ProFoUnd: Program-analysis–based Form Understanding

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

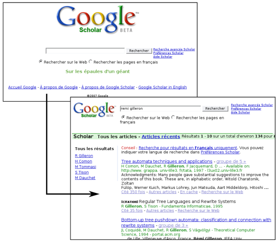
PIERRE SENELLART

IC2 Group Seminar, 16 May 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.


Google Scholar BETA
Rechercher
Recherche avancée Scholar
Préférences Scholar
Aide Scholar
Rechercher sur le Web
Rechercher les pages en français
Sur les épaules d'un géant
Accueil Google - À propos de Google - À propos de Google Scholar - Google Scholar in English
©2007 Google
remi gilleron
Scholar Tous les articles - Articles récents
Tous les résultats
R Gilleron
H Comon
M Tommasi
S Tison
M Dauchet

Résultats 1 - 10 sur un total d'environ 134 pour r

Conseil : Recherche pour résultats en Français uniquement. Vous pouvez indiquer votre langue de recherche dans Préférences Scholar.

Tree automata techniques and applications - groupe de 5 »
H Comon, M Dauchet, R **Gilleron**, F Jacquemard, D ... - Available on: http://www. grappa. univ-lille3. fr/tata, 1997 - l3ux02.univ-lille3.fr
Acknowledgments: Many people gave substantial suggestions to improve the contents of this book. These are, in alphabetic order, Witold Charatonik, Zoltan
Fülöp, Werner Kuich, Markus Lohrey, Jun Matsuda, Aart Middeldorp, Hitoshi ...
Cité 350 fois - Autres articles - En cache - Recherche sur le Web

[CITATION] Regular Tree Languages and Rewrite Systems
R **Gilleron**, S Tison - Fundamenta Informaticae, 1995
Cité 35 fois - Autres articles - Recherche sur le Web

Bottom-up tree pushdown automata: classification and connection with rewrite systems - groupe de 3 »
JL Coquidé, M Dauchet, R **Gilleron**, S Vágvölgyi - Theoretical Computer Science, 1994 - portal.acm.org
de Lille, Villeneuve d'Ascq, France. **Rémi Gilleron**, IEEA Univ.

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

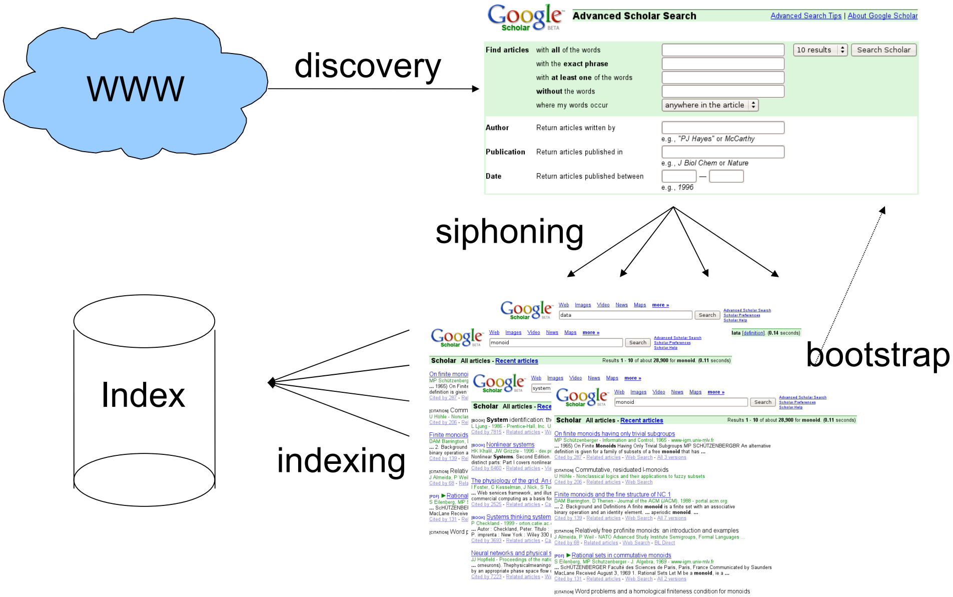
- *Yellow Pages* and other directories;
- Library catalogs;
- Weather services;
- US Census Bureau data;
- etc.

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?

Focus here: Automatic, unsupervised, methods

Extensional Approach



Advanced Scholar Search — Advanced Search Tips | About Google Scholar

Find articles with **all** of the words
with **the exact phrase**
with **at least one** of the words
without the words
where my words occur: anywhere in the article

10 results — Search Scholar

Author Return articles written by
e.g., "PJ Hayes" or McCarthy

Publication Return articles published in
e.g., J Biol Chem or Nature

Date Return articles published between —
e.g., 1996

Web Images Video News Maps **more »**
monoid — Search
Advanced Scholar Search
Scholar Preferences
Scholar Help

Scholar All articles - Recent articles — Results **1** - **10** of about **28,900** for **monoid**. (**0.11** seconds)

On finite monoids having only trivial subgroups
MP Schützenberger - Information and Control, 1965 - www-igm.univ-mlv.fr
... 1965) On Finite **Monoids** Having Only Trivial Subgroups MP SCHÜTZENBERGBR An alternative definition is given for a family of subsets of a free **monoid** that has ...
Cited by 287 - Related articles - Web Search - All 3 versions

[CITATION] Commutative, residuated l-monoids
U Höhle - Nonclassical logics and their applications to fuzzy subsets
Cited by 206 - Related articles - Web Search

Finite monoids and the fine structure of NC 1
DAM Barrington, D Therien - Journal of the ACM (JACM), 1988 - portal.acm.org
... 2. Background and Definitions A finite **monoid** is a finite set with an associative binary operation and an identity element. ... aperiodic **monoid**. ...
Cited by 139 - Related articles - Web Search - All 7 versions

[CITATION] Relatively free profinite monoids: an introduction and examples
J Almeida, P Weil - NATO Advanced Study Institute Semigroups, Formal Languages ...
Cited by 68 - Related articles - Web Search - BL Direct

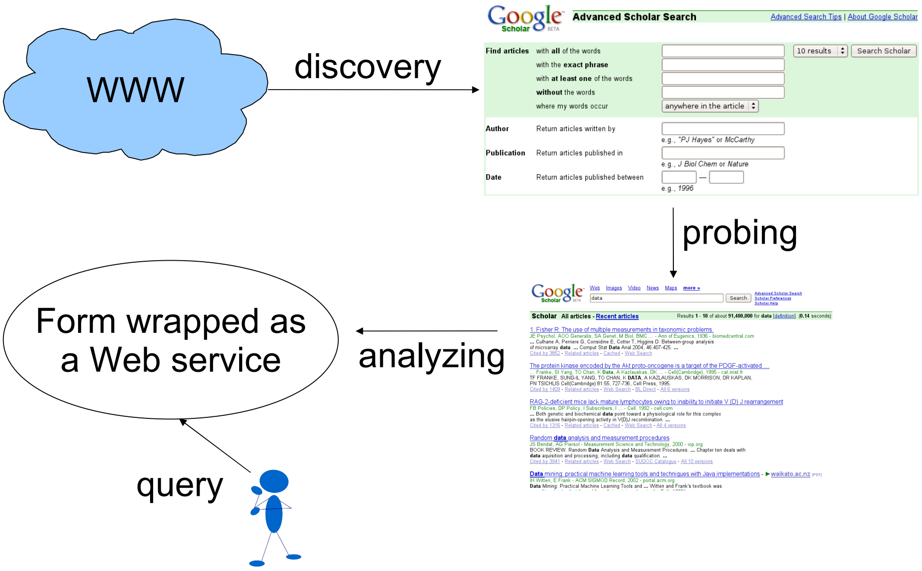
[PDF] ► Rational sets in commutative monoids
S Eilenberg, MP Schutzenberger - J. Algebra, 1969 - www-igm.univ-mlv.fr
... ScHÜTZENBERGER Faculté des Sciences de Paris, Paris, France Communicated by Saunders MacLane Received August 3, 1969 1. Rational Sets Let M be a **monoid**, ie a ...
Cited by 131 - Related articles - Web Search - All 2 versions

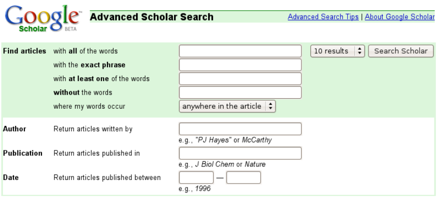
[CITATION] Word problems and a homological finiteness condition for monoids

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

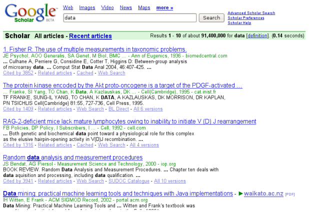



Google Scholar BETA

Advanced Scholar Search — Advanced Search Tips | About Google Scholar

Find articles	with **all** of the words	
	with the **exact phrase**	
	with **at least one** of the words	
	without the words	
	where my words occur	anywhere in the article
Author	Return articles written by	e.g., *"PJ Hayes"* or *McCarthy*
Publication	Return articles published in	e.g., *J Biol Chem* or *Nature*
Date	Return articles published between	— e.g., *1996*

10 results · Search Scholar


Google Scholar BETA — Web Images Video News Maps **more »**

data · Search — Advanced Scholar Search · Scholar Preferences · Scholar Help

Scholar **All articles** - **Recent articles** — Results **1** - **10** of about **91,400,000** for **data** [definition]. (**0.14** seconds)

1. Fisher R: The use of multiple measurements in taxonomic problems.
JE Psychol, AOO Generalis, SA Genet, M Biol, BMC ... - Ann of Eugenics, 1936 - biomedcentral.com
... Culhane A, Perriere G, Considine E, Cotter T, Higgins D: Between-group analysis
of microarray **data**. **...** Comput Stat **Data** Anal 2004, 46:407-425. **...**
Cited by 3852 - Related articles - Cached - Web Search

The protein kinase encoded by the Akt proto-oncogene is a target of the PDGF-activated ...
... Franke, SI Yang, TO Chan, K **Data**, A Kazlauskas, DK ... - Cell(Cambridge), 1995 - cat.inist.fr
TF FRANKE, SUNG-IL YANG, TO CHAN, K **DATA**, A KAZLAUSKAS, DK MORRISON, DR KAPLAN,
PN TSICHLIS Cell(Cambridge) 81:55, 727-736, Cell Press, 1995.
Cited by 1409 - Related articles - Web Search - BL Direct - All 6 versions

RAG-2-deficient mice lack mature lymphocytes owing to inability to initiate V (D) J rearrangement
FB Policies, DP Policy, I Subscribers, I ... - Cell, 1992 - cell.com
... Both genetic and biochemical **data** point toward a physiological role for this complex
as the elusive hairpin-opening activity in V(D)J recombination. **...**
Cited by 1316 - Related articles - Cached - Web Search - All 4 versions

Random **data** analysis and measurement procedures
JS Bendat, AG Piersol - Measurement Science and Technology, 2000 - iop.org
BOOK REVIEW. Random **Data** Analysis and Measurement Procedures. **...** Chapter ten deals with
data aquisition and processing, including **data** qualification. **...**
Cited by 3941 - Related articles - Web Search - SUDOC Catalogue - All 10 versions

Data mining: practical machine learning tools and techniques with Java implementations - ► waikato.ac.nz [PDF]
IH Witten, E Frank - ACM SIGMOD Record, 2002 - portal.acm.org
Data Mining: Practical Machine Learning Tools and **...** Witten and Frank's textbook was

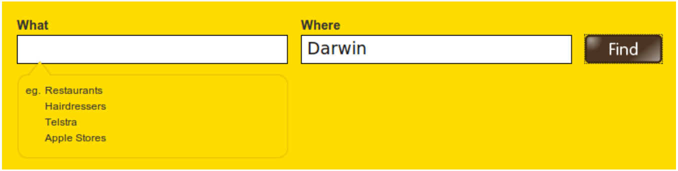
Notes on the Intensional Approach

- 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

Better Form Analysis

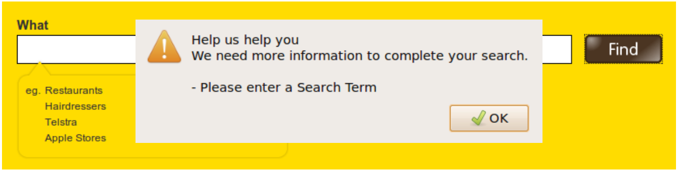

What

eg. Restaurants
Hairdressers
Telstra
Apple Stores

Where

Darwin

Better Form Analysis

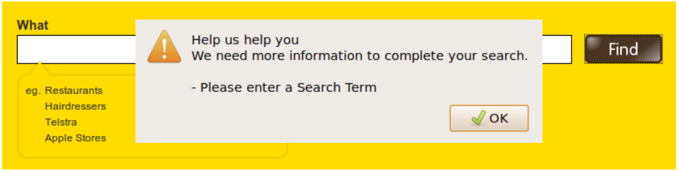

What

eg. Restaurants
Hairdressers
Telstra
Apple Stores

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

- Please enter a Search Term

Better Form Analysis


What

eg. Restaurants
Hairdressers
Telstra
Apple Stores

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

- Please enter a Search Term

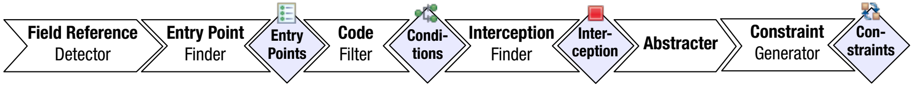
```
// Do not submit unless form is valid
$j("#searchForm").submit(function(event) {
  $j("#searchFormLocationClue").val($j("#searchFormLocationClue").val().trim());
  if ($j("#searchFormBusinessClue").val().isEmpty()) {
    alert('Help us help you\nWe need more information to
      complete your search.\n\n-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

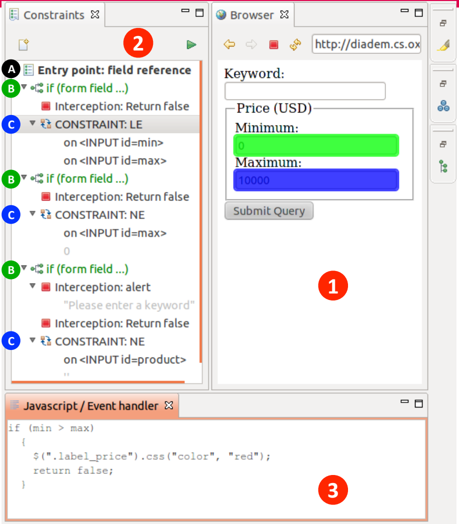
Abstracting the code

- 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 $\Rightarrow$ 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

ProFoUnd interface [Benedikt et al., 2012]



A Entry point: field reference
B if (form field ...)
Interception: Return false
C CONSTRAINT: LE
on <INPUT id=min>
on <INPUT id=max>
B if (form field ...)
Interception: Return false
C CONSTRAINT: NE
on <INPUT id=max>
0
B if (form field ...)
Interception: alert
"Please enter a keyword"
Interception: Return false
C CONSTRAINT: NE
on <INPUT id=product>
''

Keyword:	
Price (USD)	
Minimum:	0
Maximum:	10000

```
if (min > max)
  {
    $(".label_price").css("color", "red");
    return false;
  }
```

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

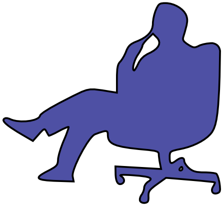
- 70 real-estate websites containing search forms
- 30 out of 70 use client-side validation, with a total of 35 constraints
- 100% precision: all identified constraints are correct
- 63% recall: 22 out of 35 JS-enforced constraints were found
- Why did we miss some?
  - Use of complex JavaScript features, 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

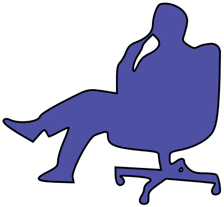


- Use a real JS parser (Rhino has lots of limitations); trying with SpiderMonkey, Mozilla's JS engine
- Large-scale evaluation, application to deep Web crawling
- Type inference for form fields: regular expressions, simple datatypes
- Combining with dynamic analysis
- Type inference for AJAX applications: 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.