

Automatic Wrapper Induction from Hidden-Web Sources with Domain Knowledge

P. Senellart^{1,2} A. Mittal³ D. Muschick⁶ R. Gilleron^{4,5} M. Tommasi^{4,5}

1



2



3



4



5



6



WIDM, 28 October 2008

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 (2001) : 500 times more content than on the **surface Web**!

Sources of the Hidden Web

Example

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

Forms

Analyzing the **structure** of HTML forms.

Authors	<input type="text"/>	<input type="text"/>	<input type="text"/>	<input type="text"/>
Title	<input type="text"/>		Year <input type="text"/>	Page <input type="text"/>
Conference	<input type="text"/>	ID <input type="text"/>		
Journal	<input type="text"/>	Volume <input type="text"/>	Number <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**.

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 Hubs Usage Date Try: Google CiteSeer Google News Yahoo! MSN SSR DBLP
7 documents found. Order: number of citations.

PAC Learning under Helpful Distributions - Denis Gilleron (1997) [Correct]
110 citations
Helpful Distributions y Francois Denis, Remi Gilleron LFL, URA 369 CNRS, Université de Lille 1 59655
1 59655 Villeneuve d'Ascq FRANCE e-mail: denis.gilleron@lfl.fr Abstract A PAC model under helpful
on Algorithmic Learning Theory ALT97 (Denis and Gilleron, 1997) Introduction It seems that many
ftp.grappa.fr

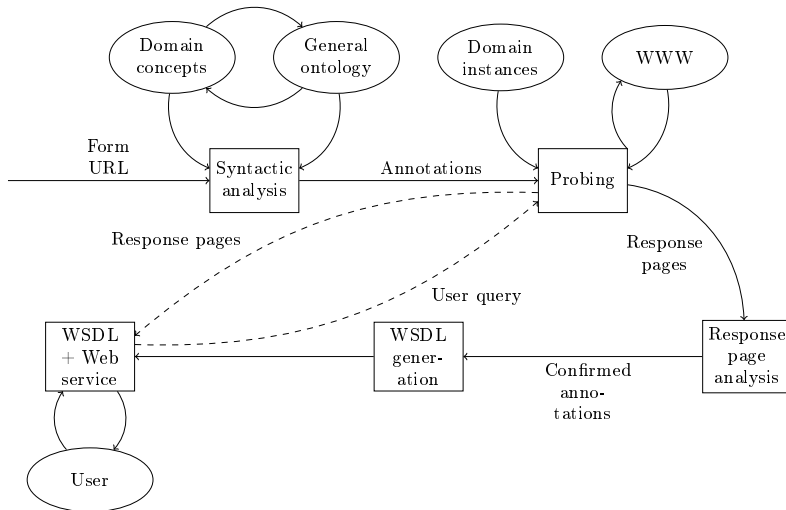
Sort by	Sort by	Sort by
Relevance	Title	Year
81%	Grindhouse	Director Screenwriter Producer
N/A	Death Proof	Director
59%	Hostel	Executive Producer
N/A	Reservoir Dogs/Bad Lieutenant	Director
N/A	Inglorious Bastards	Director
97%	Double Dare	Featured
78%	Sin City	Additional Directing
29%	The Muppets: Wizard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

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**.

General architecture



- 1 Motivation
- 2 Probing
- 3 Two-Step Wrapper Induction
- 4 Experiments
- 5 Conclusion

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

- ❶ Build a **context** for each field:
 - label tag;
 - id and name attributes;
 - text immediately before the field.
- ❷ Remove **stop words**, **stem**.
- ❸ **Match** this context with the concept names, extended with WordNet.
- ❹ 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.

- 1 Motivation
- 2 Probing
- 3 Two-Step Wrapper Induction**
- 4 Experiments
- 5 Conclusion

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 OORL

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: [Phillip 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 Cremanenco](#) (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 OORL

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: [Phillip 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 Cremanenco](#) (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.LG/0601085** [abs, ps, pdf, other] :

Title: A Formal Foundation for ODRL

Authors: **Ricardo Lucella**, **Micky Weissman**

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

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

ACM-class: H.2.7; K.4.4

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

Title: VOFitter, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Zhenzhou Lu** (1), **Mariusz Dolerański** (2), **Peter Quire** (2), **Yongheng Zhao** (1), **Francoise Genova** (3) ((1)NAO China, (2) **ESO**, (3) CDS)

Comments: Accepted for publication in ChAA (9 pages, 2 figures, 165KB)

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

Title: Matching Subsequences in Trees

Authors: **Ying Qian**, **Inge Li Goertz**

Sub-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** (**IRISA**, **Rennes**), **Stéphane Schloer** ()

Sub-class: **Information Retrieval**

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

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: **Luc Bouganim** (**IRISA**, **Rennes**), **Cassidy Cremonese** (**IRISA**, **Rennes**), **François David Nègre** (**IRISA**, **Rennes**), **PRISM - UVSQ**, **Nicolas Ollivier** (**IRISA**, **Rennes**), **Philippe Pucheral** (**IRISA**, **Rennes**), **PRISM - UVSQ**

Sub-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.LG/0601085** [abs, ps, pdf, other] :

Title: A Formal Foundation for ODRL

Authors: **Ricardo Lucella**, **Micky Weissman**

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

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

ACM-class: H.2.7; K.4.4

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

Title: VOFitter, Bridging Virtual Observatory and Industrial Office Applications

Authors: **Zhen-zhou Lu** (1), **Mariusz Dolerański** (2), **Peter Quire** (2), **Yongheng Zhao** (1), **Francoise Genova** (3) ((1)NAO China, (2) **ESO**, (3) CDS)

Comments: Accepted for publication in ChAA (9 pages, 2 figures, 165KB)

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

Title: Matching Subsequences in Trees

Authors: **Yingqin Qian**, **Inge Li Goertz**

Sub-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** (**IRISA**, **Rennes**), **Stéphane Guillou** ()

Sub-class: **Information Retrieval**

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

Title: Safe Data Sharing and Data Dissemination on Smart Devices

Authors: **Luc Bouganim** (**IRISA**, **Rennes**), **Cassidy Cremarino** (**IRISA**, **Rennes**), **François David Nappé** (**IRISA**, **Rennes**), **PRISM - UVSQ**, **Nicolas Ollivier** (**IRISA**, **Rennes**), **Philippe Pucheral** (**IRISA**, **Rennes**), **PRISM - UVSQ**

Sub-class: **Cryptography and Security**: Databases

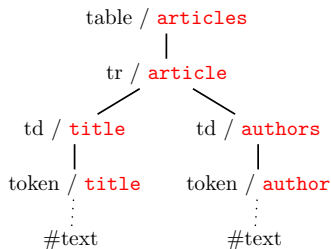
Automatic **pre-annotation** with domain knowledge (gazetteer):

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



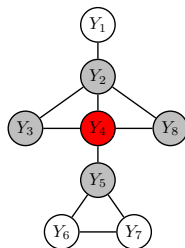
Conditional Random Fields for XML (XCRF)

Observations: 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.

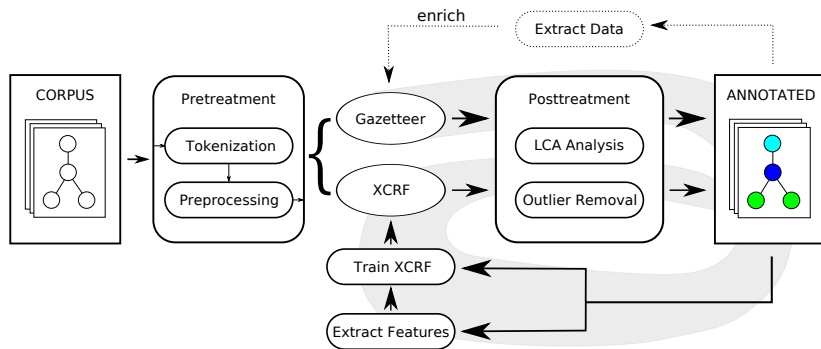
Tree probabilistic model:

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



Most **discriminative** features selected.

Architecture



- 1 Motivation
- 2 Probing
- 3 Two-Step Wrapper Induction
- 4 Experiments**
- 5 Conclusion

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.

- 1 Motivation
- 2 Probing
- 3 Two-Step Wrapper Induction
- 4 Experiments
- 5 Conclusion**

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!

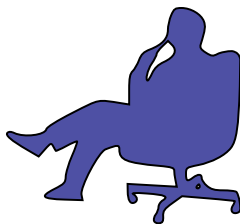
Summary

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.