolas Dies (INRIA Rocquencourt), Philippe Pucherel (INRIA Rocquencourt, PRISM - UVSQ)</td>
</tr>
<tr>
<td>Subj-class: Cryptography and Security; Databases</td>
</tr>
</tbody>
</table>

**Problems**

- **Which part** of the Web page contains the answer?
- **How to extract** structured content?
Unsupervised wrapper induction

- Automatic **pre-annotation** of pages with domain knowledge (finite automata techniques): both **imperfect** and **incomplete**.

- **Use machine learning** to generalize the result in a structural information extractor (conditional random fields techniques)
Experimental results

- 10 publication database services.
- Domain knowledge extracted from DBLP.

<table>
<thead>
<tr>
<th></th>
<th>Title</th>
<th>Author</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(F_g)</td>
<td>(F_x)</td>
<td>(F_g)</td>
</tr>
<tr>
<td>Average</td>
<td>44 63</td>
<td>64 70</td>
<td>85 76</td>
</tr>
</tbody>
</table>

- \(F_g\): \(F\)-measure (in %) of the annotation with domain knowledge.
- \(F_x\): \(F\)-measure (in %) of the annotation by the learned wrapper.
1 Introduction

2 General Framework

3 Different Modules
   - Form analysis
   - Result Page Analysis
   - Semantic Analysis of Services

4 Conclusion
Motivation

Analyze relations between different sources, or between a source and the domain knowledge.

Abstraction

Given two database instances $I$ and $J$ on different schemata, find the optimal description $\Sigma$ of $J$ given $I$ (with $\Sigma$ a finite set of formulas).

What does optimal mean?

- Concision of description.
- Validity of facts predicted by $I$ and $\Sigma$.
- Facts of $J$ explained by $I$ and $\Sigma$. 
Exemple (Tuple-generating dependancies)

\[ R \]

<table>
<thead>
<tr>
<th>( R )</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
</tr>
<tr>
<td>b</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td>d</td>
</tr>
</tbody>
</table>

\[ R' \]

<table>
<thead>
<tr>
<th>( R' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
</tr>
<tr>
<td>b</td>
</tr>
<tr>
<td>c</td>
</tr>
<tr>
<td>a</td>
</tr>
<tr>
<td>d</td>
</tr>
<tr>
<td>d</td>
</tr>
<tr>
<td>g</td>
</tr>
<tr>
<td>h</td>
</tr>
</tbody>
</table>

\[ \Sigma_0 = \emptyset \]

\[ \Sigma_1 = \{ \forall x \ R(x) \rightarrow R'(x, x) \} \]

\[ \Sigma_2 = \{ \forall x \ R(x) \rightarrow \exists y \ R'(x, y) \} \]

\[ \Sigma_3 = \{ \forall x \forall y \ R(x) \land R(y) \rightarrow R'(x, y) \} \]

\[ \Sigma_4 = \{ \exists x \exists y \ R'(x, y) \} \]
Exemple (Tuple-generating dependancies)

\[
\begin{array}{c|c}
R & R' \\
\hline
a & a \\
b & b \\
c & a \\
d & d \\
g & h \\
\end{array}
\]

\[
\Sigma_0 = \emptyset
\]

\[
\Sigma_1 = \{ \forall x \ R(x) \rightarrow R'(x, x) \}
\]

\[
\Sigma_2 = \{ \forall x \ R(x) \rightarrow \exists y \ R'(x, y) \}
\]

\[
\Sigma_3 = \{ \forall x \forall y \ R(x) \land R(y) \rightarrow R'(x, y) \}
\]

\[
\Sigma_4 = \{ \exists x \exists y \ R'(x, y) \}
\]
Exemple (Relevance computation)

\[ R \]
\[
\begin{array}{l}
a \\
b \\
c \\
d
\end{array}
\]

\[ R' \]
\[
\begin{array}{l}
a \\
b \\
c \\
d \\
g \\
h
\end{array}
\]

\[ \forall x \ R(x) \rightarrow R'(x, x) \]

\[ R'_\text{predicted} \]
\[
\begin{array}{l}
a \\
b \\
c \\
d
\end{array}
\]
### Example (Relevance computation)

\[
\begin{array}{c|c}
R & R' \\
\hline
a & a \\
b & b \\
c & a \\
d & d \\
g & h \\
\end{array}
\]

\[
\forall x \ R(x) \land x \neq c \rightarrow R'(x, x)
\]

\[
R'_{\text{predicted}}
\]

\[
\begin{array}{c|c}
& a \\
\hline
a & a \\
b & b \\
d & d \\
\end{array}
\]
Exemple (Relevance computation)

<table>
<thead>
<tr>
<th>R</th>
<th>( R' )</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>g</td>
<td>h</td>
</tr>
</tbody>
</table>

\[ \forall x \ R(x) \land x \neq c \rightarrow R'(x, x) \]

\[ R'(c, a) \]

\[ R'_{\text{predicted}} \]

| a | a |
| b | b |
| c | a |
| d | d |
### Exemple (Relevance computation)

<table>
<thead>
<tr>
<th>$R$</th>
<th>$R'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>g</td>
<td>h</td>
</tr>
</tbody>
</table>

\[ \forall x \ R(x) \land x \neq c \rightarrow R'(x, x) \]

- $R'(c, a)$
- $R'(g, h)$

$R'_{predicted}$

<p>| | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>c</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>g</td>
<td>h</td>
</tr>
</tbody>
</table>
**Exemple (Relevance computation)**

<table>
<thead>
<tr>
<th>R</th>
<th>R'</th>
<th>( R' ) predicted</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
<td>c</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>g</td>
<td>h</td>
<td>g</td>
</tr>
</tbody>
</table>

\[
\forall x \ R(x) \land x \neq c \rightarrow R'(x, x) \\
\exists x \exists y \ R'(x, y) \land x = c \land y = a \\
\exists x \exists y \ R'(x, y) \land x = g \land y = h
\]
Exemple (Relevance computation)

<table>
<thead>
<tr>
<th>$R$</th>
<th>$R'$</th>
</tr>
</thead>
<tbody>
<tr>
<td>a</td>
<td>a</td>
</tr>
<tr>
<td>b</td>
<td>b</td>
</tr>
<tr>
<td>c</td>
<td>a</td>
</tr>
<tr>
<td>d</td>
<td>d</td>
</tr>
<tr>
<td>g</td>
<td>h</td>
</tr>
</tbody>
</table>

$\forall x \ R(x) \land x \neq c \rightarrow R'(x, x)$

$\exists x \exists y \ R'(x, y) \land x = c \land y = a$

$\exists x \exists y \ R'(x, y) \land x = g \land y = h$

Relevance: 17
Results

- Description based on the minimal size of a repair of a formula that is valid and explain all facts of $J$.
- This optimality notion gives “intuitive” results for instances derived from each other with elementary operations.
- Detailed analysis of the algorithmic complexity for various logical languages. High in the polynomial hierarchy (up to $\Pi_4^P$ for the optimality of a tuple-generating dependency!).
- Even for $\forall x_1 \forall x_2 \forall x_3 \ R(x_1, x_2, x_3) \rightarrow R'(x_1)$, relevance computation is already $\text{NP}$-complete.
| 1 | Introduction               |
| 2 | General Framework          |
| 3 | Different Modules          |
| 4 | **Conclusion**             |
|   | • In Summary               |
|   | • Future Work              |
Process for the semantic interpretation of the hidden Web

1. Discovering the WWW
2. Discovering the HTML form
3. Inducing the wrapper
4. Analyzing the form and result pages
5. Semantic analysis
6. Analyzing the web service
7. Indexing the analyzed web service
8. Constructing the service index
9. Querying the service index
10. Returning results
Process for the semantic interpretation of the hidden Web

- WWW
  - discovery
- HTML form
  - probing
- Web service
  - wrapper
    - induction
- Analyzed form + result pages
- Analyzed web service
- Service index
  - indexing
  - query
    - results
  - User

Semantic analysis
Process for the semantic interpretation of the hidden Web

1. **WWW** → **discovery** → **HTML form** → **probing** → **Analyzed form + result pages** → **Service index**

   - **discovery** from WWW
   - **wrapper induction** from Web service
   - **semantic analysis** from Analyzed web service
   - **indexing** from Service index

   - **User** queries the **Service index** and receives **results**
Process for the semantic interpretation of the hidden Web

- WWW
  - discovery
  - Web service
    - wrapper
      - induction
    - semantic analysis
      - Analyzed web service
  - discovery
    - HTML form
      - probing
        - Analyzed form + result pages
          - indexing
            - Service index
              - query
                - results
  - User
A general framework for the automatic understanding of the hidden Web, with especially an unsupervised way to discover the structure of a form and of result pages.

A semi-structured probabilistic model allowing for queries and updates.

A theoretical framework and complexity study for the discovery of schema mappings, solely based on constants appearing in database instances.
Work in progress

- **Integration** of all the components to obtain a complete system for processing the hidden Web.
- Relation between schema mapping discovery and **inductive logic programming**.
- **Answering queries using views** on the semantic model.
Indexing and querying

Given a query, represented as a semantic Web service, how to know which services to query?

Problems

- **Subsumption** of input/output parameters.
- **Missing** input parameters.
- **Composition** of Web services.
Differences with classical databases

Three main differences:

- Information can only be obtained by **views** *(Local As View)*.
- Information **incomplete** and **imprecise**.
- **Nested** types.

Three sources of complexity!

Current research directions: **Magic** sets, Bucket, MiniCon...
Perspectives

- **Top-\(k\)** probabilistic results.
- A machine learning framework adapted to an imperfect annotation, that avoids overfitting (minimal-length description?).
- **Deduplication**, identification of coreferences despite slightly different information.
- **Corroboration** of informations between sources.
- Elaborated semantic analysis, based on an **understanding** of natural language text.
Merci.