ProFoUnd: Program-analysis–based Form Understanding

(joint work with M. Benedikt, T. Furche, A. Savvides)

PIERRE SENELLART

Ben Gurion University of the Negev, 28 February 2012

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.

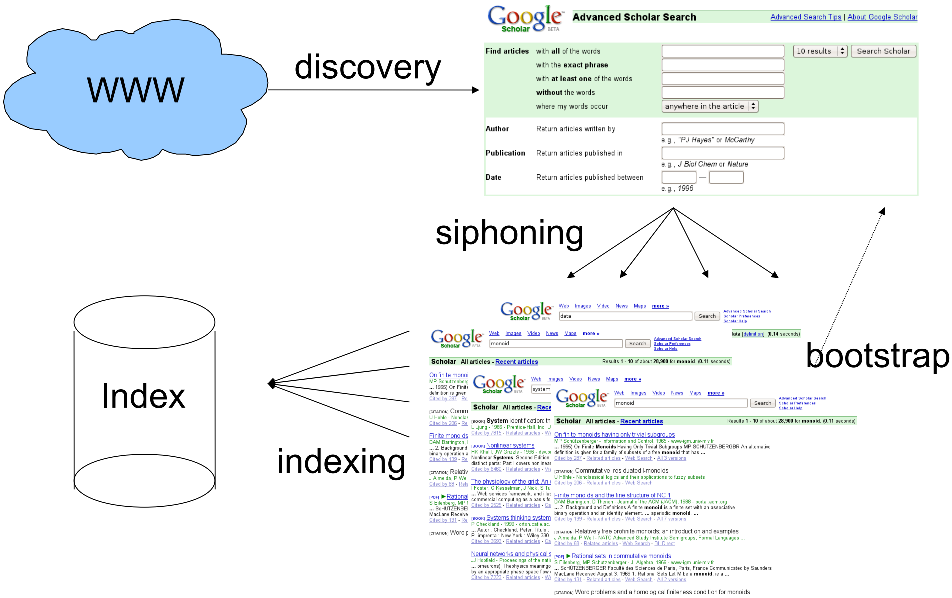
TELECOM ParisTech

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?

Focus here: Automatic, unsupervised, methods

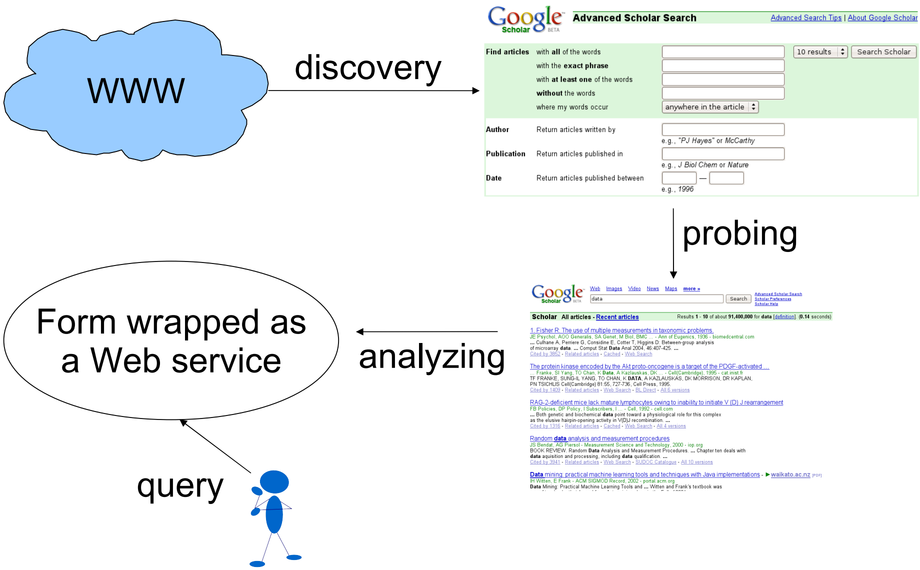
Extensional Approach



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 [Madhavan et al., 2006]
- Not always feasible (huge load on Web servers)

Intensional Approach



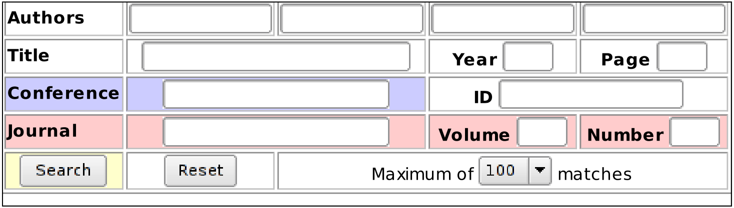
Notes on the Intensional Approach

- 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

Outline

Outline

Analyzing the structure of HTML forms.



Goal

Associating to each form field the appropriate domain concept.

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.

1^{st} 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.

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:
  - It is possible to query a field with a subword
  - All form fields are independent
  - No field is required

TELECOM ParisTech

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

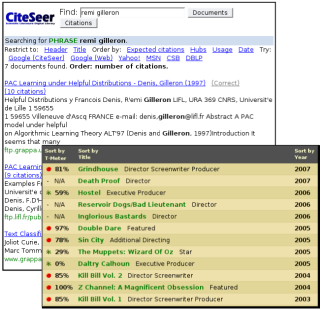
- Probing raises precision without hurting recall
- Clustering according to DOM paths: much better than previous approaches
- But some critical assumptions:
  - It is possible to query a field with a subword
  - All form fields are independent
  - No field is required

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.

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.

TELECOM ParisTech

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

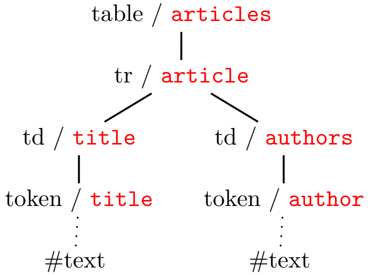
Automatic pre-annotation with domain knowledge (gazetteer):

- 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.



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.

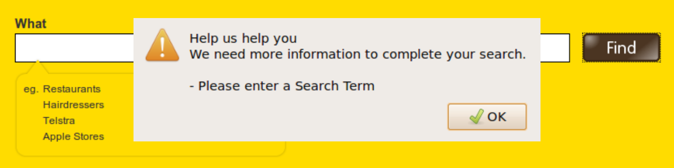
- Main issue: the machine learning assumes that the initial annotation is really the reference one

Outline

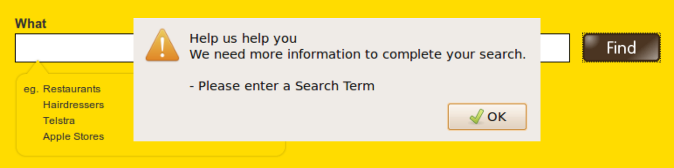
Outline

Better Form Analysis

Better Form Analysis



Better Form Analysis



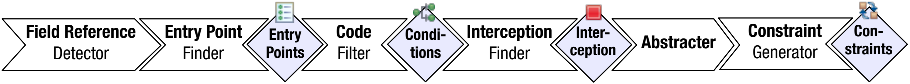

```
// Do not submit unless form is valid
$j("#searchForm").submit(function(event) {
  $j("#searchFormLocationClue").val($j("#searchFormLocationClue").val().trim());
  if ($j("#searchFormBusinessClue").val().isEmpty()) {
    alert('Help us help you\nWe need more information to
      complete your search.\n\n- 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:
  - 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?

Outline

ProFoUnd architecture



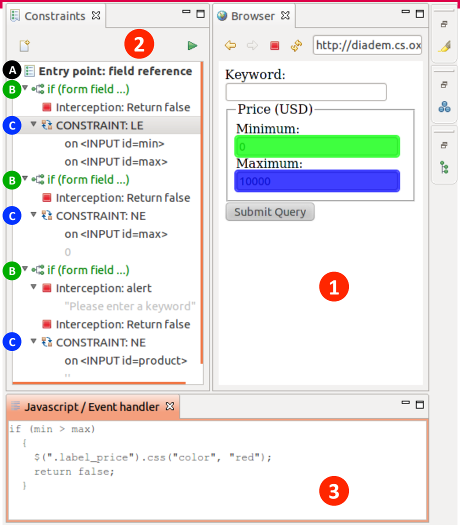
- Entry points are HTML event attributes, setting of event handlers in code, etc. (event: *click* on a submit button, *submit* on a form)
- Conditions are (in)equality tests on form field values (possibly aliased)
- Interceptions are interruptions of the form submission process (error messages, simple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

Abstracting the code

- Rice's theorem: no hope in a sound and complete constraint finder
- But that's ok! Anything that we can learn is more than what we have at the moment.
- Coarse abstraction of the JS code:
  - Only conditions on the code flow from entry points to interceptions are considered.
  - We consider only a simple subset of the JS language; anything beyond that is ignored.
  - Side-effects are mostly ignored
- As a consequence: no guarantee of either soundness or completeness ⇒ only experimental guarantees

Engineering issues to deal with

- Extracting a Web form model: DIADEM's tools http://www.diadem-project.info/
- Parsing JavaScript: Mozilla Rhino (but see later)
- JavaScript frameworks: ad-hoc support for most popular ones (jQuery, Prototype, ASP.NET generated code, YUI, Dojo, MooTools)
- Evaluating JavaScript code (e.g., to determine what a jQuery selector (`$.("form#lookup .product")`) returns): Mozilla JS engine
- Abstraction, alias references, etc.: ProFoUnd core, developed from scratch



1. Web page view, with fields highlighted
2. Constraints found: $min < max$, $max \neq 0$, $product \neq$ ''
3. JS fragment for the highlighted constraint

Preliminary evaluation

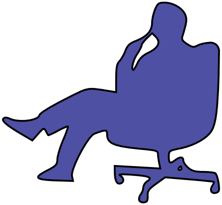
- 70 real-estate websites containing search forms
- 30 out of 70 use client-side validation, with a total of 35 constraints
- 100% precision: all identified constraints are correct
- 63% recall: 22 out of 35 JS-enforced constraints were found
- Why did we miss some?
  - Use of complex JavaScript features,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

Outline

In brief

- Exploiting data from the deep Web in an automatic manner: non-trivial, largely open problem
- Classical techniques exploit both domain knowledge and the structure of forms and result pages
- Possible to get very precise information about the behavior of Web forms by static analysis of client-side code

Perspectives

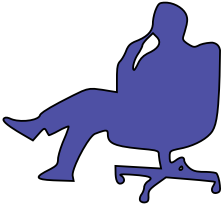


- Use a real JS parser (Rhino has lots of limitations); trying with SpiderMonkey, Mozilla's JS engine
- Large-scale evaluation, application to deep Web crawling
- Type inference for form fields: regular expressions, simple datatypes
- Combining with dynamic analysis
- Type inference for AJAX applications: static analysis of AJAX calls to determine input and output types (possibly JSON or XML types)

PhD Opportunity

PhD scholarship on this topic at U. Oxford, looking for excellent candidates!

Perspectives



- Use a real JS parser (Rhino has lots of limitations); trying with SpiderMonkey, Mozilla's JS engine
- Large-scale evaluation, application to deep Web crawling
- Type inference for form fields: regular expressions, simple datatypes
- Combining with dynamic analysis
- Type inference for AJAX applications: static analysis of AJAX calls to determine input and output types (possibly JSON or XML types)

PhD Opportunity

PhD scholarship on this topic at U. Oxford, looking for excellent candidates!

Merci.

Michael Benedikt, Tim Furche, Andreas Savvides, and Pierre Senellart. ProFoUnd: Program-analysis–based form understanding. In *Proc. WWW*, Lyon, France, April 2012. Demonstration.

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.

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.

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.

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.