

PS Web IE

<http://dbai.tuwien.ac.at/education/wie>

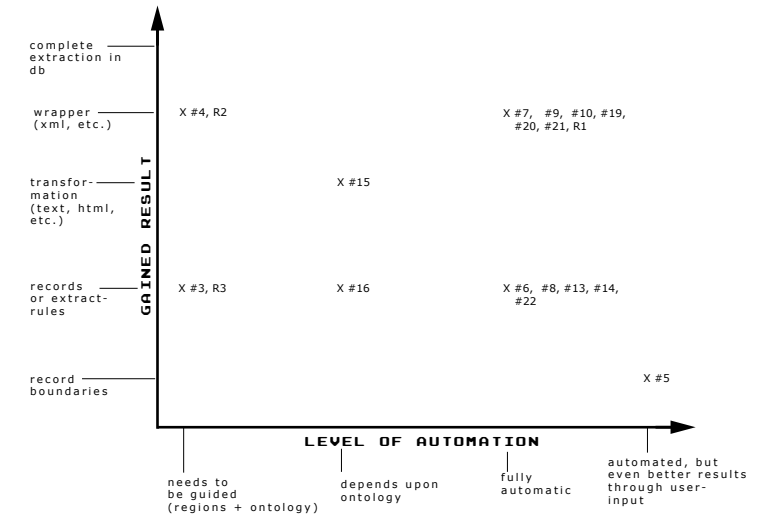
End hand-ins

Wednesday, April 27, 2005

SUMMARY: WEB INFORMATION EXTRACTION

Ch. Veigel vs. René C. Kiesler, Summer 2005

Method \ Paper	3	4	5	6	7	8	9	10	13	14	15	16	19	20	21	22	R1	R2	R3
needs user-interaction	X	X					X		X		X	X	X	X					
domain dependent											X	X						X	
uses training examples	X	X					X		X	X			X	X	X	X			
uses special input structure		X		X							X							X	X
depends on html-tags	X	X	X			X	X	X	X	X		X	X	X	X			X	
uses DOM/T-DOM								X					X	X					
uses hierarchy		X		X					X							X	X		
ontology-based											X	X							X
handles non-cont. data						X											X		
uses template-finding				X	X					X				X		X			
finds slots/data sections	X			X	X	X	X			X			X	X	X	X		X	X
finds data-tables						X	X		X			X		X	X			X	X
generates tag/token-tree		X	X			X			X				X		X				
extracts records	X		X	X	X	X					X	X	X	X		X		X	X
extracts columns				X	X						X	X	X	X				X	X
annotates data (XML)		X						X		X	X		X				X		
content partition / aggreg.								X									X	X	X
generates schema														X					X
infers grammar					X														
generates wrapper/rules	X	X			X		X						X	X	X	X			X
uses 3 rd -party tools				X	X						X			X			X		X



- 03: Stefan Schönig vs. Stalker
- 04: Marco Schönig vs. Xwrap
- 05: René C. Kiesler vs. Record Boundary Discovery
- 06: Christoph Veigel vs. Automated Segmentation
- 07: Marian Schedenig vs. Automatic Data Extraction
- 08: Paul Bohunsky vs. Mining Web Pages
- 09: Sunil Pilani vs. Flexible Learning
- 10: Friedrich Diemmel vs. Gateway HTML -> XML
- 13: Gregor Pridun vs. Table Detection
- 14: Max Arends vs. Data-rich Section Extraction
- 15: Stefan Bischof vs. Automatic Knowledge Extraction
- 16: Stefan Rümmele vs. HTML Table extraction
- 19: Leopold Redlingshofer vs. Data Extraction with Labels
- 20: Edvin Seferovic vs. Data Extraction and Annotation
- 21: Jeremy Solarz vs. Applying Pattern Mining
- 22: Tobias Doenz vs. Extracting Structured Data

- R1: Martin Zeilinger vs. Newswrapper for Brokers
- R2: Taair Hassan vs. Intelligent Text Extraction
- R3: Rainer Dobiasch vs. MOMIS Schema-Merging

<i>Paper</i>	<i>wrapper</i>	<i>DOM-based</i>	<i>equivalence classes</i>	<i>pat trie</i>	<i>HTML-based</i>	<i>non-contiguous records</i>	<i>multi-page records</i>	<i>ontology</i>	<i>heuristic</i>	<i>regexp</i>	<i>supervised</i>	<i>template discovery</i>	<i>precision</i>	<i>implementation</i>
#19	+	+	-	+?	+	-	-	-	+	+	-	+		-
#20	+	+	-	-	+	-	-	-	+	-	-	+	90%	-
#4	+	+*	-	-	+	-	-	-	+	-	+	-**		-
#3	+	-	-	-	+	-	-	-	-	-	+	-		-
#5	-	+	-	-	+	-	-	+***	+	-	-	-		-
#6	+	-	-	-	+	-	+	-	+****	-	-	+	74%-85%	-
#10	-	+	-	-	+	-	-	-	?	-	-	-		+
#21	.	-	-	+	+	-	-	-	-	-	-	-		-
#22	opt	-	+	-	-	-	-	-	-	-	-	+	80%	-
#13	-	-	-	-	+	-	-	-	+*****	-	+	-		-
#14	-	+	-	-	+	-	+	-	-	-	-	+		-
#16	+	+	-	-	+	+	+	+	-	-	-	-		+
#15	-	-	-	-	+		+	+	+	-	-	-		+
#7	+	-	-	-	-	-	-	-	+	-	-	+	70%	-
#8	-	+	-	-	+	+	-	-	-	-	-	-	99%	+

- *) unsupervised derivative exists
- ***) user
- ****) optional
- *****) Hidden Markov Model
- *****) SVM

Web Information Extraction Comparison Matrix

Paul Bohunsky (0025058)

Marian Schedenig (9725416)

PS WIE End-Hand-In Friedrich Dimmel Matr.Nr. 0302230 Kennz. 534 fritz@dimmel.at	<i>uses heuristics</i>	<i># of heuristics</i>	<i>ontologies</i>	<i>supervised</i>	<i>unsupervised</i>	<i>semi-supervised</i>	<i>wrapper</i>	<i>knowledge based</i>	<i>machine learning</i>	<i>XML compatible</i>	<i>full document extr</i>	<i>only lists & tables or user-selections</i>	<i>using DOM</i>	<i>layout based</i>
#3 Hierarchical Approach to wrapper induction				x				x				x	x	
#4 XWRAP				x			x			x		x		
#5 Rec.Bound. in Web docs	x	5			x						x			
#6 Using structure...	x	2			x							x		x
#7 Autom. DE from Lists and Tables					x		x					x		
#8 Mining Web pages for data records					x							x		
#9 FLS	x				x		x	x	x			x	x	x
#10 Gateway HTML to XML					x					x	x		x	x
#13 ML Based Approach for Table detection						x		x	x			x		x
#14 Data-rich Section extraction					x			x	x			x		
#15 Automatic ont-based Extraction			x		x		x	x	x	x	x			
#16 Automating Extraction from HTML Tables with Unknown Str.	x		x	x			x		x			x		
#19 DE and Label Assignment for Web DBs	x				x		x						x	x
#20 DE and Annotation	x				x		x			x		x	x	x
#21 Applying Pattern Mining					x									
#22 Extracting Structured Data					x									

	#3	#4	#5	#6	#7	#8	#9	#10	#13	#14	#15	#16	#19	#20	#21	#22
#3																
#4	X															
#5																
#6																
#7	X	X		X												
#8			X	X												
#9	X	X			X	X										
#10	X	X		X		X	X									
#13			X													
#14				X				X	X							
#15						X										
#16					X			X		X	X					
#19				X				X			X	X				
#20	X	X	X	X		X	X	X		X		X	X			
#21	X	X	X	X				X	X			X		X		
#22	X	X		X		X	X	X	X		X	X	X	X	X	

Legende:

X.....Ähnlichkeit in Sachgebiet ist gegeben

Max Arends 9835111

Gregor Pridun 9725153

PS WIE: End Hand-In

Low Abstraction

User Aided
Wrapper Creation
#3, #4

Data Rich
Section Discovery
#5, #6, #13, #14

Automatic Wrapper
Creation Using Visual
Characteristics
#9, #10

Automatic Wrapper
Creation Using Structure
Characteristics
#7, #8, #21, #22

Ontology Based
Information Extraction
#15, #16

Automatic Wrapper
Creation With
Label Assignment
#19, #20

High Abstraction

Stefan Bischof, Stefan Rümmele

Paper number	Supervised	Unsupervised	Human Interaction	(Token) Suffix Tree	Pattern Tree	DOM/Tag Tree	Wrapper Generation	CSP	Probabilistic methods	Knowledge Based System	Heuristics	Data Annotation	Machine Learning	Depth First Algorithm	String Matching Algorithm	WL ² -Algorithm	Normalisation of content	Ontology	Output: Record Boundary	Output: CSV	Output: Text	Output: Flat Table	Output: Data Rich Section	Output: XML
#3	x		x							x														
#4	x		x			x											x			x				x
#5		x				x	x				x							x	x					
#6		x					x	x											x					
#7		x					x																	
#8		x				x	x								x									
#9							x			x						x						x		
#10		x				x						x					x							x
#13			x										x										x	
#14		x				x																	x	
#15		x	x							x								x		x				
#16		x	x															x				x		
#19		x		x	x	x	x				x	x			x							x		
#20		x				x	x				x	x		x										x
#21		x	x	x	x		x																	
#22		x					x																	

#3 Stefan Schoenig: Ion Muslea, Steve Minton, Craig Knoblock. A Hierarchical Approach to Wrapper Induction. Proceedings of the Third International Conference on Autonomous Agents (Agents'99), Seattle, WA.

#4 Marco Schoenig: Ling Liu, Cailton Pu, Wei Han.XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources. International Conference on Data Engineering (ICDE), pages 611--621, 2000.

#5 René Kiesler: D. W. Embley, Y. Jiang, Y.-K. Ng. Record-Boundary Discovery in Web Documents. Proceedings of the 1999 ACM SIGMOD international conference on Management of data, Philadelphia, Pennsylvania, 467 - 478.

#6 Christoph Veigl: Kristina Lerman, Lise Getoor, Steven Minton, Craig Knoblock. Using the Structure of Web Sites for Automatic Segmentation of Tables. Proceedings of the 2004 ACM SIGMOD international conference on Management of data, Paris, France.

#7 Marian Schedenig: Kristina Lerman, Craig Knoblock, Steven Minton. Automatic Data Extraction from Lists and Tables in Web Sources. In Proceedings of the workshop on Advances in Text Extraction and Mining, IJCAI-2001.

#8 Paul Bohunsky: Bing Liu, R. Grossman, Yanhong Zhai. Mining Web Pages for Data Records. Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining, Washington, D.C., 2003.

#9 Sunil Piliati: William W. Cohen, Matthew Hurst, Lee S. Jensen. A Flexible Learning System for Wrapping Tables and Lists in HTML Documents. In The Eleventh International World Wide Web Conference WWW-2002, 2002.

#10 Friedrich Dimmel: Tao Fu, Mengchi Liu. A Gateway From HTML to XML. IDEAS 2004: 205-21.

#13 Gregor Pridun: Yalin Wang, Jianying Hu. A Machine Learning Based Approach for Table Detection on The Web. Proceedings of the eleventh international conference on World Wide Web, Honolulu, Hawaii, USA, 2002.

#14 Max Arends: Jiyong Wang, Fred H. Lochovsky. Data-rich Section Extraction from HTML pages. Proceedings of the 3rd International Conference on Web Information Systems Engineering, Pages: 313 - 322, 2002.

#15 Stefan Bischof: Harith Alani, Sanghee Kim, David E. Millard, Mark J. Weal, Wendy Hall, Paul H. Lewis, Nigel R. Shadbolt. Automatic Ontology-Based Extraction from Web Documents. IEEE Intelligent Systems, Volume 18, Issue 1, January 2003.

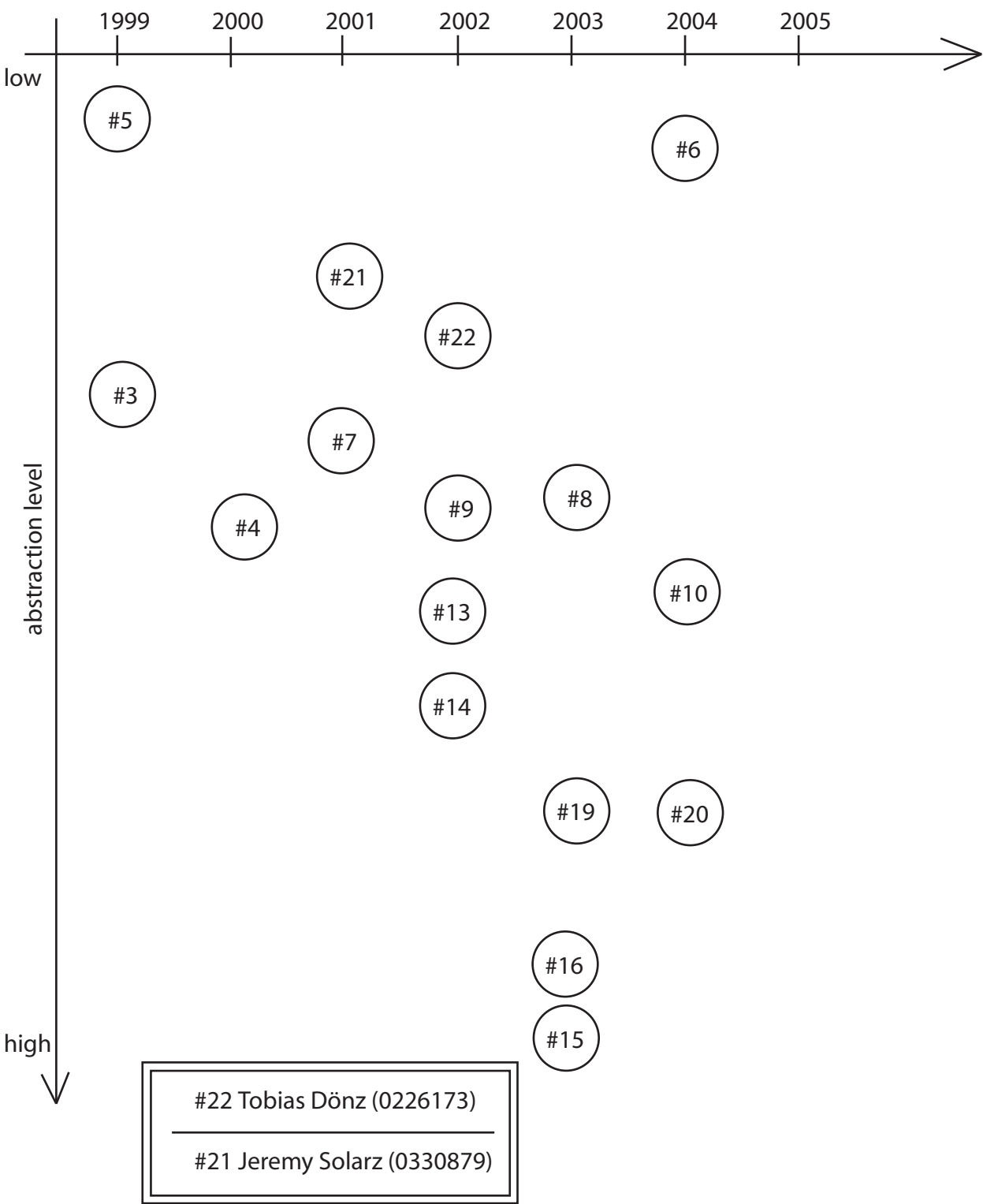
#16 Stefan Rümmele: David W. Embley, Cui Tao, Stephen W. Liddle. Automating the Extraction of Data from HTML Tables with Unknown Structure. Submitted , May 2003, (source: <http://www.deg.byu.edu/papers/>).

#19 Leopold Redlingshofer: Jiyong Wang, Frederick H. Lochovsky. Data Extraction and Label Assignment for Web Databases. Proceedings of the twelfth international conference on World Wide Web, Budapest, Hungary, 2003.

#20 Edvin Seferovic: Hui Song, Suraj Giri, Fanyuan Ma. Data Extraction and Annotation for Dynamic Web Pages. 2004 IEEE International Conference on e-Technology, e-Commerce and e-Service (EEE'04) March 28 - 31, 2004 Taipei, Taiwan.

#21 Jeremy Solarz: Chia-Hui Chang, Shao-Chen Lui, Yen-Chin Wu, Applying Pattern Mining to Web Information Extraction. Proceedings of the 5th Pacific-Asia Conference on Knowledge Discovery and Data Mining, 2001.

#22 Tobias Donz: Arvind Arasu, Hector Garcia-Molina. Extracting Structured Data from Web Pages. Proceedings of the 2003 ACM SIGMOD international conference on Management of data, San Diego, California.



Paper characteristics matrix

Nr.	Supervised Unsupervised	Annotation	Approach	* Content structure type				uses heuristics
				uses ML	table	list	compl.	
#3	×	—	finite automate	—	×	×	-/×	—
#4	×	—	tag ⁽¹⁾ tree	—	×	×	×	?
#5	—	—	tag tree	—	×	-/×	—	×
#6	—	—	CSP HMM	—	×	×	—	×
#7	—	—	token sequence	×	×	×	—	—
#8	—	—	tag tree string - match	—	×	—	—	×
#9	—	—	wrapper learning	×	×	×	-/×	?
#10	—	×	DOM tree	—	×	×	-/×	×
#13	?	-/?	ML	×	×	—	—	—
#14	—	—	data-rich sections	—	×	×	—	?
#15	—	×	ontolog.	—	×	×	×	?
#16	—	×	ontolog.	—	×	—	—	×
#19	—	×	suffix pattern	×	×	×	-/?	×
#20	—	×	tag tree comp.	×	×	—	—	×
#21	—	×	PAT Tree	—	×	×	—	—
#22	—	—	LFEQ dtoken	—	×	-/×	×	×

* ML algorithm and/or training set

(1) to #3, #4: extraction rules

-/× : teilweise
-/? : eher nicht