Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge

P. Senellart[1,2] A. Mittal[3] D. Muschick[6] R. Gilleron[4,5] M. Tommasi[4,5]

1 TELECOM ParisTech
2 INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA centre de recherche SACLAY - ÎLE-DE-FRANCE
3 INDIAN INSTITUTE OF TECHNOLOGY BOMBAY
4 Lille 3 UNIVERSITE CHARLES-DE-GAULLE
5 INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET EN AUTOMATIQUE INRIA centre de recherche LILLE - NORD EUROPE
6 TU Graz Graz University of Technology

University of Oxford, 5 May 2009

The Hidden Web

Definition (Hidden Web, Deep 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!

Hundreds of thousands of hidden Web databases.

Sources of the Hidden Web

Example

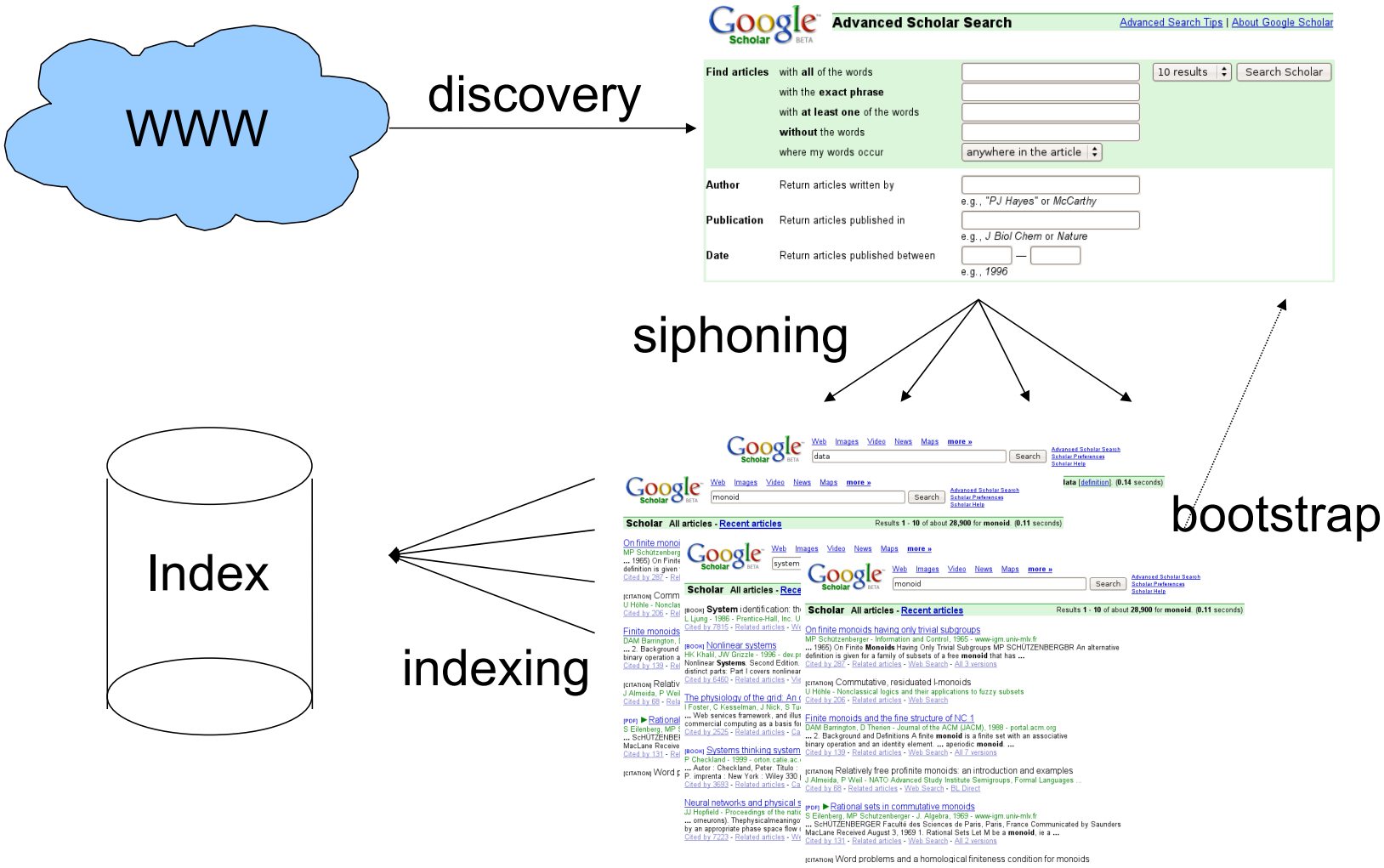
- *Yellow Pages* and other directories;
- Library catalogs;
- Weather services;
- US Census Bureau data;
- etc.

Discovering Knowledge from the Deep Web

- 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 do it automatically, in an unsupervised way?

Focus here: understanding form-based query interfaces, and corresponding result pages.

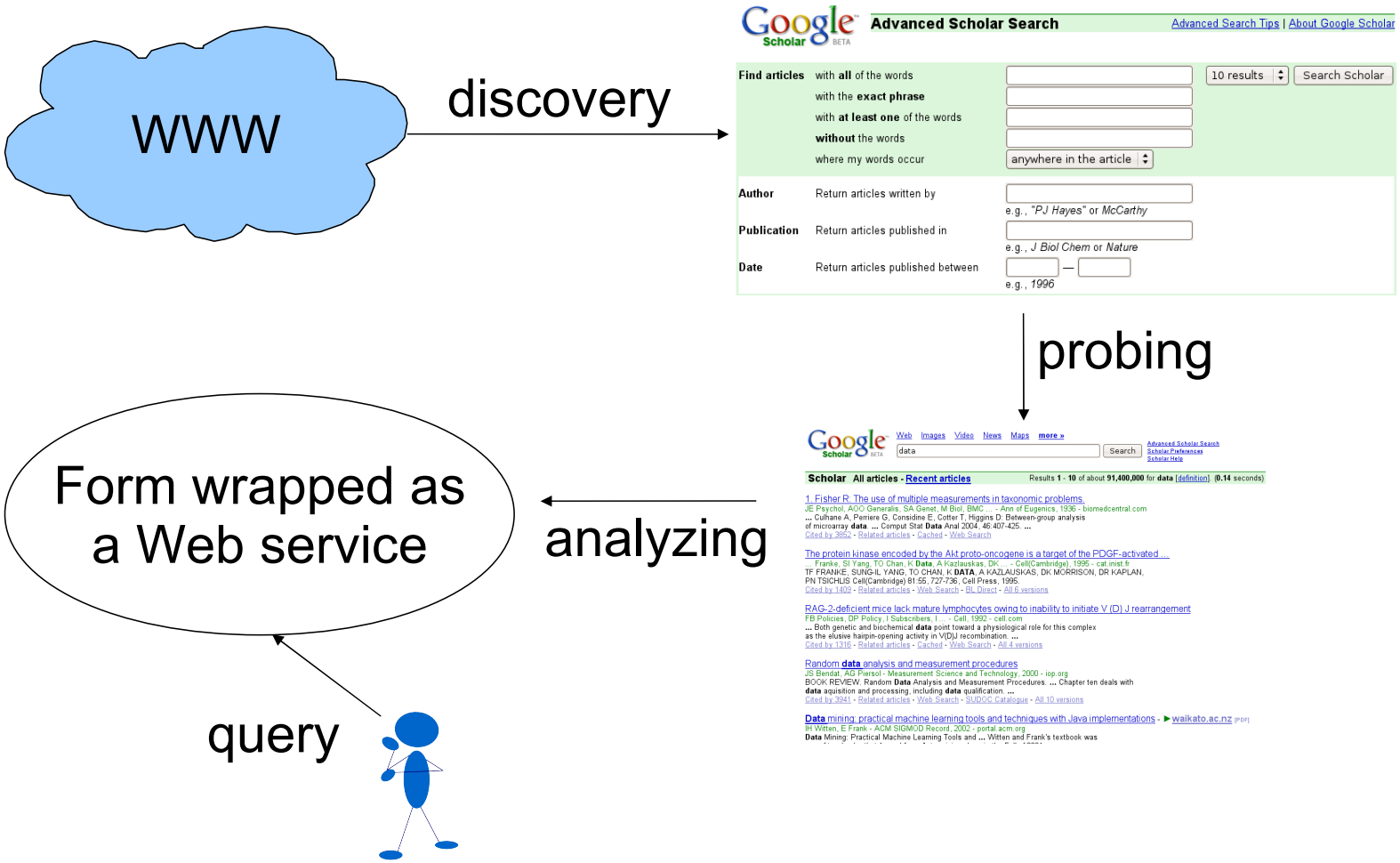
Extensional Approach



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
- Not always feasible (huge load on Web servers)

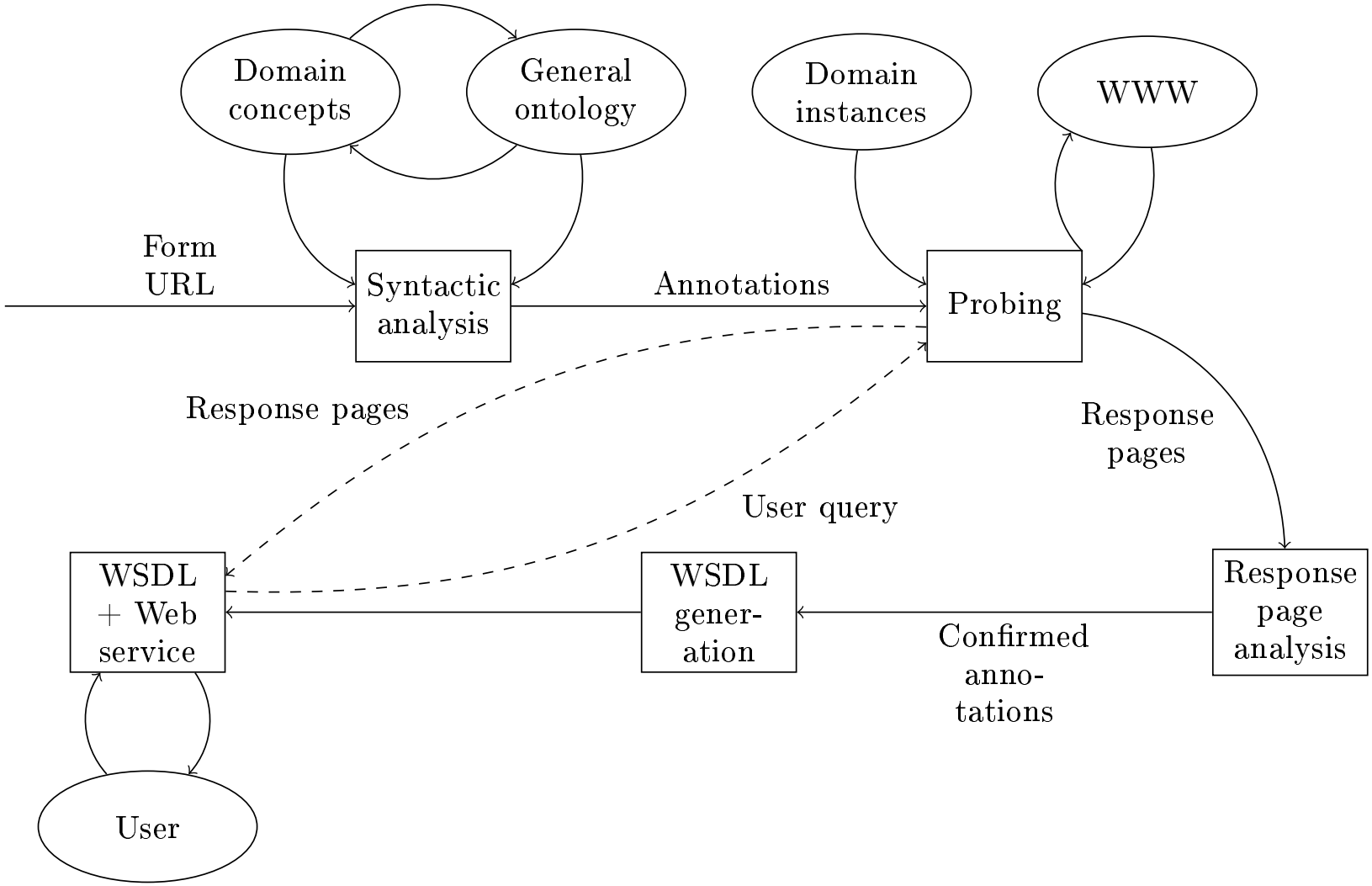
Intensional Approach



Notes on the Intensional Approach

- More ambitious
- 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

General architecture



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.

First 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.

First 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.

First 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.

First 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.

Second Step: Confirm Annotations with 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 using 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.

Second Step: Confirm Annotations with 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 using 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.

Second Step: Confirm Annotations with 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 using 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.

Second Step: Confirm Annotations with Probing

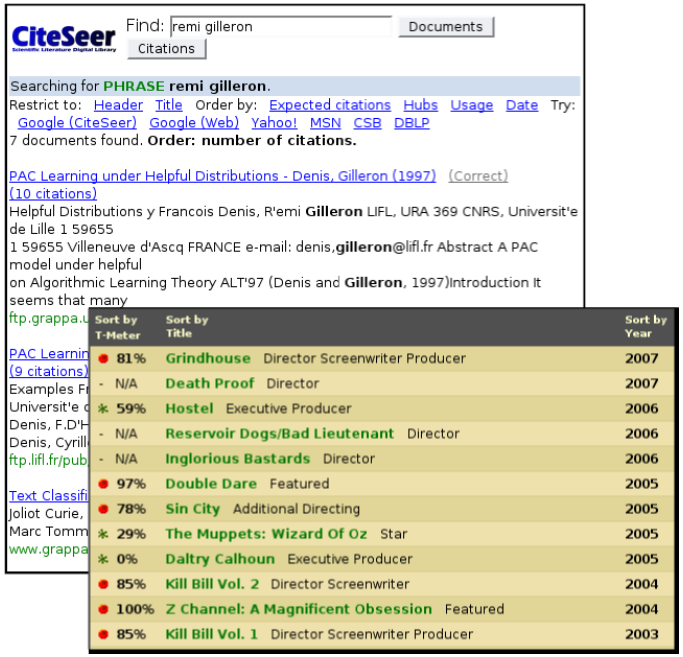
For each field annotated with a concept c:

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 using 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.

Result Pages

Pages resulting from a given form submission:

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



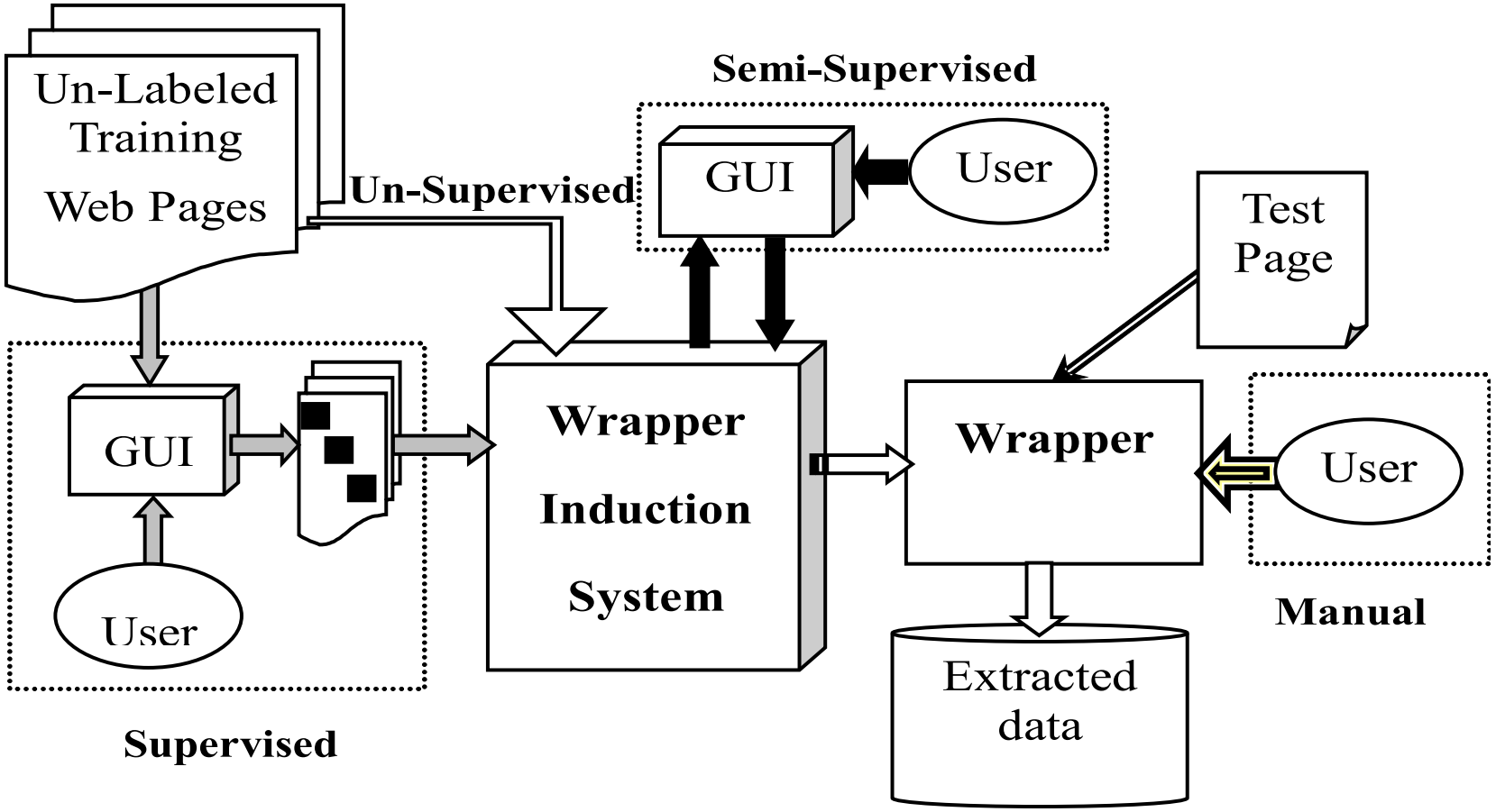
Goal

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.

Different Approaches to Information Extraction

(Chang et al., TKDE 2006)



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

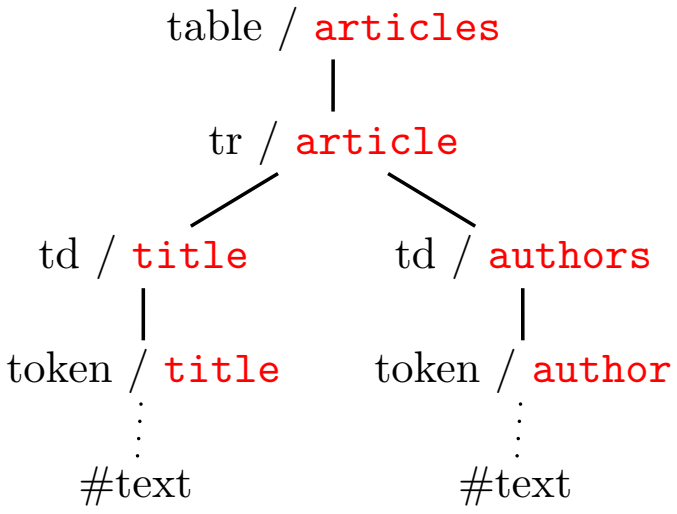
Automatic pre-annotation with domain knowledge (gazetteer):

- 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.



Conditional Random Fields

- Generalization of hidden Markov Models
- 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.

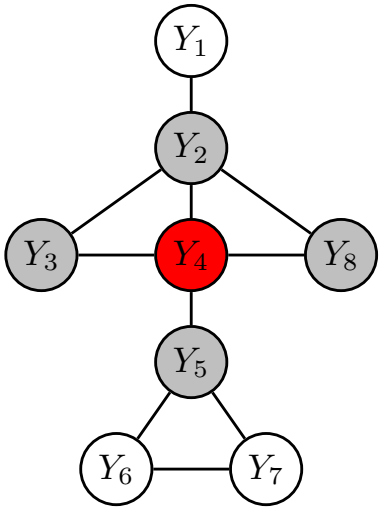
Conditional Random Fields for XML (XCRF)

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.

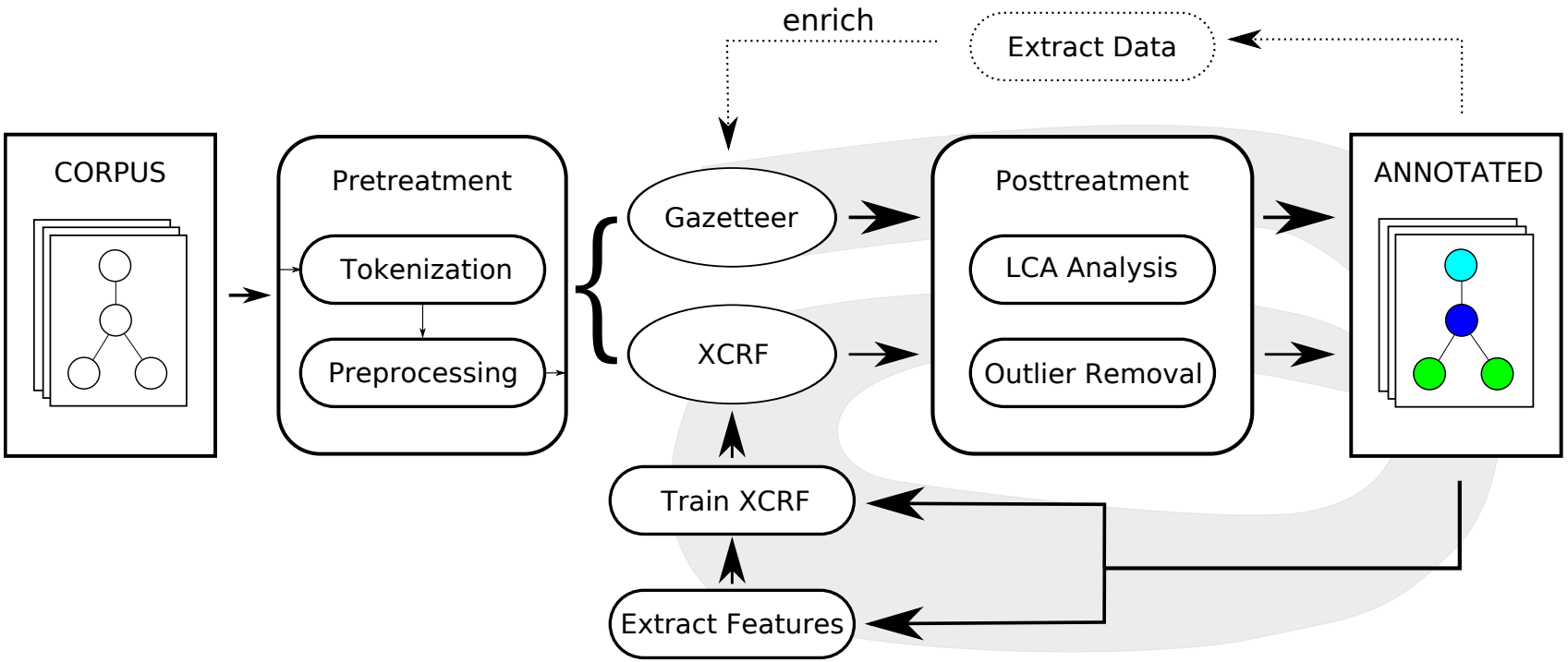
Tree probabilistic model:

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



Most discriminative features selected.

Architecture



Experimental Setup

- 10 services of research publication databases.
- Domain knowledge extracted from DBLP.
- Forms analyzed and probed (5 probes per field and candidate annotation).
- Induction of wrappers from training (unannotated) set of result pages, and evaluation of wrappers on test set of result pages.

Results for form analysis

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

- Good precision and recall.
- Probing raises precision without hurting recall.

Remark

Much better results for distinguishing error and result pages by clustering according to the paths in the DOM tree than previous approaches.

Results for wrapper induction

	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.

Summary

Important point

It is indeed possible to use content and structure together for automatic, unsupervised, information extraction!

- better than content only (gazetteer);
- better than structure only (RoadRunner).

- Content is used to bootstrap a structure-based learning: isn't it what humans do to understand the structure of such pages?
- Not perfect (yet), should be possible to get much better!

Summary

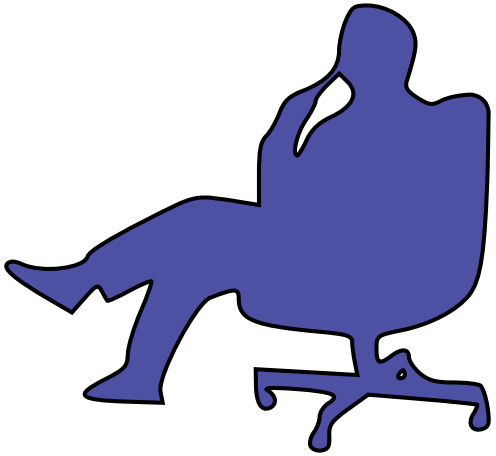
Important point

It is indeed possible to use content and structure together for automatic, unsupervised, information extraction!

- better than content only (gazetteer);
- better than structure only (RoadRunner).

- Content is used to bootstrap a structure-based learning: isn't it what humans do to understand the structure of such pages?
- Not perfect (yet), should be possible to get much better!

Perspectives



- More intelligent gazetteer: use NL features to extract noun phrases that look like titles?
- A machine learning framework adapted to a noisy and incomplete annotation, without overfitting: minimal-length description?.
- Exploit probabilities that come with learned features to produce ranked information extractor.

Merci.