



# Challenges in Deep Web Data Extraction

PIERRE SENELLART





# The Deep Web

## Definition (Deep Web, Hidden Web, Invisible Web)

All the content on the Web that is not directly accessible through **hyperlinks**. In particular: HTML forms, Web services.



**Size estimate:** 500 times more content than on the **surface Web**! [BrightPlanet, 2001]. Hundreds of thousands of deep Web databases [Chang et al., 2004]



# Sources of the Deep Web

## Example

- *Yellow Pages* and other directories;
- Library catalogs;
- Weather services;
- US Census Bureau data;
- etc.



# Discovering Knowledge from the Deep Web

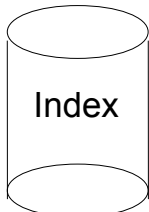
## [Nayak et al., 2012]

- Content of the deep Web hidden to classical Web search engines (they just follow links)
- But very valuable and high quality!
- Even services allowing access through the surface Web (e.g., e-commerce) have more semantics when accessed from the deep Web
- How to **benefit** from this information?
- How to **analyze**, **extract** and **model** this information?

**Focus here:** Automatic, unsupervised, methods, for a given domain of interest

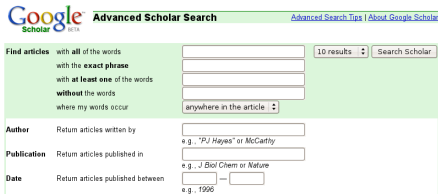
WWW

discovery

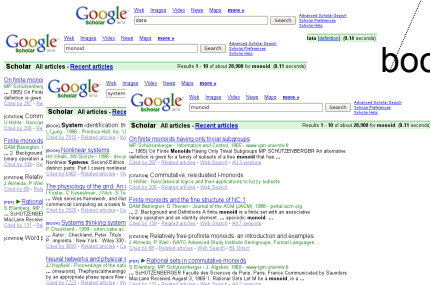


# discovery

## indexing



# siphoning



# bootstrap

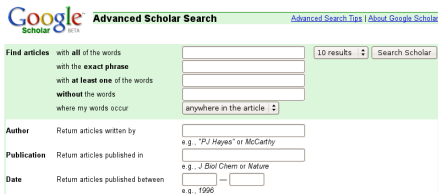
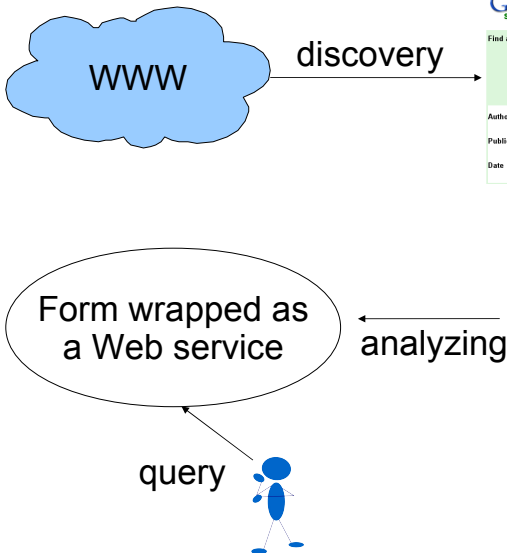
science Word problems and a homological finiteness condition for monoid



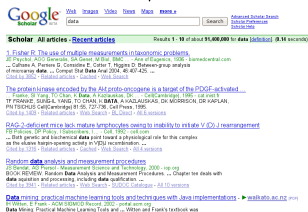
# Notes on the Extensional Approach

- Main issues:
  - Discovering services
  - Choosing appropriate data to submit forms
  - Use of data found in result pages to bootstrap the siphoning process
  - Ensure good coverage of the database
- Approach **favored by Google**, used in production [Madhavan et al., 2006]
- Not always feasible (huge load on Web servers)

# Intensional Approach



## probing





# Notes on the Intensional Approach

- More **ambitious** [Chang et al., 2005, Senellart et al., 2008]
- Main issues:
  - Discovering services
  - Understanding the structure and semantics of a form
  - Understanding the structure and semantics of result pages
  - Semantic analysis of the service as a whole
  - Query rewriting using the services
- No significant load imposed on Web servers





# Outline

Introduction

Analysis of Deep Web Forms

Information Extraction from Deep Web Pages

Modelling Uncertainty in XML

Querying the Deep Web

Conclusion



# Forms

Analyzing the **structure** of HTML forms.

<b>Authors</b>	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
<b>Title</b>	<input type="text"/>		<b>Year</b> <input type="text"/>	<b>Page</b> <input type="text"/>
<b>Conference</b>	<input type="text"/>		<b>ID</b> <input type="text"/>	
<b>Journal</b>	<input type="text"/>	<b>Volume</b> <input type="text"/>	<b>Number</b> <input type="text"/>	
<input type="button" value="Search"/>	<input type="button" value="Reset"/>	Maximum of <input type="text" value="100"/> matches		

## Goal

Associating to each form field the appropriate **domain concept**.



# 1<sup>st</sup> Step: Structural Analysis

1. Build a **context** for each field:
  - label tag;
  - id and name attributes;
  - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



# 1<sup>st</sup> Step: Structural Analysis

1. Build a **context** for each field:
  - label tag;
  - id and name attributes;
  - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



# 1<sup>st</sup> Step: Structural Analysis

1. Build a **context** for each field:
  - label tag;
  - id and name attributes;
  - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



# 1<sup>st</sup> Step: Structural Analysis

1. Build a **context** for each field:
  - label tag;
  - id and name attributes;
  - text immediately before the field.
2. Remove **stop words**, **stem**.
3. **Match** this context with the concept names, extended with WordNet.
4. Obtain in this way **candidate annotations**.



## 2<sup>nd</sup> Step: Confirm Annotations w/ Probing

For each field annotated with a concept  $c$ :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of  $c$  (chosen representatively of the frequency distribution of  $c$ ).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.



## 2<sup>nd</sup> Step: Confirm Annotations w/ Probing

For each field annotated with a concept  $c$ :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of  $c$  (chosen representatively of the frequency distribution of  $c$ ).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.





## 2<sup>nd</sup> Step: Confirm Annotations w/ Probing

For each field annotated with a concept  $c$ :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of  $c$  (chosen representatively of the frequency distribution of  $c$ ).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.



## 2<sup>nd</sup> Step: Confirm Annotations w/ Probing

For each field annotated with a concept  $c$ :

1. Probe the field with nonsense word to get an **error page**.
2. **Probe** the field with instances of  $c$  (chosen representatively of the frequency distribution of  $c$ ).
3. Compare pages obtained by probing with the error page (by clustering along the DOM tree structure of the pages), to distinguish error pages and **result pages**.
4. **Confirm** the annotation if enough result pages are obtained.



## How well does this work?

- **Good results** in practice [Senellart et al., 2008]

	Initial annot.		Confirmed annot.	
	$p(\%)$	$r(\%)$	$p(\%)$	$r(\%)$
Average	49	73	82	73

- Probing raises precision **without hurting recall**
- Clustering according to **DOM paths**: much better than previous approaches
- But some critical assumptions:
  - All form fields are **independent**
  - It is possible to query a field with a **subword**
  - No field is **required**



# How well does this work?

- **Good results** in practice [Senellart et al., 2008]

	Initial annot.		Confirmed annot.	
	$p(\%)$	$r(\%)$	$p(\%)$	$r(\%)$
Average	49	73	82	73

- Probing raises precision **without hurting recall**
- Clustering according to **DOM paths**: much better than previous approaches
- But some critical assumptions:
  - All form fields are **independent**
  - It is possible to query a field with a **subword**
  - No field is **required**



# Better Form Analysis

What

Where

eg. Restaurants  
Hairdressers  
Telstra  
Apple Stores




# Better Form Analysis


What

Find

eg. Restaurants  
Hairdressers  
Telstra  
Apple Stores

 Help us help you  
We need more information to complete your search.

- Please enter a Search Term


 OK




# Better Form Analysis

What

eg. Restaurants  
Hairdressers  
Telstra  
Apple Stores

 Help us help you  
We need more information to complete your search.  
- Please enter a Search Term

 OK

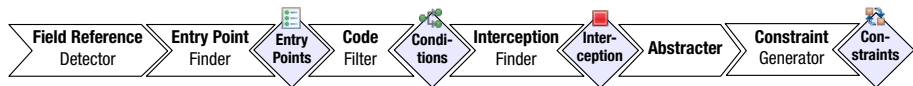
```
// Do not submit unless form is valid
$j("#searchForm").submit(function(event) {
    $j("#searchFormLocationClue").val($j("#searchFormLocationClue").val().trim());
    if ($j("#searchFormBusinessClue").val().isEmpty()) {
        alert('Help us help you\nWe need more information to\ncomplete your search.\n\n- Please enter a Search Term');
        return false;
    } else {
        return true;
    }
});
```



# JavaScript: the Data Language of the Web

- Lots of JavaScript code on the Web (source is always available!)
- Lots of information can be gained by **static analysis** of this code:
  - **Required** fields
  - **Dependencies** between fields (if  $x$  is filled in, so should be  $y$ ; the value of  $x$  should be less than that of  $y$ ; etc.)
  - **Datatype** of each fields (regular expressions, numeric types, dates, etc.)
- Is this feasible in practice?





- **Entry points** are HTML event attributes, setting of event handlers in code, etc. (event: *click* on a submit button, *submit* on a form)
- **Conditions** are (in)equality tests on form field values (possibly aliased)
- **Interceptions** are interruptions of the form submission process (error messages, simple return false; in event handler, etc.)



# Preliminary evaluation

- 70 real-estate websites containing search forms
- 30 out of 70 use client-side validation, with a total of 35 constraints
- **100% precision**: all identified constraints are correct
- **63% recall**: 22 out of 35 JS-enforced constraints were found
- Why did we miss some?
  - Use of complex JavaScript features, such as eval
  - Code obfuscation by introducing extra layers of computation
  - Limitations of the abstracter – work in progress!



# Outline

Introduction

Analysis of Deep Web Forms

Information Extraction from Deep Web Pages

Modelling Uncertainty in XML

Querying the Deep Web

Conclusion



# Result Pages

Pages resulting from a given form submission:

- share the **same structure**;
- set of **records** with fields;
- **unknown** presentation!

Find: remi gilleron Documents Citations

Searching for PHRASE remi gilleron.  
Restrict to: Header Title Order by: Expected citations Title Usage Date Try: Google CiteSeer Google Web Yahoo! MSN CSE DBLP  
7 documents found. Order: number of citations.

PAC Learning under Helpful Distributions - Denis, Gilleron (1997) [Correct]  
(10 citations)  
Helpful Distributions y Francois Denis, Remi Gilleron LRI, URA 369 CNRS, Université de Lille 1 59655  
1 59655 Villeneuve d'Ascq FRANCE e-mail: denis.gilleron@lri.fr Abstract A PAC model under helpful  
on Algorithmic Learning Theory ALT-97 (Denis and Gilleron, 1997) Introduction it seems that many  
ftp.grappa.fr

Sort by	Sort by	Sort by
Title	Title	Year
● 81%	Grindhouse Director Screenwriter Producer	2007
● 78%	Death Proof Director	2007
● 59%	Hostel Executive Producer	2006
● 59%	Reservoir Dogs/Wild Lieutenant Director	2006
● 59%	Inglorious Bastards Director	2006
● 97%	Double Dare Featured	2005
● 78%	Sin City Additional Directing	2005
● 29%	The Muppets: Wizard Of Oz Star	2005
● 0%	Daltry Calhoun Executive Producer	2005
● 85%	Kill Bill Vol. 2 Director Screenwriter	2004
● 100%	2 Channel: A Magnificent Obsession Featured	2004
● 85%	Kill Bill Vol. 1 Director Screenwriter Producer	2003

Test Classif  
Joliet Curio  
Marc Tourn  
www.grappa

## Goal

Building **wrappers** for a given kind of result pages, in a fully automatic, **unsu-**  
**pervised**, way.

**Simplification:** restriction to a domain of interest, with some **domain knowl-**  
**edge**.



# Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for [all:xml](#)

**1. cs.LO/0601085 [abs, ps, pdf, other] :**

Title: A Formal Foundation for ODRL

Authors: [Riccardo Pucella](#), [Vicky Weissman](#)

Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004

Subj-class: Logic in Computer Science: Cryptography and Security

ACM-class: H.2.7; K.4.4

**2. astro-ph/0512493 [abs, pdf] :**

Title: VOFfilter, Bridging Virtual Observatory and Industrial Office Applications

Authors: [Chen-zhou Cui](#) (1), [Markus Dolensky](#) (2), [Peter Quinn](#) (2), [Yong-heng Zhao](#) (1), [Francoise Genova](#) (3) ((1)NAO China, (2) ESO, (3) CDS)

Comments: Accepted for publication in ChJA (9 pages, 2 figures, 185KB)

**3. cs.DS/0512061 [abs, ps, pdf, other] :**

Title: Matching Subsequences in Trees

Authors: [Philip Bille](#), [Inge Li Goertz](#)

Subj-class: Data Structures and Algorithms

**4. cs.IR/0510025 [abs, ps, pdf, other] :**

Title: Practical Semantic Analysis of Web Sites and Documents

Authors: [Thierry Despeyroux](#) (INRIA Rocquencourt / INRIA Sophia Antipolis)

Subj-class: Information Retrieval

**5. cs.CR/0510013 [abs, pdf] :**

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: [Luc Boupanin](#) (INRIA Rocquencourt), [Cosmin Cremarencu](#) (INRIA Rocquencourt), [François Dang Ngoc](#) (INRIA Rocquencourt, PRISM - UVSQ),

[Nicolas Dieu](#) (INRIA Rocquencourt), [Philippe Pucheral](#) (INRIA Rocquencourt, PRISM - UVSQ)

Subj-class: Cryptography and Security: Databases

Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.



# Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for [all:xml](#)

**1. cs.LO/0601085 [abs, ps, pdf, other] :**

Title: A Formal Foundation for ODRL

Authors: [Riccardo Pucella](#), [Vicky Weissman](#)

Comments: 30 pgs, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2004

Subj-class: Logic in Computer Science: Cryptography and Security

ACM-class: H.2.7; K.4.4

**2. astro-ph/0512493 [abs, pdf] :**

Title: VOFfilter, Bridging Virtual Observatory and Industrial Office Applications

Authors: [Chen-zhou Cui](#) (1), [Markus Dolensky](#) (2), [Peter Quinn](#) (2), [Yong-heng Zhao](#) (1), [Francoise Genova](#) (3) ((1)NAO China, (2) ESO, (3) CDS)

Comments: Accepted for publication in ChJA (9 pages, 2 figures, 185KB)

**3. cs.DS/0512061 [abs, ps, pdf, other] :**

Title: Matching Subsequences in Trees

Authors: [Philip Bille](#), [Inge Li Goertz](#)

Subj-class: Data Structures and Algorithms

**4. cs.IR/0510025 [abs, ps, pdf, other] :**

Title: Practical Semantic Analysis of Web Sites and Documents

Authors: [Thierry Despeyroux](#) (INRIA Rocquencourt / INRIA Sophia Antipolis)

Subj-class: Information Retrieval

**5. cs.CR/0510013 [abs, pdf] :**

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: [Luc Bouganim](#) (INRIA Rocquencourt), [Cosmin Cremarencu](#) (INRIA Rocquencourt), [François Dang Ngoc](#) (INRIA Rocquencourt, PRISM - UVSQ),

[Nicolas Dieu](#) (INRIA Rocquencourt), [Philippe Pucheral](#) (INRIA Rocquencourt, PRISM - UVSQ)

Subj-class: Cryptography and Security: Databases

Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.



# Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for **all:xml**

1. **cs.LO/0601085** [abs, ps, pdf, other] :

Title: A Formal Foundation for ODRL

Authors: **Riccardo Baccus**, **Vicky Weissman**

Comments: 30 pp, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security), 2006

Subj-class: **Logic in Computer Science**: Cryptography and Security

ACM-class: H.2.7; K.4.4

2. **astro-ph/0512493** [abs, pdf] :

Title: VOFiler, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Chen-zhou Cui** (1), **Markus Dolensky** (2), **Peter Quinn** (2), **Yong-heng Zhao** (1), **Francoise Genov** (3) ((1)NAO China, (2) **ESO**, (3) CDS)

Comments: Accepted for publication in CHJAA (9 pages, 2 figures, 185KB)

3. **cs.DS/0512061** [abs, ps, pdf, other] :

Title: Matching Subsequences in Trees

Authors: **Philip Bille**, **Inge Li Goertz**

Subj-class: **Data Structures and Algorithms**

4. **cs.IR/0510025** [abs, ps, pdf, other] :

Title: Practical Semantic Analysis of Web Sites and Documents

Authors: **Jierry Despreux** (**IRISA**, **IRISA-Fontaine**, **IRISA**, **IRISA**), **Philippe Bichard** (**IRISA**, **IRISA**)

Subj-class: **Information Retrieval**

5. **cs.CR/0510013** [abs, pdf] :

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: **Luc Bouganim** (**IRISA**, **IRISA**), **Cosmin Creangă** (**IRISA**, **IRISA**), **François Dang Ngoc** (**IRISA**, **IRISA**), **PRISM - UVSQ**, **Nicolas Béd** (**IRISA**, **IRISA**), **Philippe Bichard** (**IRISA**, **IRISA**), **PRISM - UVSQ**

Subj-class: Cryptography and Security; Databases

Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.



# Annotation by domain knowledge

Showing results 1 through 25 (of 94 total) for **all:xml**

**1. cs.LO/0601085 [abs, ps, pdf, other] :**

Title: A Formal Foundation for ODRL

Authors: **Bernardo Cossio**, **Vicky Weissman**

Comments: 30 pp, preliminary version presented at WITS-04 (Workshop on Issues in the Theory of Security) 2006

Subj-class: **Logic in Computer Science**: Cryptography and Security

ACM-class: H.2.7; K.4.4

**2. astro-ph/0512493 [abs, pdf] :**

Title: VOFilter, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Chen-zhou Cui** (1), **Mariusz Dolensky** (2), **Peter Quinn** (2), **Yong-heng Zhao** (1), **Francoise Genov** (3) ((1)NAO China, (2) **ESO**, (3) CDS)

Comments: Accepted for publication in ChJAA (9 pages, 2 figures, 185KB)

**3. cs.DS/0512061 [abs, ps, pdf, other] :**

Title: Matching Subsequences in Trees

Authors: **Philip Bille**, **Inge Li Goertz**

Subj-class: **Data Structures and Algorithms**

**4. cs.IR/0510025 [abs, ps, pdf, other] :**

Title: Practical Semantic Analysis of Web Sites and Documents

Authors: **Jierrry Despreux** (**Paris Lodron Universität Salzburg**), **Stéphane Aubin**

Subj-class: **Information Retrieval**

**5. cs.CR/0510013 [abs, pdf] :**

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: **Luc Bouganim** (**Paris Lodron Universität Salzburg**), **Cosmin Creangă** (**Paris Lodron Universität Salzburg**), **François Dang Ngoc** (**Paris Lodron Universität Salzburg**), **PRISM - UVSQ**,

**Nicolas Béd** (**Paris Lodron Universität Salzburg**), **Philippe Buchera** (**Paris Lodron Universität Salzburg**), **PRISM - UVSQ**

Subj-class: Cryptography and Security; Databases

Automatic **pre-annotation** with domain knowledge (gazetteer):

- Entity recognizers for dates, person names, etc.
- Titles of articles, conference names, etc.: those that are in the knowledge base.

Both **incomplete** and **imprecise**!

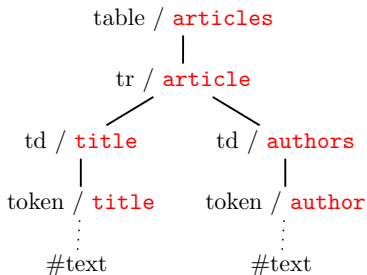






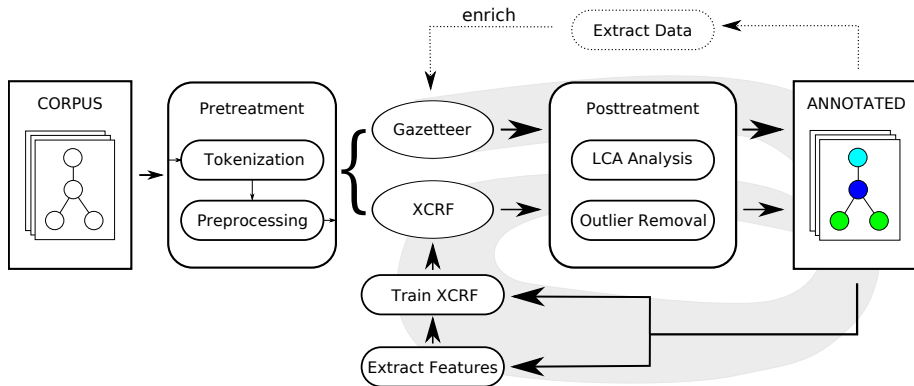
# Unsupervised Wrapper Induction

- Use the pre-annotation as the input of a structural supervised machine learning process.
- Purpose: remove outliers, generalize incomplete annotations.





# Architecture





# How well does this work?

- Good, but not great, results [Senellart et al., 2008]

	Title		Author		Date	
	$F_g$	$F_x$	$F_g$	$F_x$	$F_g$	$F_x$
Average	44	63	64	70	85	76

- $F_g$ :  $F$ -measure (%) of the annotation by the gazetteer.
  - $F_x$ :  $F$ -measure (%) of the annotation by the induced wrapper.
- Main issue: the machine learning assumes that the initial annotation is really the reference one



# Handling Uncertainty

- The outcome of an annotation process, of machine learning, is inherently **imprecise**
- Even more so for conditional random fields: we get **probabilities** that an item is given an annotation
- **Issue:** usually, these confidence scores, probabilities, etc., are **disregarded** and just used for ranking or top- $k$  selection
- **What we would like:** to deal with these scores in a **rigorous** manner, and keep them **throughout** a long process
- Web data is usually loosely structured and tree shaped  $\Rightarrow$  **XML-like**



# Handling Uncertainty

- The outcome of an annotation process, of machine learning, is inherently **imprecise**
- Even more so for conditional random fields: we get **probabilities** that an item is given an annotation
- **Issue:** usually, these confidence scores, probabilities, etc., are **disregarded** and just used for ranking or top- $k$  selection
- **What we would like:** to deal with these scores in a **rigorous** manner, and keep them **throughout** a long process
- Web data is usually loosely structured and tree shaped  $\Rightarrow$  **XML**-like



# Outline

Introduction

Analysis of Deep Web Forms

Information Extraction from Deep Web Pages

Modelling Uncertainty in XML

Querying the Deep Web

Conclusion



# Uncertain data

Numerous sources of **uncertain data**:

- Measurement errors
- Data integration from contradicting sources
- Imprecise mappings between heterogeneous schemata
- Imprecise automatic process (information extraction, natural language processing, etc.)
- Imperfect human judgment



# Managing this imprecision

## Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

Especially:

- Use **probabilities** to represent the confidence in the data
- Query data and retrieve **probabilistic** results
- Allow adding, deleting, modifying data in a **probabilistic** way





# Managing this imprecision

## Objective

Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

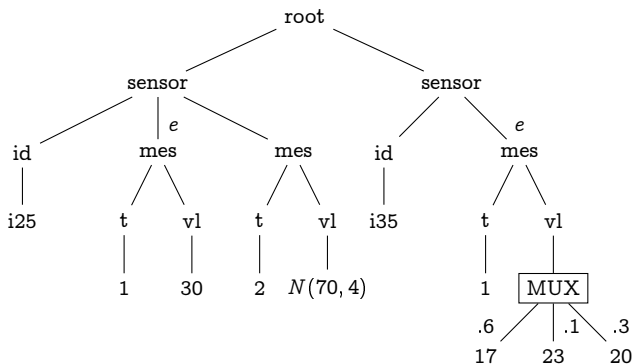
Especially:

- Use **probabilities** to represent the confidence in the data
- Query data and retrieve **probabilistic** results
- Allow adding, deleting, modifying data in a **probabilistic** way



# A General Probabilistic XML Model

[Abiteboul et al., 2009, Kimelfeld and Senellart 2013]



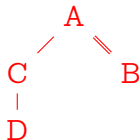
- $e$ : event “it did not rain” at time 1
- MUX: mutually exclusive options
- $N(70, 4)$ : normal distribution

- Compact representation of a **set of possible worlds**
- Two kinds of dependencies: global ( $e$ ) and local (MUX)
- Generalizes **existing models** of the literature



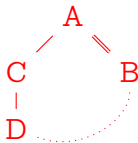
# Query languages on trees

Tree-pattern queries (TP) `/A[C/D]//B`



Tree-pattern queries with joins (TPJ) for `$x` in `$doc/A/C/D`

`return $doc/A//B[.= $x]`



Monadic second-order queries (MSO) generalization of TP, do not cover TPJ unless the size of the alphabet is bounded

But also: **updates** (insertion, deletions), **aggregate queries** (count, sum, max, avg. . .)



# Querying probabilistic XML

Semantics of a (Boolean) query = **probability**:

1. Generate **all possible worlds** of a given probabilistic document
2. In each world, **evaluate the query**
3. **Add up** the probabilities of the worlds that make the query true

**EXPTIME** algorithm! Can we do better, i.e., can we apply directly the algorithm on the probabilistic document?

We shall talk about **data complexity** of query answering.



# Querying probabilistic XML

Semantics of a (Boolean) query = **probability**:

1. Generate **all possible worlds** of a given probabilistic document (possibly exponentially many)
2. In each world, **evaluate the query**
3. **Add up** the probabilities of the worlds that make the query true

**EXPTIME** algorithm! Can we do better, i.e., can we apply directly the algorithm on the probabilistic document?

We shall talk about **data complexity** of query answering.



# Complexity of Query Evaluation

## ■ Boolean queries:

	Local dependencies	Global dependencies
TP	$\text{PTIME}$ [Kimelfeld et al., 2009]	$\text{FP}^{\#P}$ -complete
TPJ	$\text{FP}^{\#P}$ -complete	$\text{FP}^{\#P}$ -complete
MSO	$\text{PTIME}$ [Cohen et al., 2009]	$\text{FP}^{\#P}$ -complete

- Aggregate queries: (somewhat) tractable on local dependencies when the aggregate function is a **monoid** function; **continuous distributions** do not add complexity [Abiteboul et al., 2010]
- Not the same kind of updates are tractable for local and global dependencies [Kharlamov et al., 2010]



# Outline

Introduction

Analysis of Deep Web Forms

Information Extraction from Deep Web Pages

Modelling Uncertainty in XML

Querying the Deep Web

Conclusion



Basic query language with recursion.

$$ReachGood() \leftarrow Start(x), Reach(x, y), Good(y)$$
$$Reach(x, y) \leftarrow Reach(x, z), Reach(z, y)$$
$$Reach(x, y) \leftarrow G(x, y)$$

- Rules consisting of **Horn clauses**.
- Heads of rules are **intensional** predicates.
- Other predicates are **extensional** (input) predicates.
- Distinguished **goal** predicate.

Given an instance of the input predicates, computes the goal predicate using a least fixed point semantics.

**Monadic Datalog (MDL)** = all intensional predicates are unary.





Basic query language with recursion.

$$ReachGood() \leftarrow Start(x), Reach(x, y), Good(y)$$
$$Reach(x, y) \leftarrow Reach(x, z), Reach(z, y)$$
$$Reach(x, y) \leftarrow G(x, y)$$

- Rules consisting of **Horn clauses**.
- Heads of rules are **intensional** predicates.
- Other predicates are **extensional** (input) predicates.
- Distinguished **goal** predicate.

Given an instance of the input predicates, computes the goal predicate using a least fixed point semantics.

**Monadic Datalog (MDL)** = all intensional predicates are unary.



Basic query language with recursion.

$$ReachGood() \leftarrow Start(x), Reach(x, y), Good(y)$$
$$Reach(x, y) \leftarrow Reach(x, z), Reach(z, y)$$
$$Reach(x, y) \leftarrow G(x, y)$$

- Rules consisting of **Horn clauses**.
- Heads of rules are **intensional** predicates.
- Other predicates are **extensional** (input) predicates.
- Distinguished **goal** predicate.

Given an instance of the input predicates, computes the goal predicate using a least fixed point semantics.

**Monadic Datalog (MDL)** = all intensional predicates are unary.



$ReachGood() \leftarrow Start(x), Reach(x, y), Good(y)$

$Reach(x, y) \leftarrow Reach(x, z), Reach(z, y)$

$Reach(x, y) \leftarrow G(x, y)$

DL query, not MDL

$ReachGood() \leftarrow Reachable(x), Good(x)$

$Reachable(y) \leftarrow G(x, y), Reachable(x)$

$Reachable(x) \leftarrow Start(x)$

(Equivalent) MDL query



# Containment of Datalog

$Q \subseteq Q'$  iff for every input instance  $D$ ,  $Q(D) \subseteq Q'(D)$

One can use containment to decide equivalence, giving natural way to optimize recursive queries.

Bad news [Shmueli, 1987]

Datalog containment and equivalence are **undecidable**

But important special cases known to be decidable, e.g., MDL containment is in 2EXPTIME [Cosmadakis et al., 1988].



# Containment of Datalog

$Q \subseteq Q'$  iff for every input instance  $D$ ,  $Q(D) \subseteq Q'(D)$

One can use containment to decide equivalence, giving natural way to optimize recursive queries.

Bad news [Shmueli, 1987]

Datalog containment and equivalence are **undecidable**

But important special cases known to be decidable, e.g., MDL containment is in 2EXPTIME [Cosmadakis et al., 1988].



# MDL containment and Restricted Interfaces

## Restricted Access Scenario

We have a relational schema with relations  $R_1 \dots R_n$ .

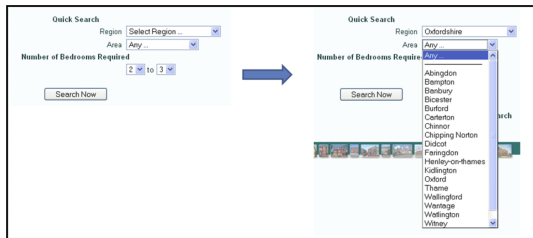
Each  $R_i$  has some arity  $ar_i$  and is additionally **restricted** in that access is only via a set of **access methods**  $m_1 \dots m_{n_i}$ . An access method has a set of “input positions”  $S \subseteq \{1 \dots ar_i\}$  that require known values.

An **access** to method  $m_i$  is a binding of the input positions of  $m_i$ , which returns an output.

Given an instance  $I$  for the schema, a set of initial constants  $C_0$  the access patterns define a collection of **valid access paths**: sequences of accesses  $ac_1 \dots ac_k$  and responses such that each value in the binding to  $ac_i$  is either in  $C_0$  or is an output of  $ac_j$  with  $j < i$ . Facts that are returned by valid paths are the **accessible data**.



# Access Methods



Method [ApartmentFind](#):

[Region](#), [Area](#), [NumBeds](#) → [Address](#), [Price](#), [Description](#), [Link](#)

Above the input fields have enum domains – but in general the domains can be infinite (e.g., textboxes). Querying over limited interfaces arises in many other data management settings: web services, legacy database managers.



# Equivalence with Access Patterns

Given two conjunctive queries  $Q$ ,  $Q'$  and a schema with access patterns, determine whether  $Q$  and  $Q'$  agree on the accessible data. Similarly  $Q$  is contained in  $Q'$  relative to the access patterns if whenever  $Q$  is true on the accessible data, then so is  $Q'$ .

## Question

What is the complexity of query equivalence, containment under access patterns?

Containment can be used to solve a number of other static analysis questions about limited access schemas, such as whether an access is **relevant** to a query. [Benedikt et al., 2011]





# Equivalence with Access Patterns

Given two conjunctive queries  $Q$ ,  $Q'$  and a schema with access patterns, determine whether  $Q$  and  $Q'$  agree on the accessible data. Similarly  $Q$  is contained in  $Q'$  relative to the access patterns if whenever  $Q$  is true on the accessible data, then so is  $Q'$ .

## Question

What is the complexity of query equivalence, containment under access patterns?

Containment can be used to solve a number of other static analysis questions about limited access schemas, such as whether an access is **relevant** to a query. [Benedikt et al., 2011]



# Limited Access Containment and MDL

## [Li and Chang, 2001]

### Axiomatizing accessibility

$$\begin{aligned} \text{Accessible}(x_j) &\leftarrow (R(\vec{x}) \wedge \bigwedge_{i \in \text{input}(m)} \text{Accessible}(x_i)) \\ \text{Accessible}(c) &\leftarrow \end{aligned}$$

$c$  a constant or value in some enum datatype of the schema.

An MDL program that computes the accessible values: those obtainable via a valid access path.

$\Rightarrow$  For any UCQ query  $Q$  one can write an MDL query  $Q_{acc}$  that computes the value of  $Q$  restricting to accessible values.

$Q$  contained in  $Q'$  under access patterns  $\Leftrightarrow$

$Q_{acc}$  contained in  $Q'$  on all databases.

Containment of a Monadic Datalog Query in a UCQ!



# Limited Access Containment and MDL

## [Li and Chang, 2001]

### Axiomatizing accessibility

$$\begin{aligned} \text{Accessible}(x_j) &\leftarrow (R(\vec{x}) \wedge \bigwedge_{i \in \text{input}(m)} \text{Accessible}(x_i)) \\ \text{Accessible}(c) &\leftarrow \end{aligned}$$

$c$  a constant or value in some enum datatype of the schema.

An MDL program that computes the accessible values: those obtainable via a valid access path.

$\Rightarrow$  For any UCQ query  $Q$  one can write an MDL query  $Q_{acc}$  that computes the value of  $Q$  restricting to accessible values.

$Q$  contained in  $Q'$  under access patterns  $\Leftrightarrow$

$Q_{acc}$  contained in  $Q'$  on all databases.

Containment of a Monadic Datalog Query in a UCQ!



## (Formerly) Open Questions

- Is the **2EXPTIME** bound on UCQ containment tight?  
Only known lower-bound was **PSPACE**.  
Yes, the bound is tight. [Benedikt et al., 2012a]
- What about containment under limited access patterns?  
Only obvious lower bound of **NP**; **coNEXPTIME** upper bound  
proved for special cases [Calì and Martinenghi, 2008]  
**coNEXPTIME**-complete [Benedikt et al., 2011, 2012b]

The use of these bounds to get practical query rewriting algorithms is largely open.



## (Formerly) Open Questions

- Is the **2EXPTIME** bound on UCQ containment tight?  
Only known lower-bound was **PSPACE**.  
Yes, the bound is tight. [Benedikt et al., 2012a]
- What about containment under limited access patterns?  
Only obvious lower bound of **NP**; **coNEXPTIME** upper bound  
proved for special cases [Calì and Martinenghi, 2008]  
**coNEXPTIME-complete** [Benedikt et al., 2011, 2012b]

The use of these bounds to get practical query rewriting algorithms is largely open.



## (Formerly) Open Questions

- Is the **2EXPTIME** bound on UCQ containment tight?  
Only known lower-bound was **PSPACE**.  
Yes, the bound is tight. [Benedikt et al., 2012a]
- What about containment under limited access patterns?  
Only obvious lower bound of **NP**; **coNEXPTIME** upper bound  
proved for special cases [Calì and Martinenghi, 2008]  
**coNEXPTIME-complete** [Benedikt et al., 2011, 2012b]

The use of these bounds to get practical query rewriting algorithms is largely open.



## (Formerly) Open Questions

- Is the  $2EXPTIME$  bound on UCQ containment tight?  
Only known lower-bound was  $PSPACE$ .  
Yes, the bound is tight. [Benedikt et al., 2012a]
- What about containment under limited access patterns?  
Only obvious lower bound of  $NP$ ;  $coNEXPTIME$  upper bound  
proved for special cases [Calì and Martinenghi, 2008]  
 $coNEXPTIME$ -complete [Benedikt et al., 2011, 2012b]

The use of these bounds to get practical query rewriting algorithms is largely open.



# Outline

Introduction

Analysis of Deep Web Forms

Information Extraction from Deep Web Pages

Modelling Uncertainty in XML

Querying the Deep Web

Conclusion





Exploiting deep Web data in a rigorous manner requires combining techniques:

- Information retrieval
- Information extraction
- Machine learning
- Database systems
- Database theory
- Static analysis



Exploiting deep Web data in a rigorous manner requires combining techniques:

- Information retrieval
- Information extraction
- Machine learning
- Database systems
- Database theory
- Static analysis

Help is most welcome!



Merci.



# Outline

Complements

References



# Conditional Random Fields

- Generalization of hidden Markov Models [Lafferty et al., 2001]
- Probabilistic **discriminative** model: models the probability of an annotation **given an observable** (different from **generative** models)
- **Graphical model**: every annotation can depends on the neighboring annotations (as well as the observable); dependencies measured through (boolean or integer) **feature functions**.
- Features are automatically assigned a weight and combined to find the **most probable annotation** given the observable.



# Conditional Random Fields for XML (XCRF)

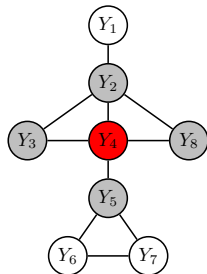
[Gilleron et al., 2006]

**Observables:** various structural and content-based features of nodes (tag names, tag names of ancestors, type of textual content...).

**Annotations:** domain concepts assigned to nodes of the tree.

**Tree** probabilistic model:

- models **dependencies** between annotations;
- conditional independence: annotations of nodes only depend on their **neighbors** (and on observables).



Most **discriminative** features selected.



# Why Probabilistic XML?

- Extensive literature about probabilistic relational databases [Dalvi et al., 2009, Widom, 2005, Koch, 2009]
- Different typical querying languages: conjunctive queries vs tree-pattern queries (possibly with joins)
- Cases where a tree-like model might be appropriate:
  - No schema or few constraints on the schema
  - Independent modules **annotating** freely a content warehouse
  - Inherently tree-like data (e.g., mailing lists, parse trees) with naturally occurring queries involving the descendant axis

## Remark

Some results can be transferred from one model to the other. In other cases, connection much trickier!



# Outline

Complements

References



Serge Abiteboul, Benny Kimelfeld, Yehoshua Sagiv, and Pierre Senellart. On the expressiveness of probabilistic XML models. *VLDB Journal*, 18(5):1041–1064, October 2009.

Serge Abiteboul, T-H. Hubert Chan, Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Aggregate queries for discrete and continuous probabilistic xml. In *Proc. ICDT*, Lausanne, Switzerland, March 2010.

Michael Benedikt, Georg Gottlob, and Pierre Senellart. Determining relevance of accesses at runtime. In *Proc. PODS*, pages 211–222, Athens, Greece, June 2011.

Michael Benedikt, Pierre Bourhis, and Pierre Senellart. Monadic datalog containment. In *Proc. ICALP*, pages 79–91, Warwick, United Kingdom, July 2012a.

- Michael Benedikt, Tim Furche, Andreas Savvides, and Pierre Senellart. ProFoUnd: Program-analysis-based form understanding. In *Proc. WWW*, pages 313–316, Lyon, France, April 2012b. Demonstration.
- BrightPlanet. The deep Web: Surfacing hidden value. White Paper, July 2001.
- Andrea Cali and Davide Martinenghi. Conjunctive query containment under access limitations. In *ER*, 2008.
- Kevin Chen-Chuan Chang, Bin He, Chengkai Li, Mitesh Patel, and Zhen Zhang. Structured databases on the Web: Observations and implications. *SIGMOD Record*, 33(3):61–70, September 2004.
- Kevin Chen-Chuan Chang, Bin He, and Zhen Zhang. Toward large scale integration: Building a metaquerier over databases on the Web. In *Proc. CIDR*, Asilomar, USA, January 2005.

Sara Cohen, Benny Kimelfeld, and Yehoshua Sagiv. Running tree automata on probabilistic XML. In *Proc. PODS*, Providence, RI, USA, June 2009.

Stavros S. Cosmadakis, Haim Gaifman, Paris C. Kanellakis, and Moshe Y. Vardi. Decidable optimization problems for database logic programs. In *STOC*, 1988.

Nilesh Dalvi, Christopher Ré, and Dan Suciu. Probabilistic databases: Diamonds in the dirt. *Communications of the ACM*, 52(7), 2009.

Rémi Gilleron, Patrick Marty, Marc Tommasi, and Fabien Torre. Interactive tuples extraction from semi-structured data. In *Proc. Web Intelligence*, Hong Kong, China, December 2006.

Evgeny Kharlamov, Werner Nutt, and Pierre Senellart. Updating probabilistic XML. In *Proc. Updates in XML*, Lausanne, Switzerland, March 2010.

Benny Kimelfeld and Pierre Senellart. Probabilistic XML: Models and complexity. In Zongmin Ma and Li Yan, editors, *Advances in Probabilistic Databases for Uncertain Information Management*, pages 39–66. Springer-Verlag, May 2013.

Benny Kimelfeld, Yuri Kosharovsky, and Yehoshua Sagiv. Query evaluation over probabilistic XML. *VLDB Journal*, 18(5): 1117–1140, October 2009.

Christoph Koch. MayBMS: A system for managing large uncertain and probabilistic databases. In Charu Aggarwal, editor, *Managing and Mining Uncertain Data*. Springer-Verlag, 2009.

John Lafferty, Andrew McCallum, and Fernando Pereira. Conditional Random Fields: Probabilistic models for segmenting and labeling sequence data. In *Proc. ICML*, Williamstown, USA, June 2001.

Chen Li and Edward Chang. Answering queries with useful bindings. *TODS*, 26(3):313–343, 2001.

Jayant Madhavan, Alon Y. Halevy, Shirley Cohen, Xin Dong, Shawn R. Jeffery, David Ko, and Cong Yu. Structured data meets the Web: A few observations. *IEEE Data Engineering Bulletin*, 29(4):19–26, December 2006.

Richi Nayak, Pierre Senellart, Fabian M. Suchanek, and Aparna Varde. Discovering interesting information with advances in Web technology. *SIGKDD Explorations*, 14(2), December 2012.

Pierre Senellart, Avin Mittal, Daniel Muschick, Rémi Gilleron, and Marc Tommasi. Automatic wrapper induction from hidden-Web sources with domain knowledge. In *Proc. WIDM*, pages 9–16, Napa, USA, October 2008.

Oded Shmueli. Decidability and Expressiveness Aspects of Logic Queries. In *PODS*, pages 237–249, 1987.

Jennifer Widom. Trio: A system for integrated management of data, accuracy, and lineage. In *Proc. CIDR*, Asilomar, CA, USA, January 2005.