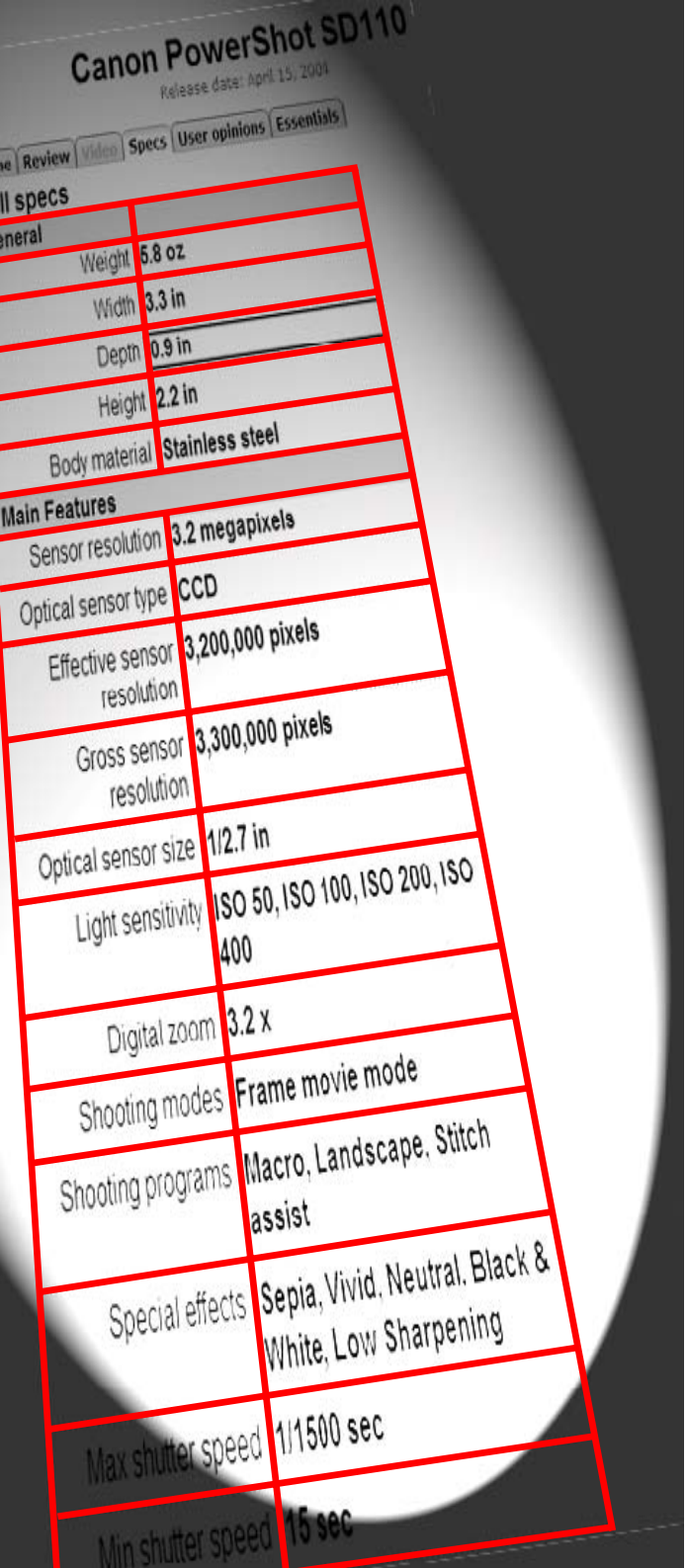


Updated version:
July 23, 2006

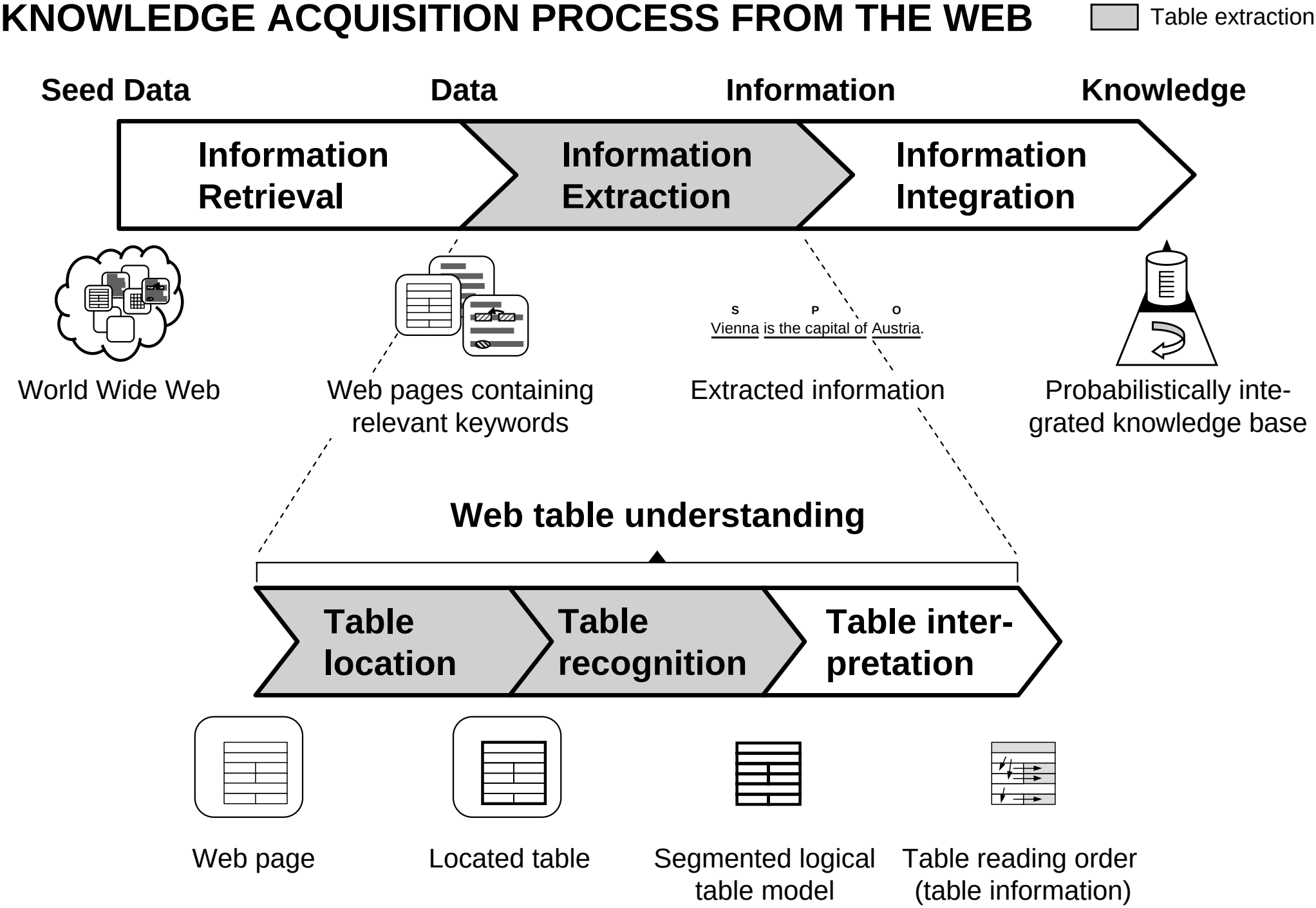
Table Extraction Using Spatial Reasoning on the CSS2 Visual Box Model

AAAI-06
Boston, July 18, 2006

Wolfgang Gatterbauer
Paul Bohunsky

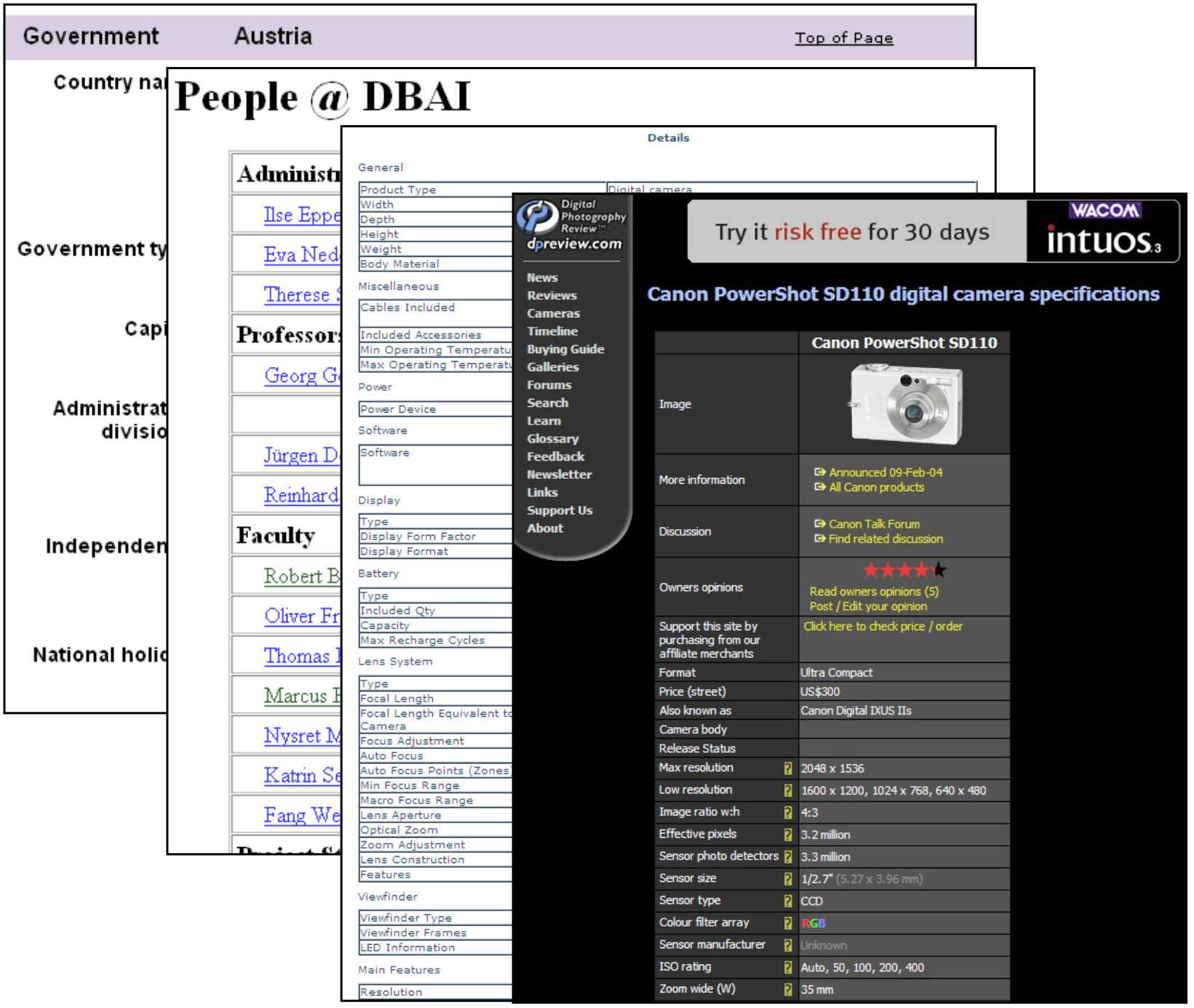
DBAI
Database & Artificial Intelligence Group
Vienna University of Technology

KNOWLEDGE ACQUISITION PROCESS FROM THE WEB



Source: Team VENTrec

EXAMPLE WEB INFORMATION IN TABLES



Government	Austria	Top of Page

Country nam

Government ty

Capi

Administrat division

Independen

National holid

People @ DBAI

Administr
Ilse Eppe
Eva Ned
Therese S
Professors
Georg G
Jürgen D
Reinhard
Faculty
Robert B
Oliver Fr
Thomas I
Marcus F
Nysret M
Katrin Se
Fang We

Details

General	
Product Type	Digital camera
Width	
Depth	
Height	
Weight	
Body Material	
Miscellaneous	
Cables Included	
Included Accessories	
Min Operating Temperatu	
Max Operating Temperatu	
Power	
Power Device	
Software	
Software	
Display	
Type	
Display Form Factor	
Display Format	
Battery	
Type	
Included Qty	
Capacity	
Max Recharge Cycles	
Lens System	
Type	
Focal Length	
Focal Length Equivalent to Camera	
Focus Adjustment	
Auto Focus	
Auto Focus Points (Zones	
Min Focus Range	
Macro Focus Range	
Lens Aperture	
Optical Zoom	
Zoom Adjustment	
Lens Construction	
Features	
Viewfinder	
Viewfinder Type	
Viewfinder Frames	
LED Information	
Main Features	
Resolution	

Digital Photography Review™ dpreview.com

News
Reviews
Cameras
Timeline
Buying Guide
Galleries
Forums
Search
Learn
Glossary
Feedback
Newsletter
Links
Support Us
About

Try it risk free for 30 days — WACOM intuos3

Canon PowerShot SD110 digital camera specifications

	Canon PowerShot SD110
Image	
More information	Announced 09-Feb-04 All Canon products
Discussion	Canon Talk Forum Find related discussion
Owners opinions	★★★★½ Read owners opinions (5) Post / Edit your opinion
Support this site by purchasing from our affiliate merchants	Click here to check price / order
Format	Ultra Compact
Price (street)	US$300
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	RGB
Sensor manufacturer	Unknown
ISO rating	Auto, 50, 100, 200, 400
Zoom wide (W)	35 mm

Source: Team VENTrec

PROBLEM OF CODE-BASED TABLE RECOGNITION

Web page with interesting tabular information

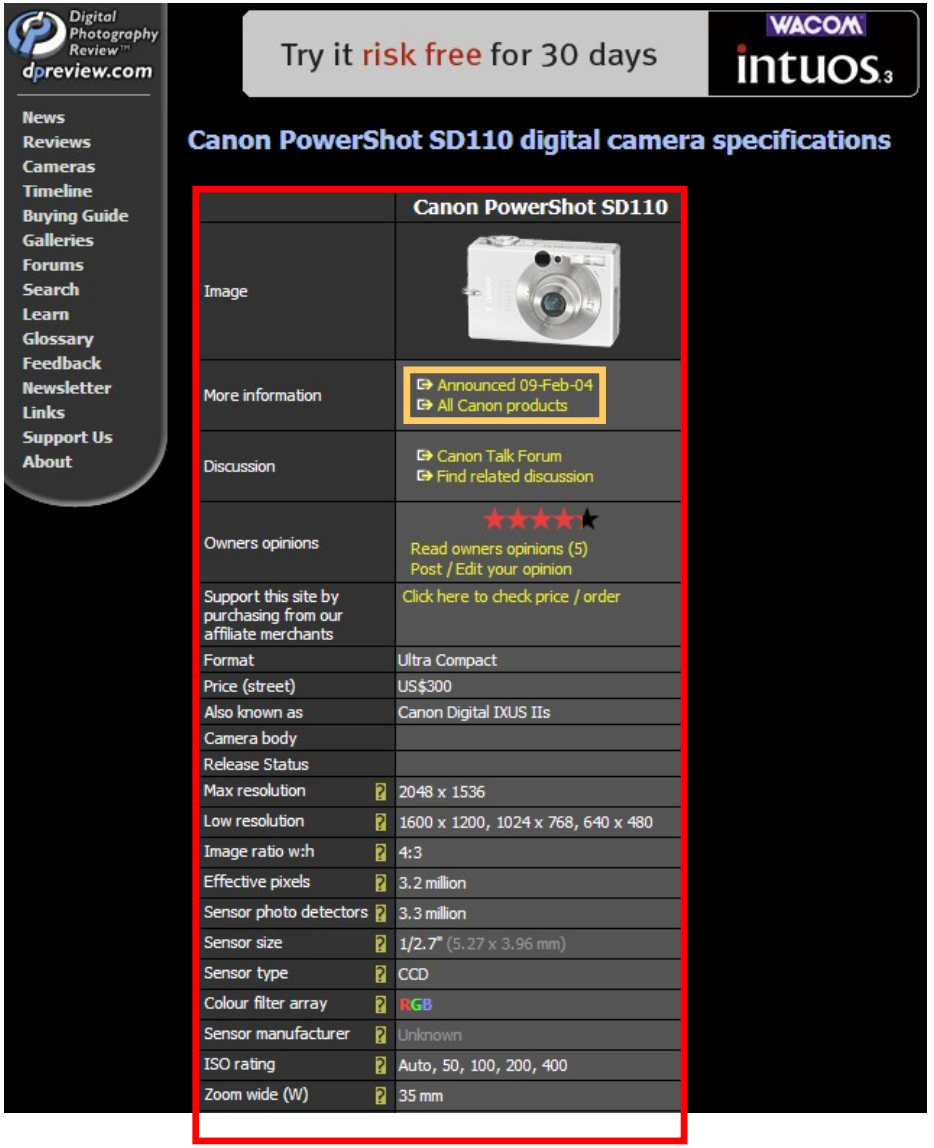


Table in the DOM tree

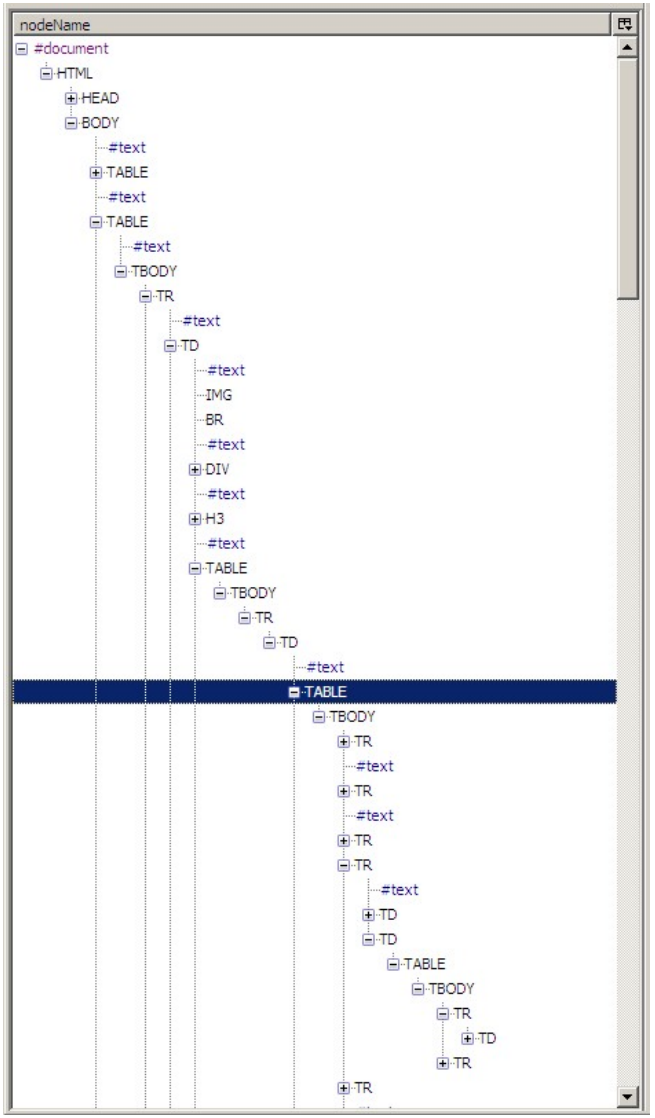


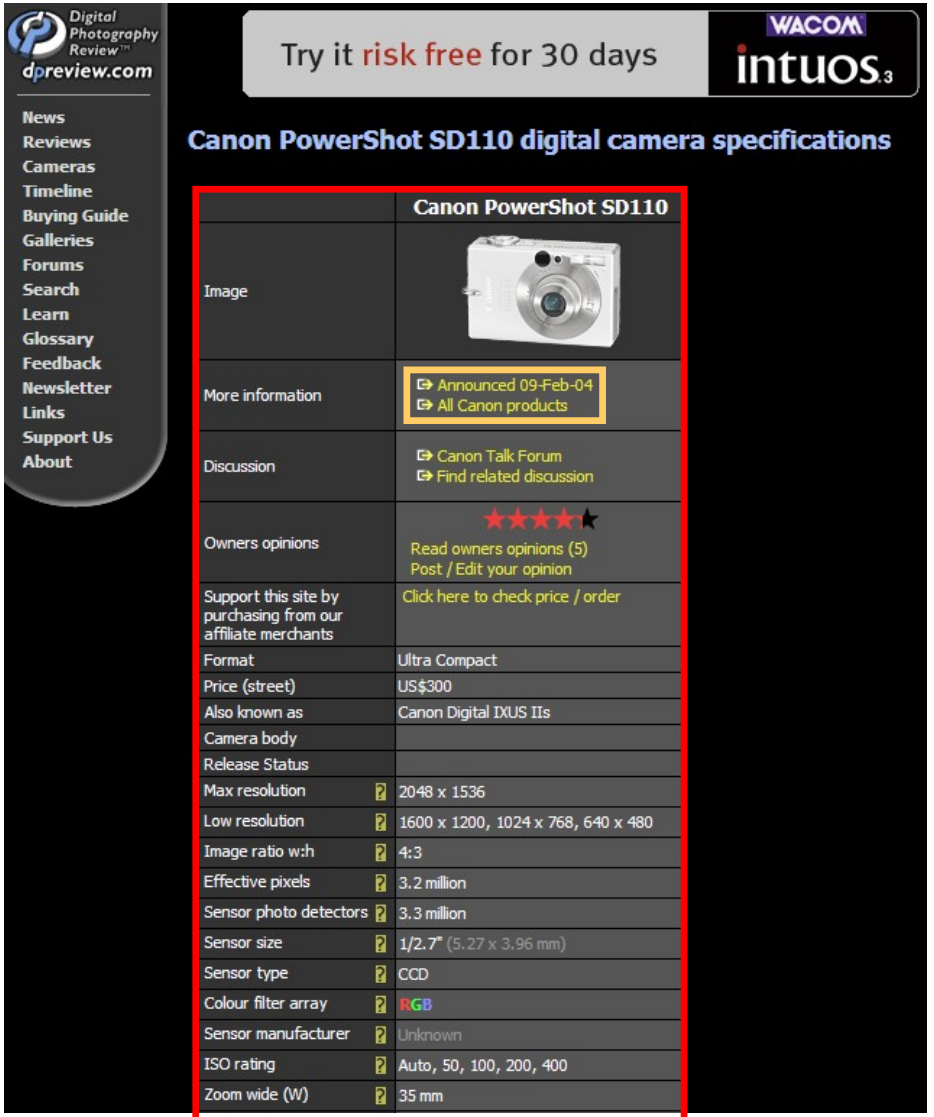
Table of interest

Nested table

Source: Team VENTrec

PROBLEM OF CODE-BASED TABLE RECOGNITION

Nested non-leaf <TABLE> tables



Digital Photography Review
dpreview.com

News
Reviews
Cameras
Timeline
Buying Guide
Galleries
Forums
Search
Learn
Glossary
Feedback
Newsletter
Links
Support Us
About

Try it risk free for 30 days

WACOM intuos3

Canon PowerShot SD110 digital camera specifications

	Canon PowerShot SD110
Image	
More information	Announced 09-Feb-04 All Canon products
Discussion	Canon Talk Forum Find related discussion
Owners opinions	★★★★½ Read owners opinions (5) Post / Edit your opinion
Support this site by purchasing from our affiliate merchants	Click here to check price / order
Format	Ultra Compact
Price (street)	US$300
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	RGB
Sensor manufacturer	Unknown
ISO rating	Auto, 50, 100, 200, 400
Zoom wide (W)	35 mm

Sequentially aligned <TABLE> tables

Details

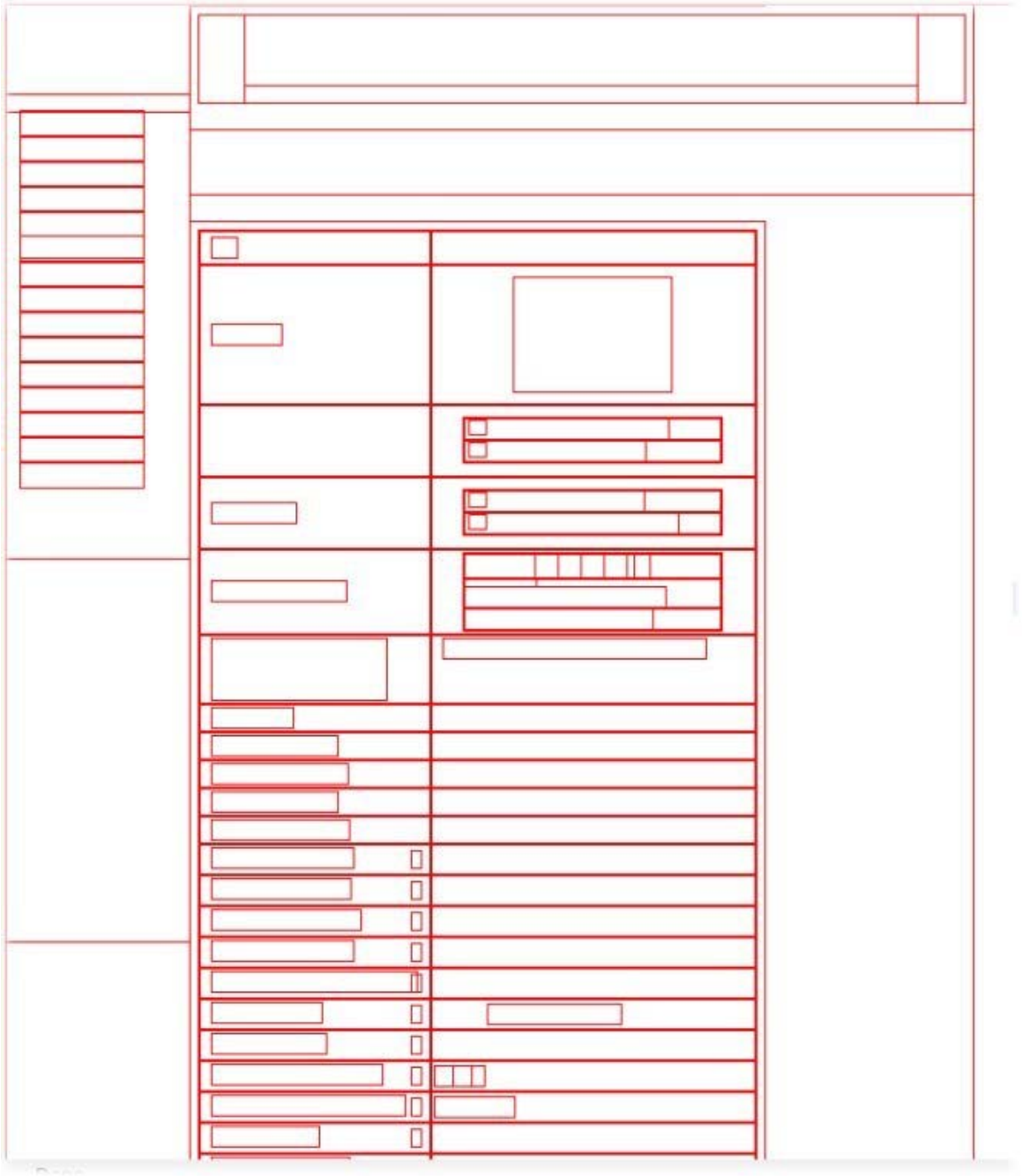
General	
Product Type	Digital camera
Width	8.5 cm
Depth	2.4 cm
Height	5.6 cm
Weight	165 g
Body Material	Stainless steel
Miscellaneous	
Cables Included	1 x A/V cable 1 x USB cable
Included Accessories	Wrist strap
Min Operating Temperature	0 °C
Max Operating Temperature	40 °C
Power	
Power Device	Battery charger - external
Software	
Software	Drivers & Utilities, Canon PhotoStitch, Canon ZoomBrowser EX, ArcSoft PhotoImpression, ArcSoft VideoImpression, Canon ImageBrowser
Display	
Type	LCD display - TFT active matrix - 1.5" - colour
Display Form Factor	Built-in
Display Format	118,000 pixels
Battery	
Type	1 x camera battery - rechargeable - Lithium Ion
Included Qty	1
Capacity	790 mAh
Max Recharge Cycles	300
Lens System	
Type	Zoom lens
Focal Length	5.4 mm - 10.8 mm
Focal Length Equivalent to 35mm Camera	35 - 70mm
Focus Adjustment	Automatic
Auto Focus	TTL contrast detection
Auto Focus Points (Zones)	9
Min Focus Range	47 cm
Macro Focus Range	10-47cm
Lens Aperture	F/2.8-3.9
Optical Zoom	2 x
Zoom Adjustment	Motorised drive
Lens Construction	6 group(s) / 6 element(s)
Features	Built-in lens shield, aspherical lens
Viewfinder	
Viewfinder Type	Optical - real-image zoom
Viewfinder Frames	Autofocus frame
LED Information	Flash ready, autofocus ready
Main Features	
Resolution	3.2 Megapixel

Source: Team VENTrec

VENTrec: Visualized Element Nodes Table RECognition

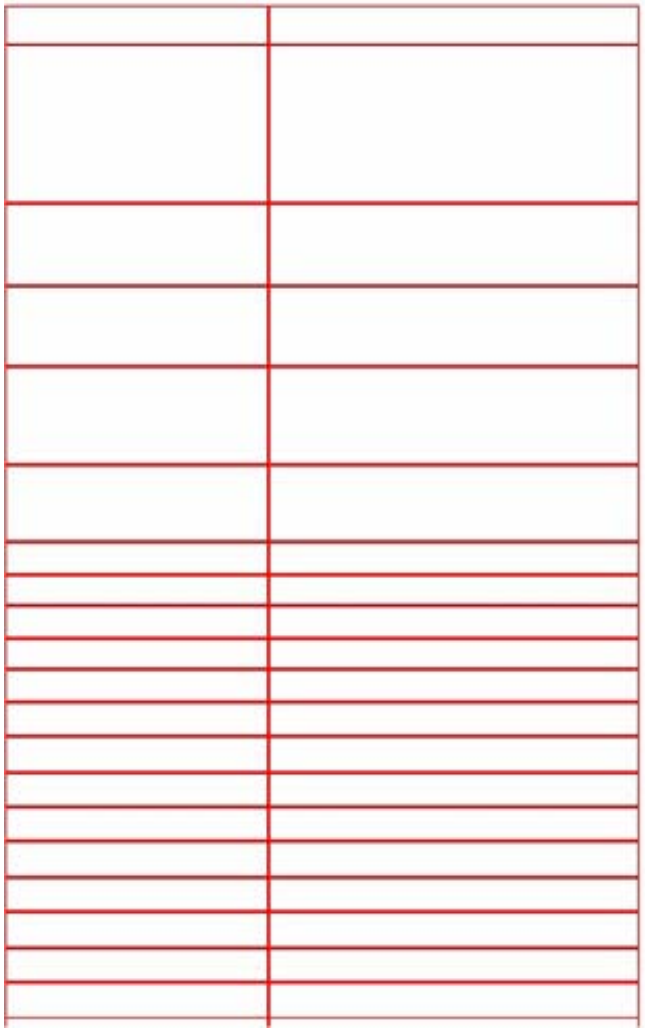
Source: Team VENTrec

VENTrec: Visualized Element Nodes Table RECognition



Source: Team VENTrec

VENTrec: Visualized Element Nodes Table RECognition



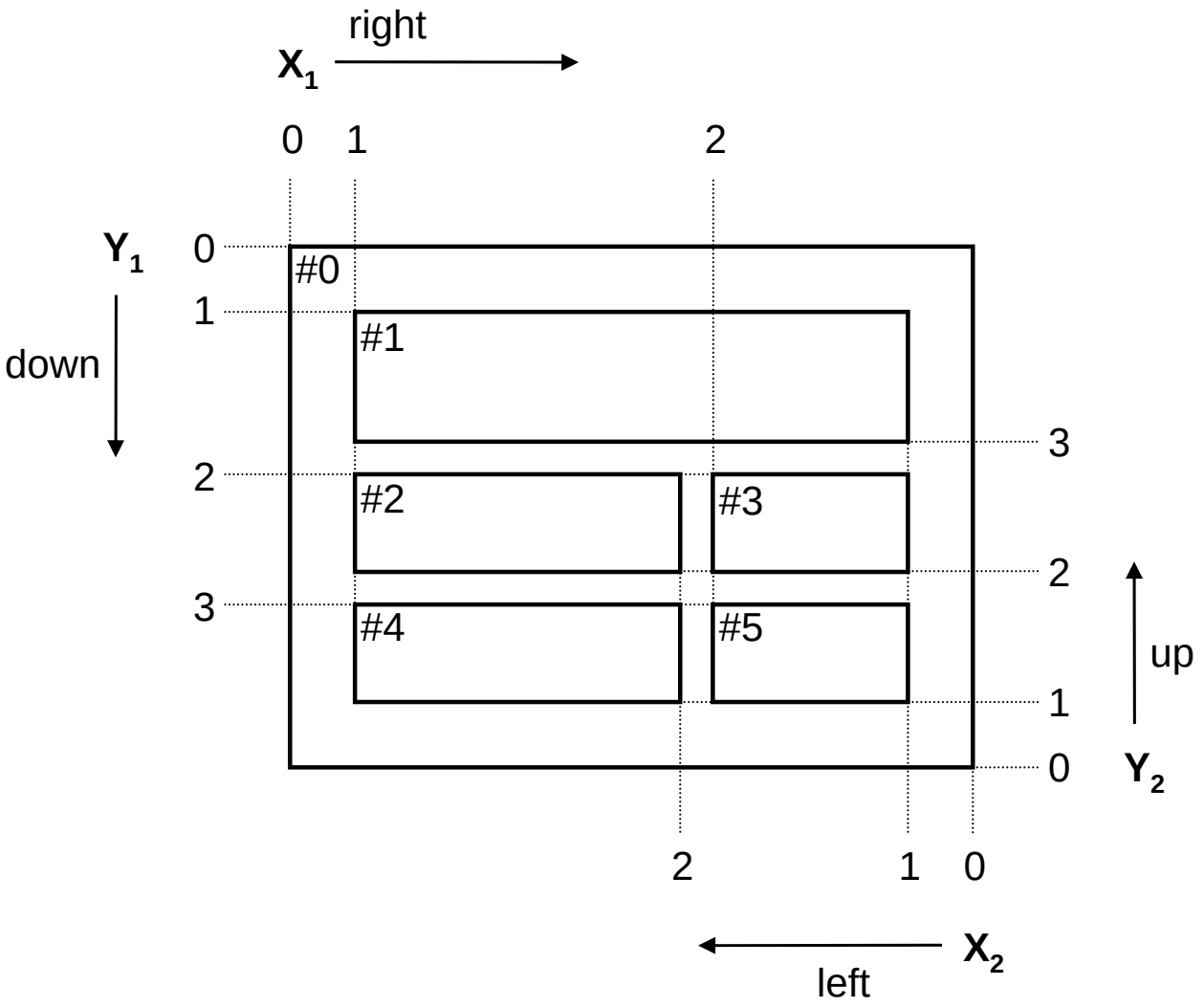
Source: Team VENTrec

VENTrec: Visualized Element Nodes Table RECognition

	Canon PowerShot SD110
Image	
More information	Announced 09-Feb-04 All Canon products
Discussion	Canon Talk Forum Find related discussion
Owners opinions	Read owners opinions (5) Post / Edit your opinion
Support this site by purchasing from our affiliate merchants	Click here to check price / order
Format	Ultra Compact
Price (street)	US$300
Also known as	Canon Digital IXUS IIs
Camera body	
Release Status	
Max resolution	2048 x 1536
Low resolution	1600 x 1200, 1024 x 768, 640 x 480
Image ratio w:h	4:3
Effective pixels	3.2 million
Sensor photo detectors	3.3 million
Sensor size	1/2.7" (5.27 x 3.96 mm)
Sensor type	CCD
Colour filter array	R G B
Sensor manufacturer	Unknown

Source: Team VENTrec

SUPERIMPOSED MINIMAL DOUBLE TOPOGRAPHICAL GRID (DTG)



Visualized Element Nodes

VEN	X_1	Y_1	X_2	Y_2
# 0	0	0	0	0
# 1	1	1	1	3
# 2	1	2	2	2
# 3	2	2	1	2
# 4	1	3	2	1
# 5	2	3	1	1

Grid structure

X_1	coord.
0	20
1	100
2	540

X_2	coord.
0	860
1	780
2	500

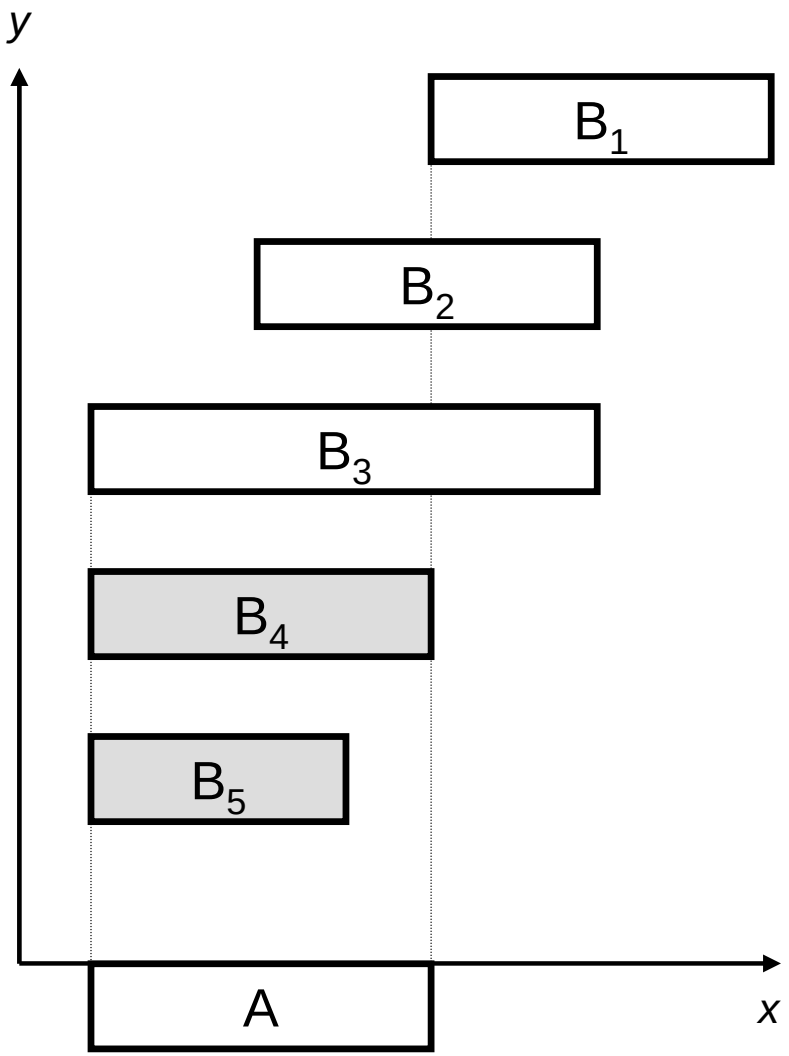
Y_1	coord.
0	20
1	100
2	300
3	460

Y_1	coord.
0	660
1	580
2	420
3	260

Source: Team VENTrec

ADJACENCY & ALIGNMENT

For each of the 4 dimensions, adjacent boxes are categorized according to 5 alignment relationships

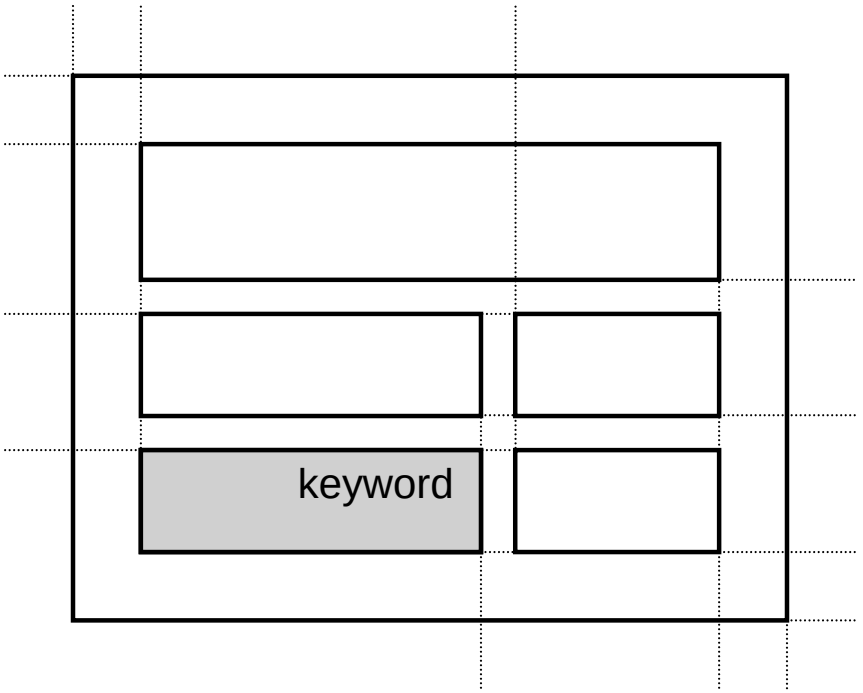


X-neighbor relationship of B(A)	**Allen's intervall relations** (Allen 1983; Aiello 2002)
B_1 = **no** neighbor (A)	13 A before B 12 B before A 11 A meets B 10 B meets A
B_2 = **step** neighbor (A)	9 A overlaps B 8 B overlaps A
B_3 = **bigger** neighbor (A)	7 A during B 6 A starts B 5 A finishes B
B_4 = **twin** neighbor (A)	4 B equal A
B_5 = **smaller** neighbor (A)	3 B finishes A 2 B starts A 1 B during A

2 neighbor relation-ships are of relevance for the expansion algorithm

Source: Team VENTrec

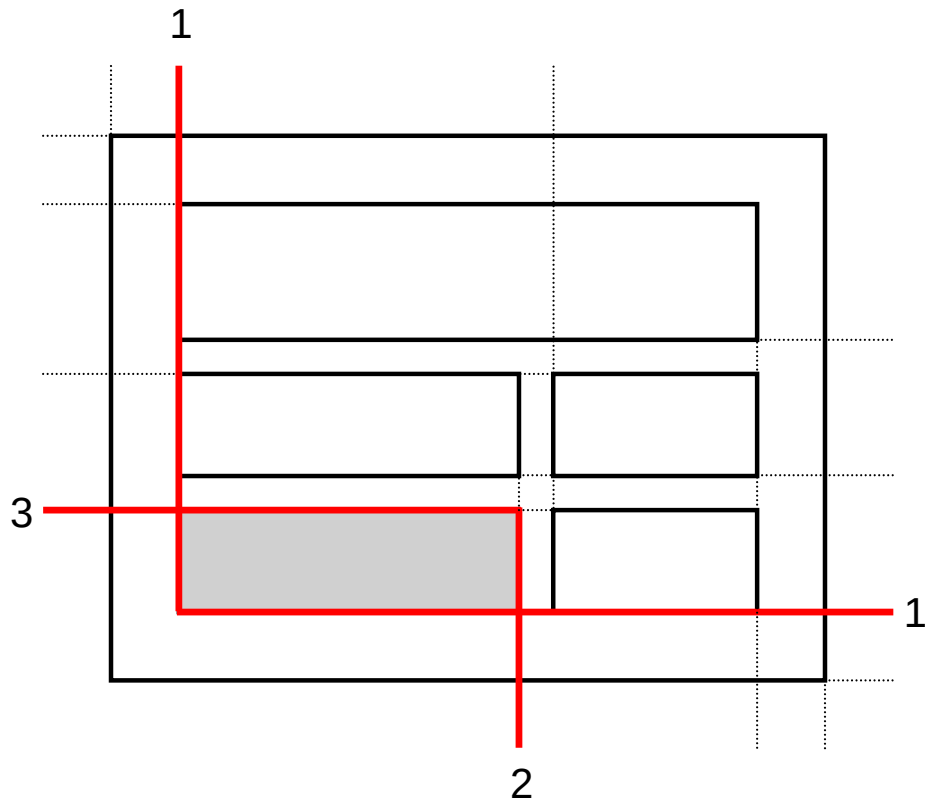
WORKING OF THE EXPANSION ALGORITHM



Keyword projected into Element Nodes

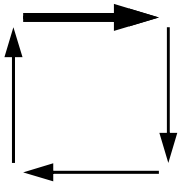
Source: Team VENTrec

WORKING OF THE EXPANSION ALGORITHM



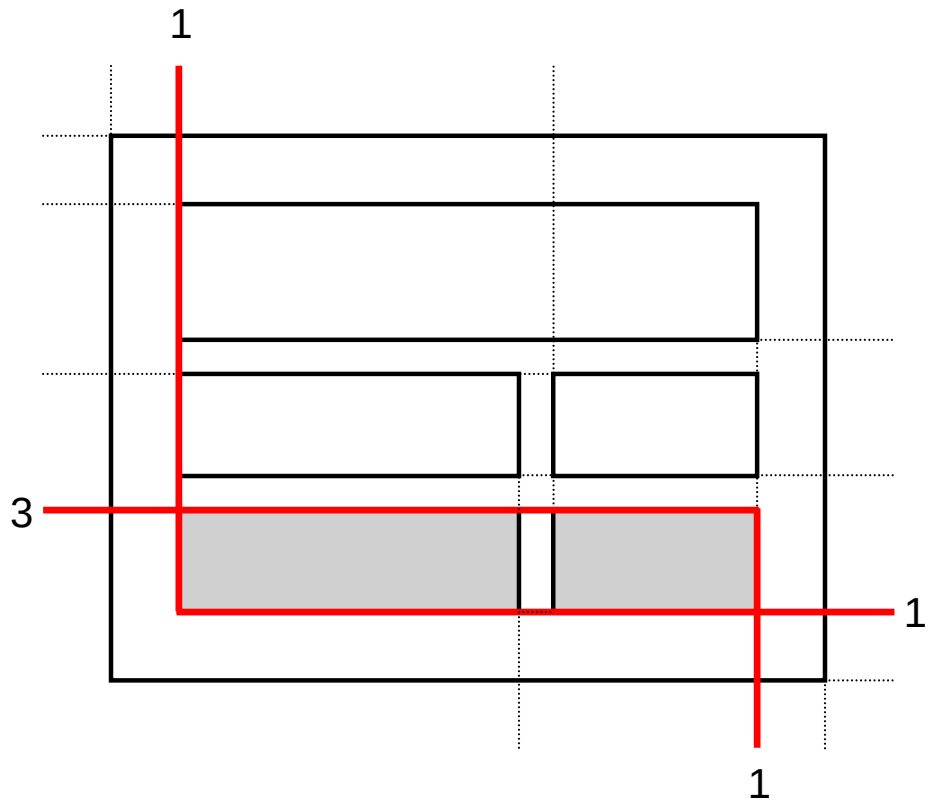
Keyword projected into Element Nodes

Circulating HyperBox Expansion Algorithm



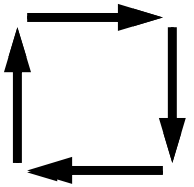
Source: Team VENTrec

WORKING OF THE EXPANSION ALGORITHM



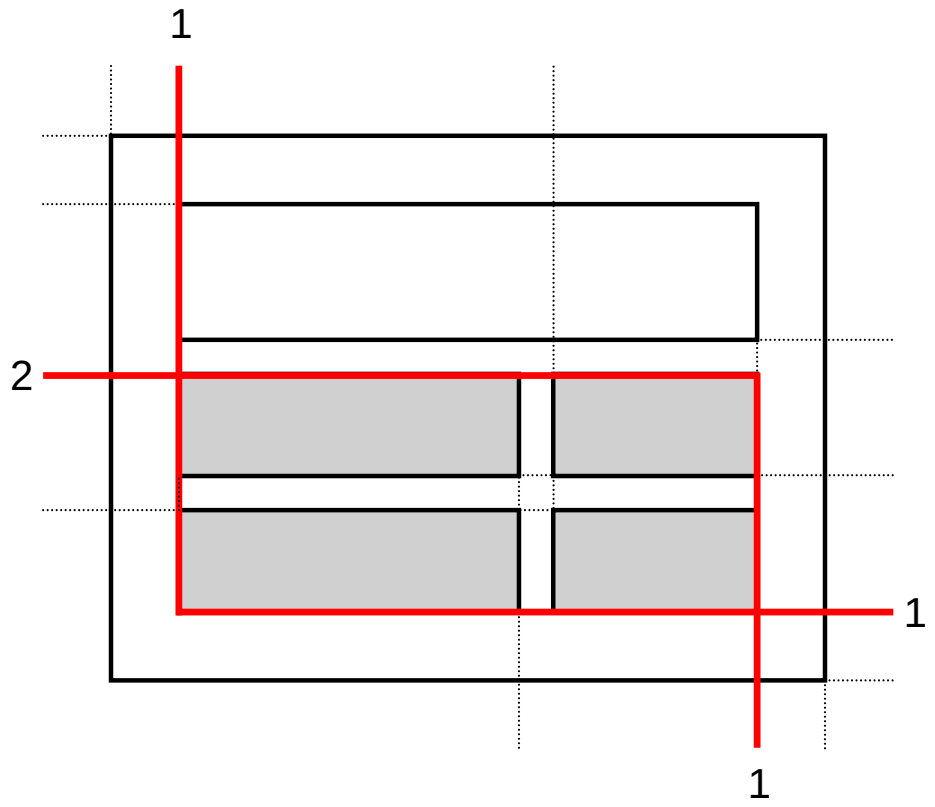
Keyword projected into Element Nodes

Circulating HyperBox Expansion Algorithm



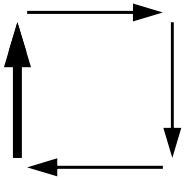
Source: Team VENTrec

WORKING OF THE EXPANSION ALGORITHM



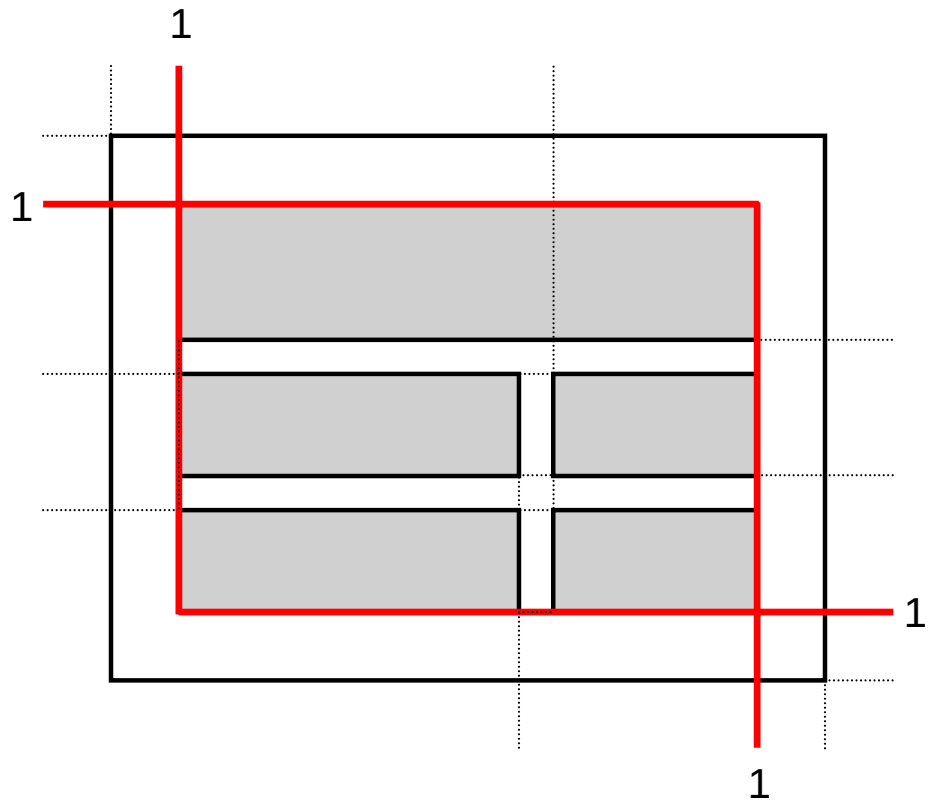
Keyword projected into Element Nodes

Circulating HyperBox Expansion Algorithm



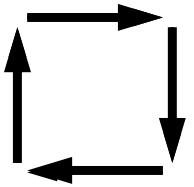
Source: Team VENTrec

WORKING OF THE EXPANSION ALGORITHM



Keyword projected into Element Nodes

Circulating HyperBox Expansion Algorithm



STOP

Source: Team VENTrec

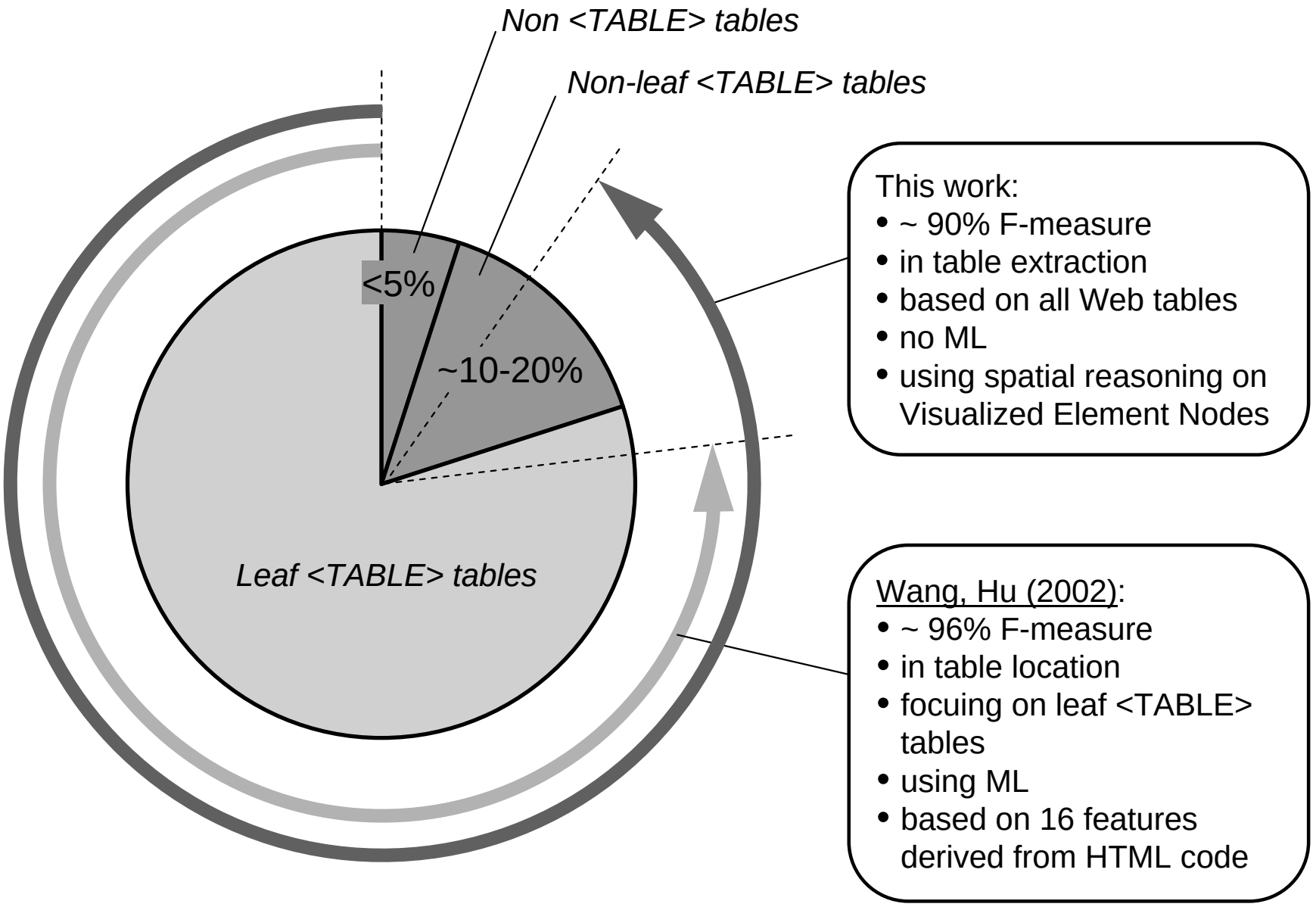
TIME COMPLEXITY

n ... # of element nodes
k ... # of keyword appearances

Positional data gathering	n
Circulating Expansion Algorithm	$k\sqrt{n}$
System	$n + k\sqrt{n}$

PERFORMANCE COMPARISON

In Percent of Web tables

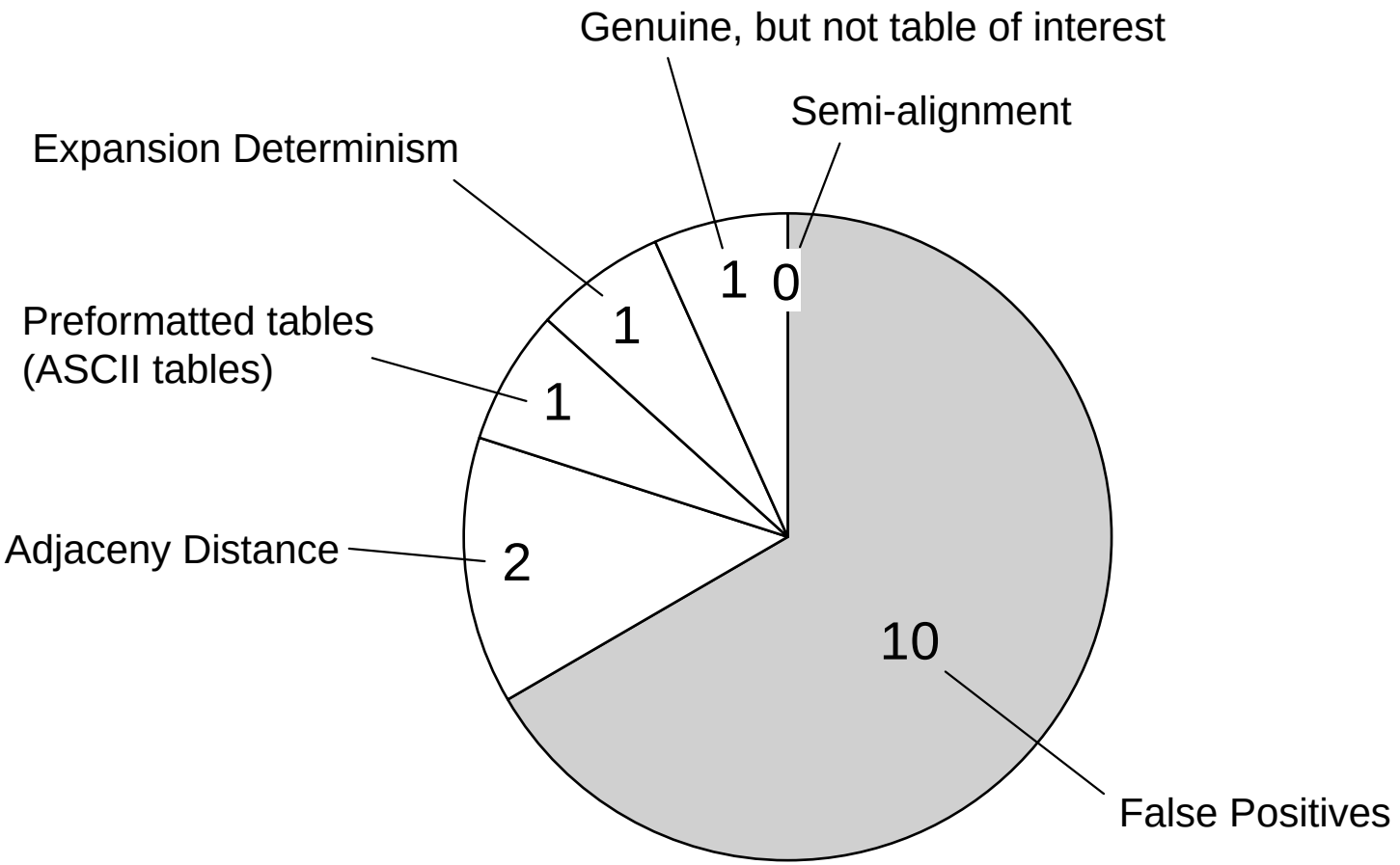


Source: Team VENTrec

DOMINANT REASONS FOR WRONG RESULTS

Total = 15

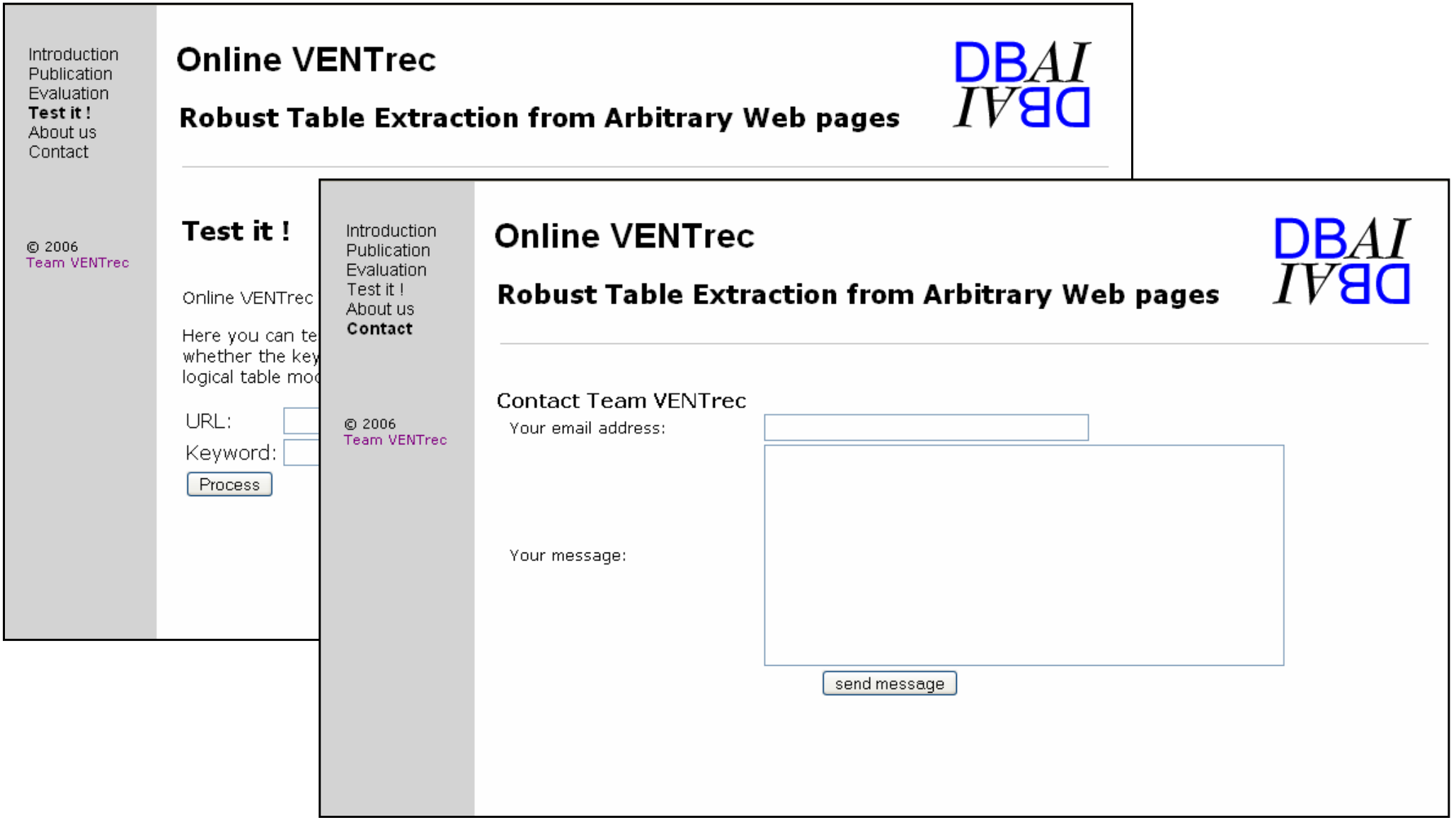
~10% of quantitative test set



Source: Team VENTrec

ONLINE VENTrec

Test it: http://education.dbai.tuwien.ac.at/ventrec/



Source: Team VENTrec

SUMMARY

Web table understanding

- As one approach for automated knowledge acquisition from the Web
- Table extraction from HTML code, however, is difficult
- Alternative: extracting tables from *rendered Web pages*

Our approach

- Reasons on *spatial relationships*
- Between *Visualized Element Nodes*
- Working upon a *Double Topographical Grid* Struture

Experiments

- ~90% F-measure yet without any form of learning
- Includes ~20% of Web tables missed by previous approaches

Source: Team VENTrec

NEXT STEPS

VENTrec II

- Objectivation of table model
- Ambiguity resolution

} Theoretical foundations

- Heuristics for non-alignment
- ML-improved version
- Methodology for web table ground truthing

Table interpretation

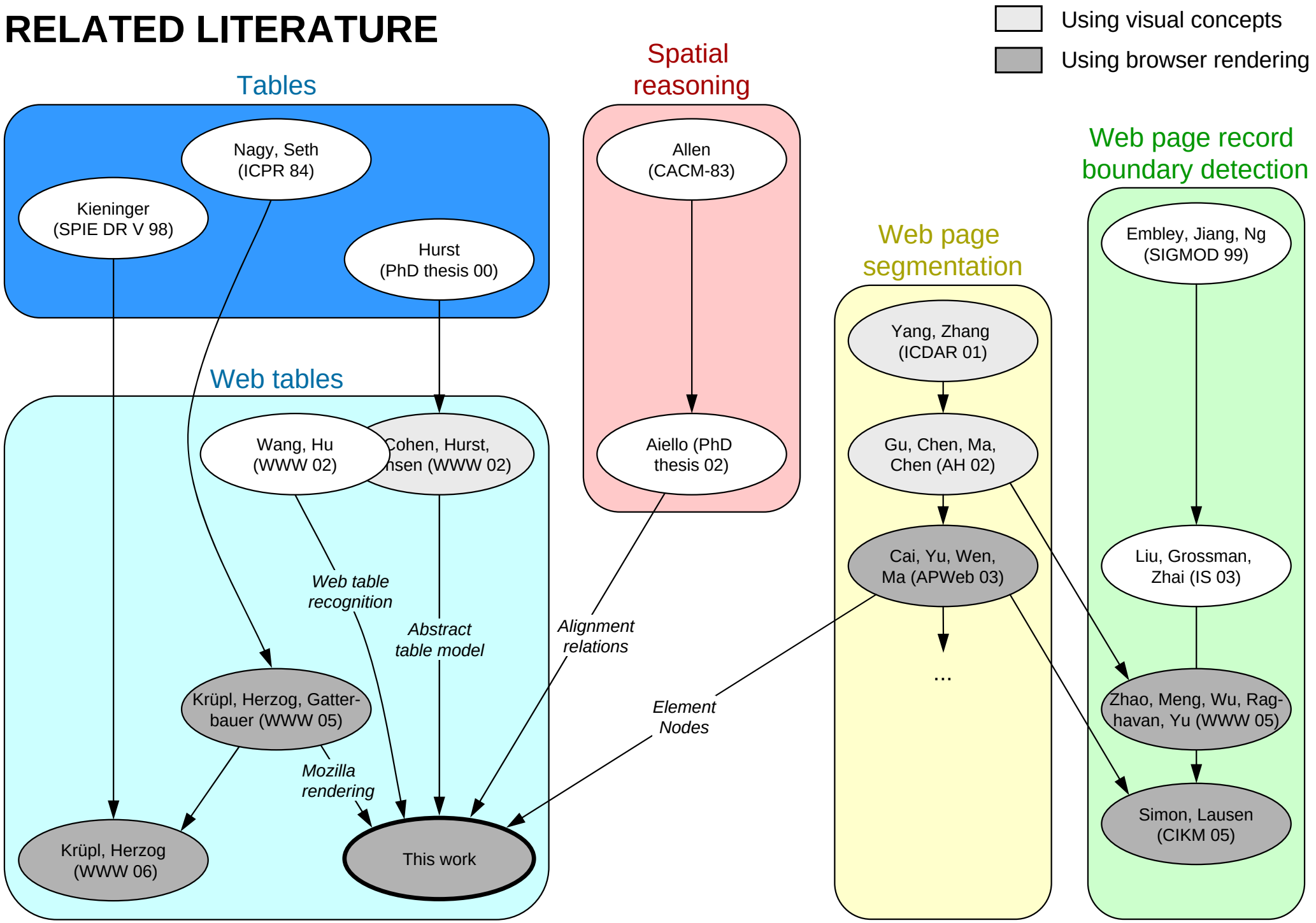
- Agent-based bottom-up learning of reading order = extraction of contained information

Visit us on the Web: http://education.dbai.tuwien.ac.at/ventrec/

Source: Team VENTrec

Backup

RELATED LITERATURE



Source: Team VENTrec

RELATED LITERATURE

- Aiello, M. 2002. *Spatial Reasoning: Theory and Practice*. Ph.D. thesis, ILLC, University of Amsterdam.
- Allen, J.F. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM* 26(11):832-843.
- Cai, D.; Yu, S.; Wen, J.-R.; and Ma, W.-Y. 2003. Extracting content structure for web pages based on visual representation. In *Proc. APWeb'03*, 406-417. Springer.
- Cohen, W.W.; Hurst, M.; and Jensen, L.S. 2002. A flexible learning system for wrapping tables and lists in HTML documents. In *Proc. WWW'02*, 232-241. ACM.
- Embley, D.W.; Jiang, Y.S.; and Ng, Y.-K. 1999. Record-Boundary Discovery in Web Documents. In *Proc. SIGMOD'99*, 467-478, ACM.
- Gu, X; Chen, J; Ma, W.-Y.; and Chen, G. 2002. Visual Based Content Understanding towards Web Adaptation. In Proc. AH'02, 164-173, Springer.
- Hurst, M. 2000. *The Interpretation of Tables in Texts*. PhD. thesis, University of Edinburgh.
- Kieninger, T. G. 1998. Table structure recognition based on robust block segmentation. *Proc. SPIE Vol. 3305, Document Recognition V: 22-32*.
- Krüpl, B.; Herzog, M.; and Gatterbauer, W. 2005. Using visual cues for extraction of tabular data from arbitrary HTML documents. In *Poster Proc. WWW'05*, 1000-1001. ACM.
- Krüpl, B., and Herzog, M. 2006. Visually guided bottom-up table detection and segmentation in web documents. In *Poster Proc. WWW'06*. 933-934. ACM.
- Liu, B; Grossman, R. L.; Zhai, Y. 2003. Mining data records in Web pages. In *Proc. SIGKDD'03*, 601-606, ACM.
- Nagy, G., and Seth, S.C. 1984. Hierarchical representation of optically scanned documents. In *Proc. ICPR'84*, 347-349. IEEE.
- Simon, K., and Lausen, G. 2005. ViPER: augmenting automatic information extraction with visual perceptions. In *Proc. CIKM'05*, 381-388. ACM
- Wang, Y., and Hu, J. 2002. A machine learning based approach for table detection on the Web In *Proc. WWW'02*, 242-250. ACM.
- Yang, Y. and Zang, H. 2001. HTML Page Analysis Based on Visual Cues. In Proc. ICDAR'99, 859-864, IEEE.
- Zhao, H.; Meng, W.; Wu, Z.; Raghavan, V.; and Yu, C. 2005. Fully automatic wrapper generation for search engines. In *Proc. WWW'05*, 66-75. ACM

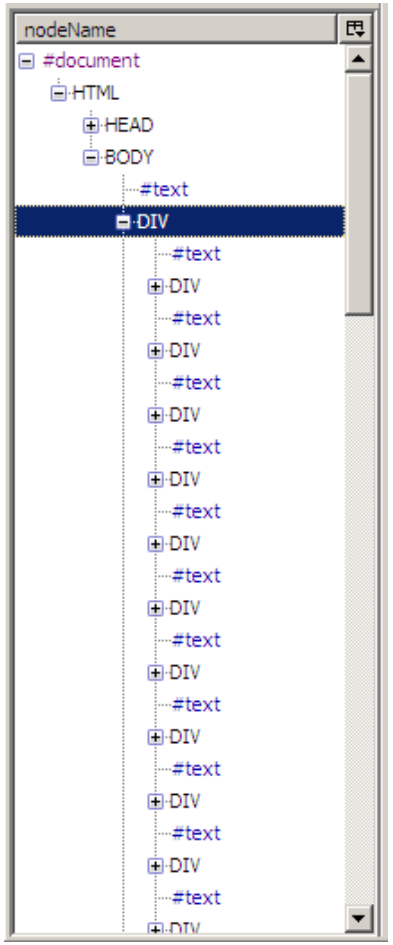
This work:

- Gatterbauer, W., and Bohunsky, P. 2006. Table extraction using spatial reasoning on the CSS2 visual box model. In *Proc. AAAI'06*, 1313-1318. AAAI, MIT Press.

Source: Team VENTrec

HTML RENDERING AS NON-INJECTIVE MAPPING

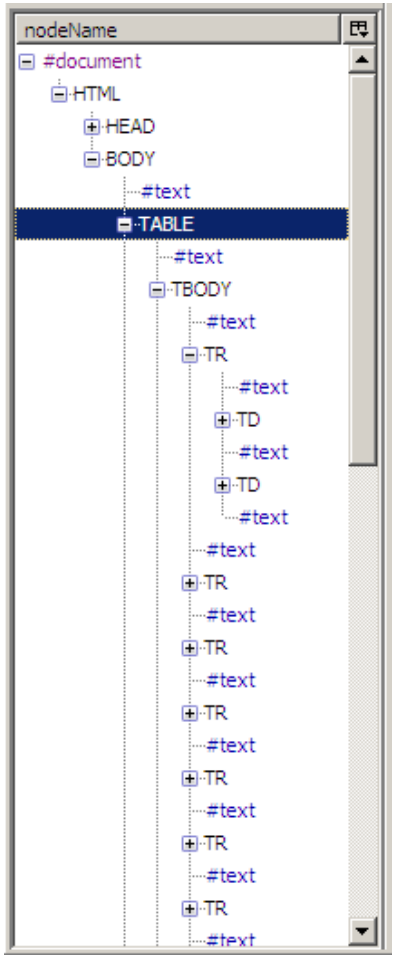
<DIV> table



Visual rendering

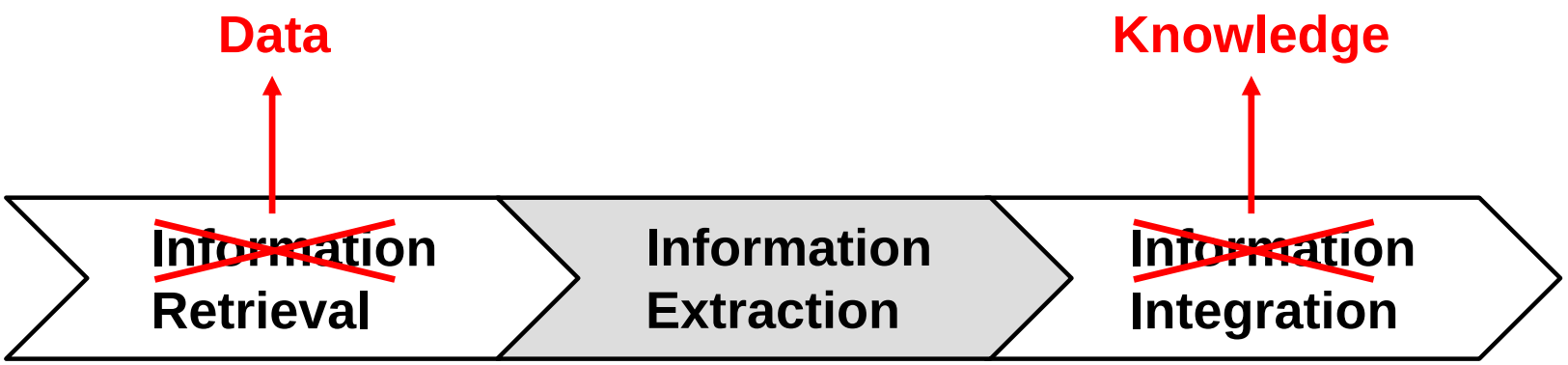
Athlete	Country
AAMODT Kjetil Andre	NOR
ABRAMASHVILI Iason	GEO
ACTON Brigitte	CAN
AGUIRRE Facundo	ARG
AHUJA Neha	IND
ALBRECHT Daniel	SUI
ALCOTT Chemmy	GBR
ALIEVA Olesja	RUS
ANGUITA Daniela	CHI
ANTOR Alex	AND
ARNHOLD Mirella	BRA
AUFDENBLATTEN Fraenzi	SUI
BABUSIAK Jaroslav	SVK
BANK Ondrej	CZE
BARAHONA Noelle	CHI

<TABLE> table



Source: Team VENTrec

KNOWLEDGE ACQUISITION PROCESS



?

Source: http://education.dbai.tuwien.ac.at/wie/WS05/