Deep Web Data

Analysis, Extraction, and Modelling

PIERRE SENELLART

Microsoft Research, 26 April 2010

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 [Varde et al., 2009]

- 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

Outline

Outline

Forms

Analyzing the structure of HTML forms.

Authors				
Title			**Year**	**Page**
Conference			**ID**	
Journal			**Volume**	**Number**
Search	Reset	Maximum of 100 matches		

Goal

Associating to each form field the appropriate domain concept.

Outline

1st 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.

1st 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.

1st 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.

1st 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.

2nd 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.

2nd 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.

2nd 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.

2nd 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.

Outline

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?

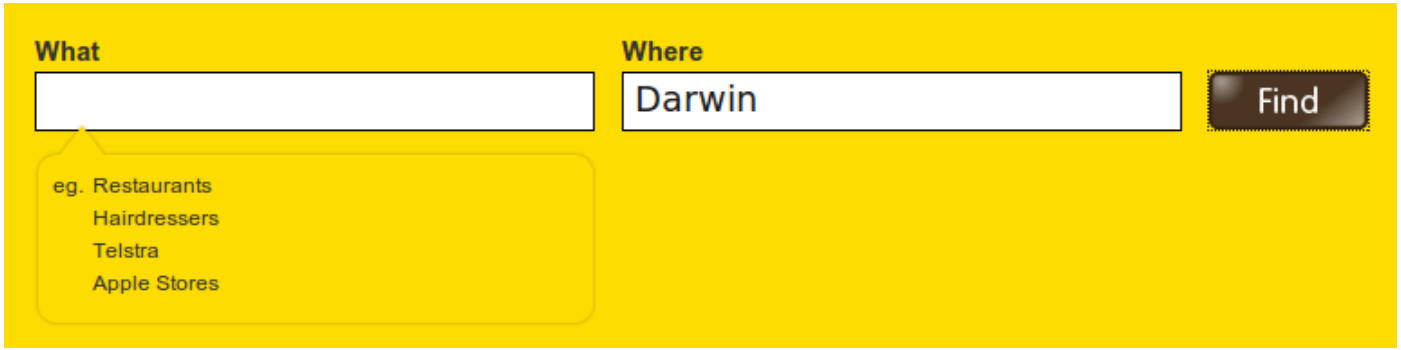
- Good results in practice [Senellart et al., 2008]

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

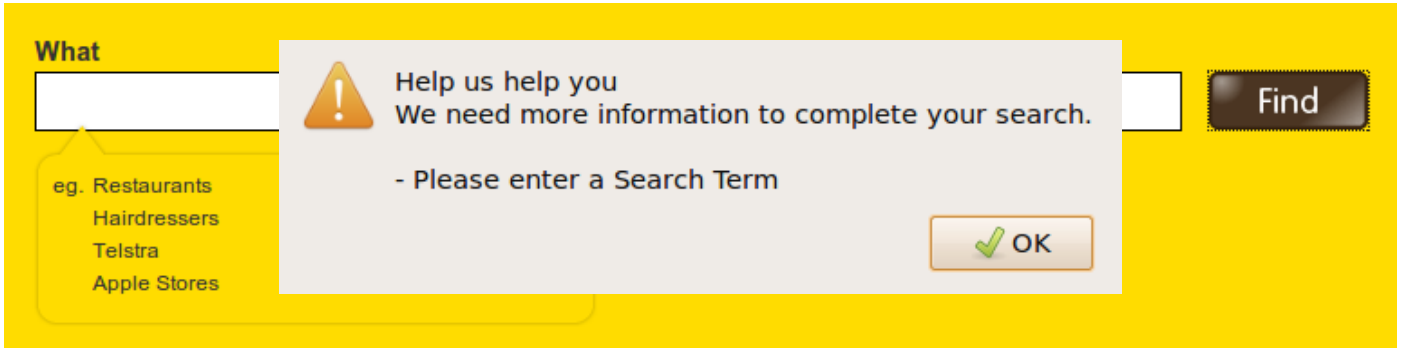
- Probing raises precision without hurting recall
- Clustering according to DOM paths: much better than previous approaches
- But some critical assumptions:
  - All form fields are independent
  - It is possible to query a field with a subword
  - No field is required

Outline

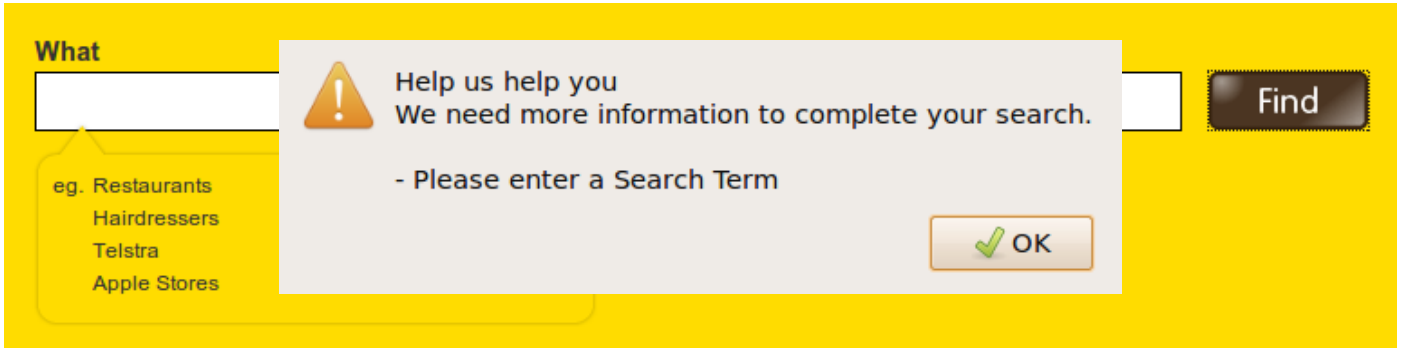
How to Do Better?



How to Do Better?



How to Do Better?




```
// 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
      complete 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:
  - Form understanding
    - Required fields
    - Dependencies between fields
    - Datatype of each fields (regular expressions, numeric types, dates, etc.)
  - Security testing [Guha et al., 2009]
  - AJAX application crawling [Mesbah et al., 2008]
  - ...
- Tools? Methods? Coping with frameworks?

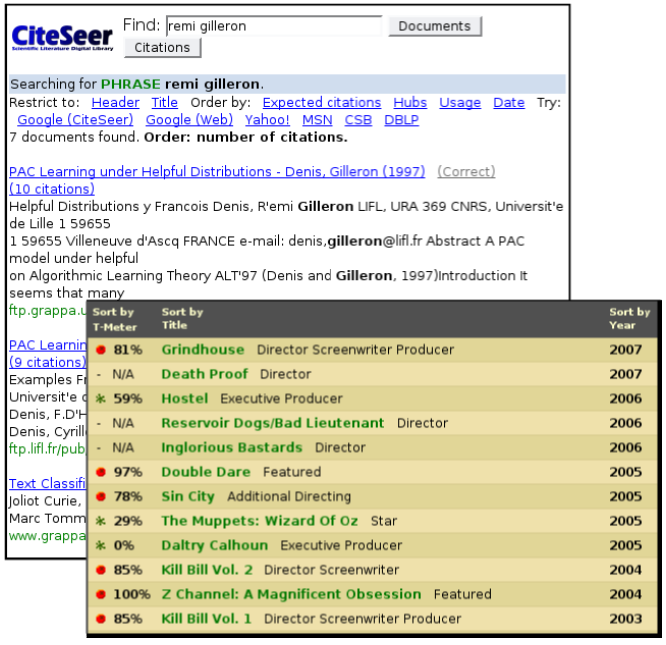
Outline

Outline

Result Pages

Pages resulting from a given form submission:

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



Goal

Building wrappers for a given kind of result pages, in a fully automatic, unsupervised, way.

Simplification: restriction to a domain of interest, with some domain knowledge.

Outline

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: **VOFilter, 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 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: 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 Cremarenco (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: **VOFilter, 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 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: 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 Cremarenco (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: **VOFilter, 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 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: 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 Cremarenco (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: **VOFilter, 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 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: 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 Cremarenco (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

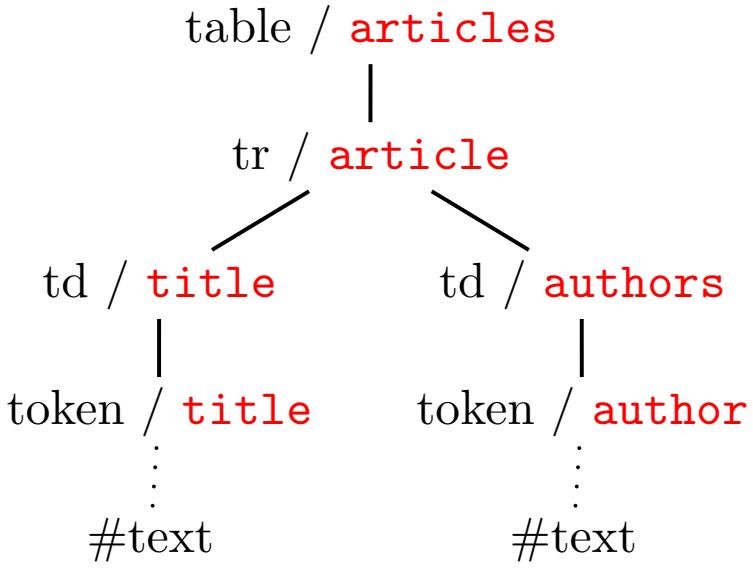
Automatic pre-annotation with domain knowledge (gazetteer):

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

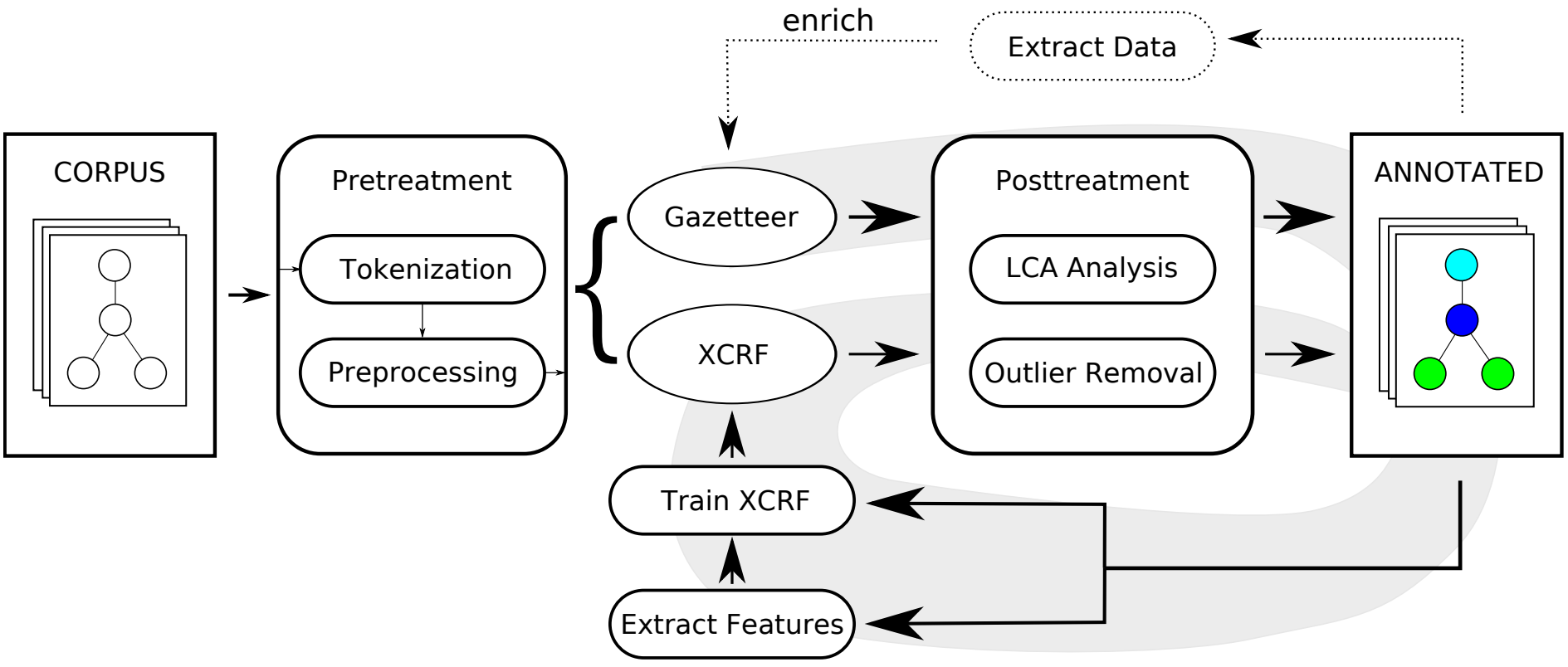
Both incomplete and imprecise!

Unsupervised Wrapper Induction

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



Architecture



Outline

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

Outline

How to Do Better?

- A machine learning approach always look for a trade-off between:
  - fitting the data
  - having a concise model
- But trade-off never resolved in a generic, parameter-less, way
- Idea: minimum-length description principle. The best model is the one that, together with the description of what is not covered by the model, the most concise.
- But people have tried for decades to do Kolmogorov-complexity–based learning!
- Idea: (very) restricted languages for describing the model and repairs, e.g., the formalism of data exchange [Fagin et al., 2003]
- Initial theoretical work in that direction [Gottlob and Senellart, 2010], much more to do!

How to Do Better?

- A machine learning approach always look for a trade-off between:
  - fitting the data
  - having a concise model
- But trade-off never resolved in a generic, parameter-less, way
- Idea: minimum-length description principle. The best model is the one that, together with the description of what is not covered by the model, the most concise.
- But people have tried for decades to do Kolmogorov-complexity–based learning!
- Idea: (very) restricted languages for describing the model and repairs, e.g., the formalism of data exchange [Fagin et al., 2003]
- Initial theoretical work in that direction [Gottlob and Senellart, 2010], much more to do!

How to Do Better?

- A machine learning approach always look for a trade-off between:
  - fitting the data
  - having a concise model
- But trade-off never resolved in a generic, parameter-less, way
- Idea: minimum-length description principle. The best model is the one that, together with the description of what is not covered by the model, the most concise.
- But people have tried for decades to do Kolmogorov-complexity–based learning!
- Idea: (very) restricted languages for describing the model and repairs, e.g., the formalism of data exchange [Fagin et al., 2003]
- Initial theoretical work in that direction [Gottlob and Senellart, 2010], much more to do!

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

Outline

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

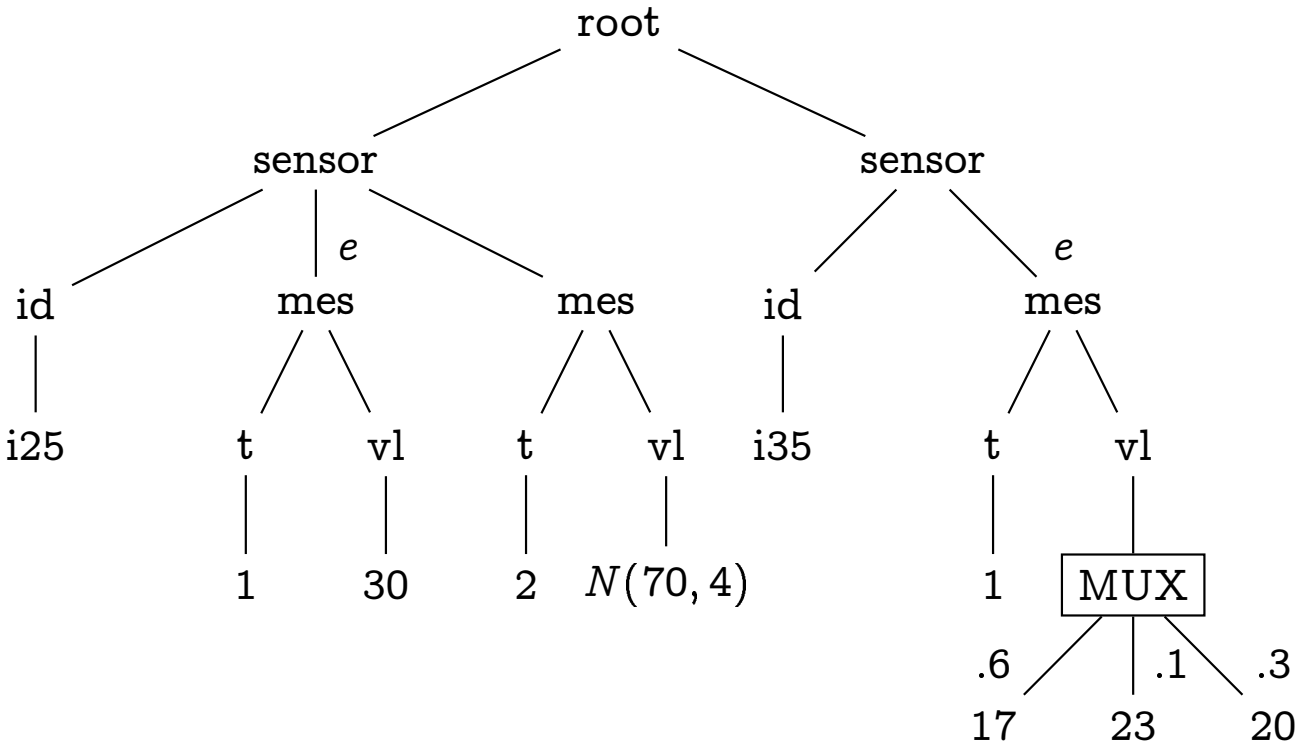
Not to pretend this imprecision does not exist, and manage it as rigorously as possible throughout a long, automatic and human, potentially complex, process.

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

Outline

A General Probabilistic XML Model [Abiteboul et al., 2009]

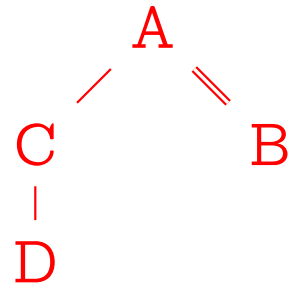


- e: event "it did not rain" at time 1
- MUX: mutually exclusive options
- $N(70, 4)$: normal distribution

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

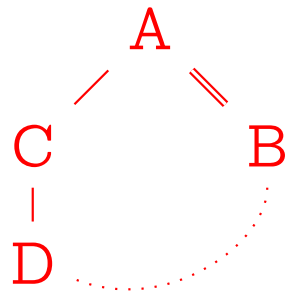
Query languages on trees

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.

Outline

Complexity of Query Evaluation

- Boolean queries:

	Local dependencies	Global dependencies
TP	PTIME [Kimelfeld et al., 2009]	$\mathrm{FP}^{\#\mathrm{P}}$-complete
TPJ	$\mathrm{FP}^{\#\mathrm{P}}$-complete	$\mathrm{FP}^{\#\mathrm{P}}$-complete
MSO	PTIME [Cohen et al., 2009]	$\mathrm{FP}^{\#\mathrm{P}}$-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., 2010b]
- Not the same kind of updates are tractable for local and global dependencies [Kharlamov et al., 2010]

Outline

Probabilistic XML vs. XML Schema

- A probabilistic XML model with local and global dependencies represents a finite (weighted) set of bounded possible trees
- An XML schema represents an infinite (unweighted) set of unbounded possible trees

What's the connection?

Tree Automata are Everywhere

- A MSO query is a tree automaton [Thomas, 1997]
- Schemas on trees (à la XML Schema) are tree automata
- The local dependencies model, and infinite, unbounded, extensions thereof (derived from recursive Markov chains [Etessami and Yannakakis, 2009]), are probabilistic tree automata [Benedikt et al., 2010]

Computing the probability of a MSO query over the set of worlds of a probabilistic XML model constrained by a schema is just a matter of composing and intersecting tree automata!

. . . at least theoretically speaking

Tree Automata are Everywhere

- A MSO query is a tree automaton [Thomas, 1997]
- Schemas on trees (à la XML Schema) are tree automata
- The local dependencies model, and infinite, unbounded, extensions thereof (derived from recursive Markov chains [Etessami and Yannakakis, 2009]), are probabilistic tree automata [Benedikt et al., 2010]

Computing the probability of a MSO query over the set of worlds of a probabilistic XML model constrained by a schema is just a matter of composing and intersecting tree automata!

. . . at least theoretically speaking

So, What Reamins to Be Done?

- Practical algorithms for:
  - query evaluation over recursive Markov chains
  - schema validation
  - sampling constrained by a schema
- Verification issues:
  - Potential effect of an update?
  - Is the result of an update still valid against a schema?
  - Same questions, but with probability thresholds

Outline

In Brief

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

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

In Brief

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

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

This is not a one-man job ☺.

Other Recent Works

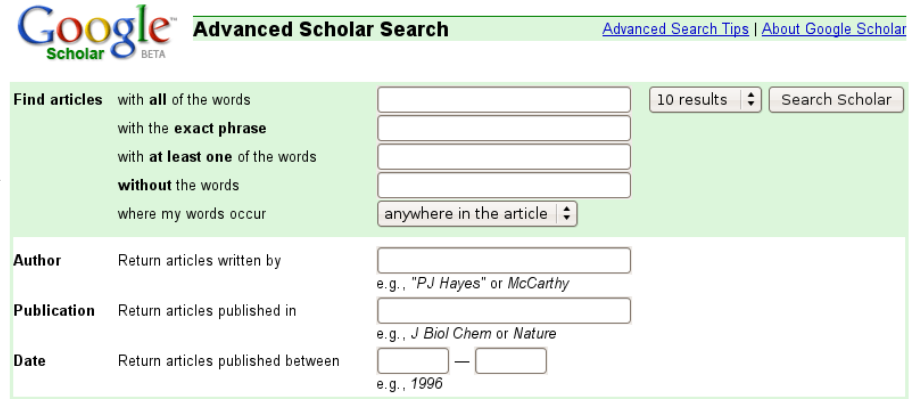
- Determining the truthfulness of facts and the trustworthiness of sources on the Web via fixpoint methods [Galland et al., 2010]
- A general query language for social networks [Abiteboul et al., 2010a] (work in progress)
- Archiving of a living Web; freshness and consistency of crawls [Spaniol et al., 2009]
- PageRank prediction with hidden Markov models [Vazirgiannis et al., 2008]
- Generalization of Kleinberg's HITS algorithm to the comparison of two arbitrary graphs; application to synonym extraction in a dictionary [Senellart and Blondel, 2008]
- Finding similar nodes in a network with fixpoint computation of Green measures [Ollivier and Senellart, 2007]

Merci.

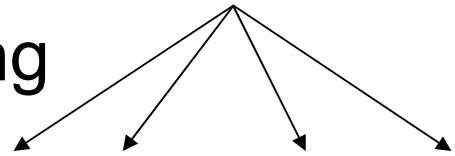
Outline

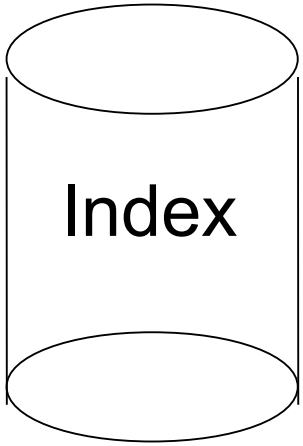
Extensional Approach

discovery

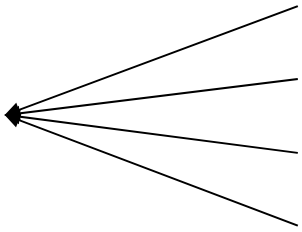


siphoning





indexing



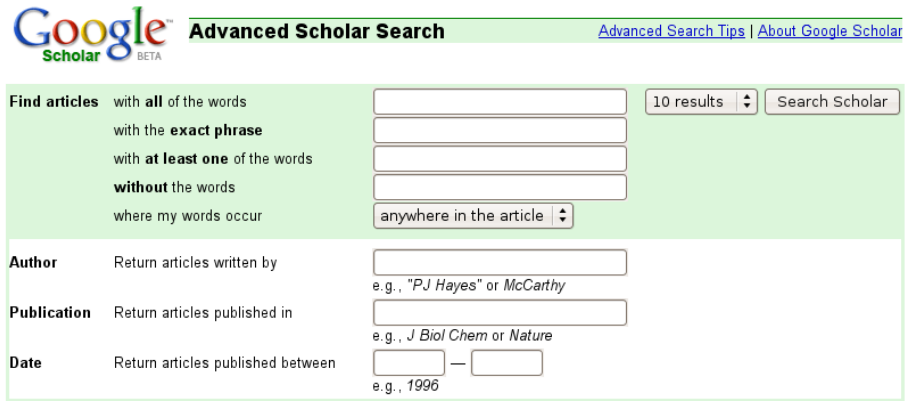
bootstrap

Notes on the Extensional Approach

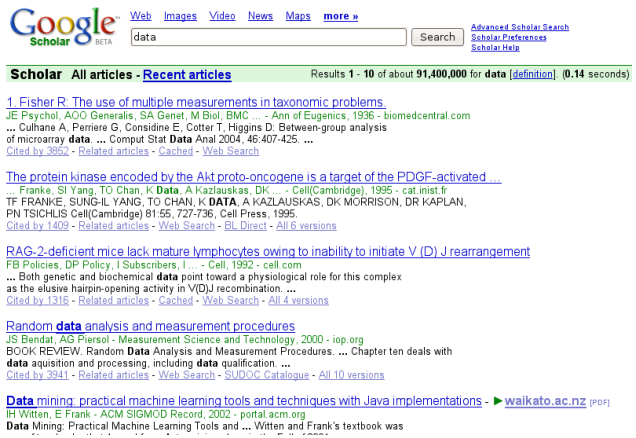
- Main issues:
  - Discovering services
  - Choosing appropriate data to submit forms
  - Use of data found in result pages to bootstrap the siphoning process
  - Ensure good coverage of the database
- Approach favored by Google, used in production [Madhavan et al., 2006]
- Not always feasible (huge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

Intensional Approach

discovery



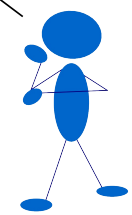
probing



analyzing

Form wrapped as a Web service

query



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
- No significant load imposed on Web servers

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.

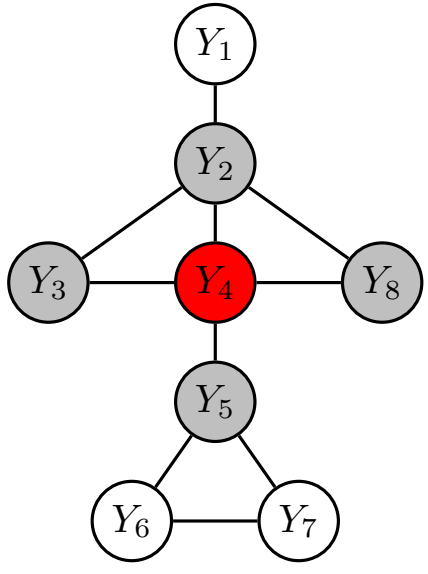
Conditional Random Fields for XML (XCRF) [Gilleron et al., 2006]

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

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, Sihem Amer-Yahia, Amélie Marian, Alban Galland, and Pierre Senellart. Birds of a tag flock together. In *Proc. SSM*, New York, USA, February 2010a. Workshop without formal proceedings.

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 2010b.

Michael Benedikt, Evgeny Kharlamov, Dan Olteanu, and Pierre Senellart. Probabilistic XML via Markov chains, March 2010. Preprint.

BrightPlanet. The deep Web: Surfacing hidden value. White Paper, July 2001.

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.

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

Kousha Etessami and Mihalis Yannakakis. Recursive Markov chains, stochastic grammars, and monotone systems of nonlinear equations. *JACM*, 56(1), 2009. ISSN 0004-5411. doi: http://doi.acm.org/10.1145/1462153.1462154.

Ronald Fagin, Phokion G. Kolaitis, Renée J. Miller, and Lucian Popa. Data exchange: Semantics and query answering. In *Proc. ICDT*, Siena, Italy, January 2003.

Alban Galland, Serge Abiteboul, Amélie Marian, and Pierre Senellart. Corroborating information from disagreeing views. In *Proc. WSDM*, pages 1041–1064, New York, USA, February 2010.

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.

Georg Gottlob and Pierre Senellart. Schema mapping discovery from data instances. *Journal of the ACM*, 57(2), January 2010.

Arjun Guha, Shriram Krishnamurthi, and Trevor Jim. Using static analysis for ajax intrusion detection. In *Proc. WWW*, Madrid, Spain, 2009.

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

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.

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.

Ali Mesbah, Engin Bozdag, and Arie van Deursen. Crawling ajax by inferring user interface state changes. In *Proc. ICWE*, 2008.

Yann Ollivier and Pierre Senellart. Finding related pages using Green measures: An illustration with Wikipedia. In *Proc. AAAI*, pages 1427–1433, Vancouver, Canada, July 2007.

Pierre Senellart and Vincent D. Blondel. Automatic discovery of similar words. In Michael W. Berry and Malu Castellanos, editors, *Survey of Text Mining II: Clustering, Classification and Retrieval*, pages 25–44. Springer-Verlag, January 2008.

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.

Marc Spaniol, Dimitar Denev, Arturas Mazeika, Pierre Senellart, and Gerhard Weikum. Data quality in Web archiving. In *Proc. WICOW*, pages 19–26, Madrid, Spain, April 2009.

Wolfgang Thomas. Languages, automata, and logic. In G. Rozenberg and A. Salomaa, editors, *Handbook of Formal Languages*. Springer-Verlag, 1997.

Aparna Varde, Fabian M. Suchanek, Richi Nayak, and Pierre Senellart. Knowledge discovery over the deep Web, semantic Web and XML. In *Proc. DASFAA*, pages 784–788, Brisbane, Australia, April 2009. Tutorial.

Michalis Vazirgiannis, Dimitris Drosos, Pierre Senellart, and Akrivi Vlachou. Web page rank prediction with Markov models. In *Proc. WWW*, pages 1075–1076, Beijing, China, April 2008. Poster.

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