

DISCOVERING RECORD BOUNDARIES

based on 1999's paper by
Embley / Jiang / Ng

ABOUT RENE KIESLER

- Consultant, focusing on Portals / CMS
- Bachelor for Software Engineering
- Master study for Information / Knowledge Management
- see www.kiesler.at



MOTIVATION

Why would someone want to detect record boundaries?

- tons of lists online (classifieds, member listings, ...)
- mostly in the web, not in a database

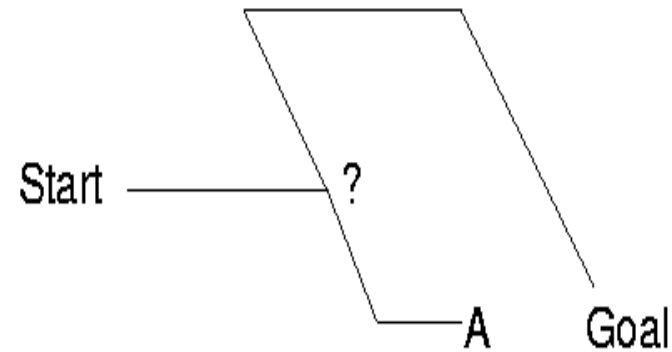
HEURISTIC 1/3

Embley / Liang / Ng chose the heuristic approach to finding boundaries.

- Origin: Greece
- „Eureka“ – I find
- different Meanings in Psychology, Philosophy, Law and Computer Science

HEURISTICS 2/3

- Psychology: Rules of Thumbs
- Philosophy: Describe something with something else
- Law: if case-by-case basis not feasible

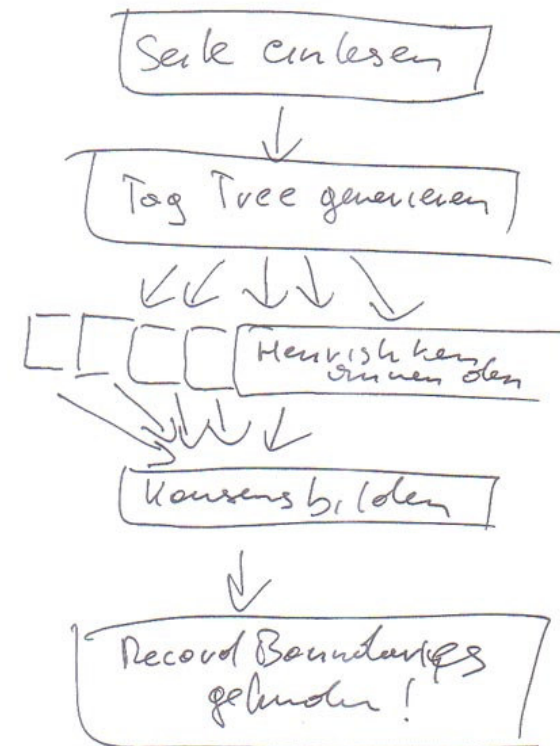


HEURISTICS 3/3

- Computer Science: gain speed at cost of precision
- heuristics give good – but not perfect results (estimates)
- popular examples: shortest-path and Artificial Intelligence

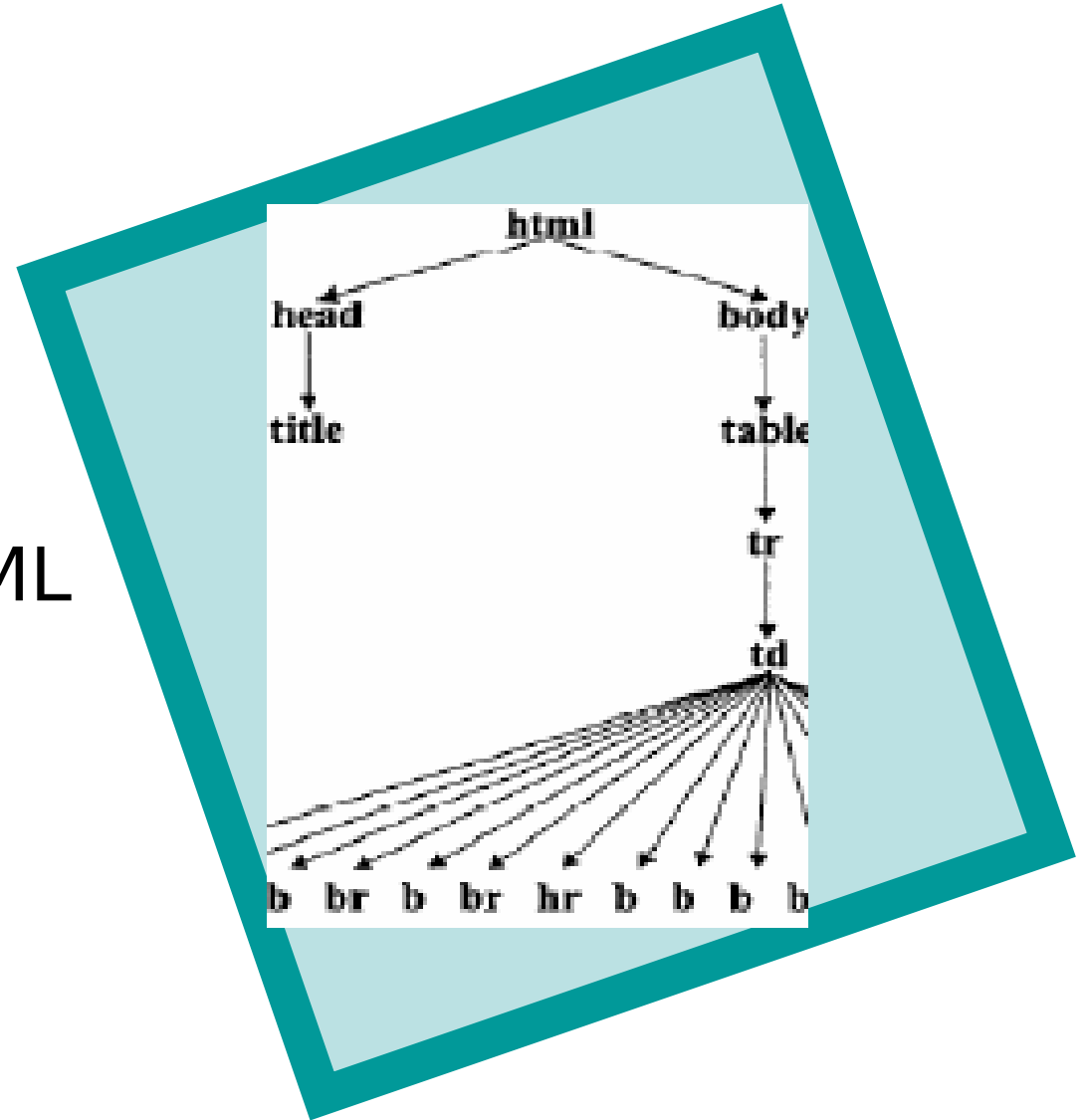
CONCEPT OF DRB

- parse single page
- generate tag tree
- use heuristics
- build consensus
- finished



HTML TO TREE

- HTML has an inherent tree structure
- root element HTML
- nested tags



USED HEURISTICS 1/2

- HT: Highest Count Tag
count fan-outs
- IT: Identifiable Separator Tags
hr, tr, td, a, table, p, br, ...
- SD: Standard Deviation
how many chars are between occurrences
of hr / b / br / ...?

USED HEURISTICS 2/2

- RT: Repeating-Tag Pattern

can we find reoccurring constructs like

`<hr>title</hr>?`


- OM: Ontology Matching

can we detect words that fit in our given ontology?

can be skipped, if no Ontology available

EXAMPLE: TUWIS 1/2

- after searching for „informatik“, we get this list
- traditional table
- let's look at the html code.



The screenshot shows the TUWIS++ LVA search results page. The header includes navigation links like 'Übersicht', 'Institute', 'Suche im Lehrangebot', 'Studienpläne', 'Hörsaalbelegung', and 'Hilfe'. There is a search bar with 'TUWIS++:' and a 'login' button. The main content area displays the search results for 'informatik' in a table format.

Ergebnis der Suche nach: informatik
Zurück zur Suche

Nr	Typ	Titel	Sem	Stunden	Vortragender
102.185	SE	AKTHI Neuroinformatik	2004W	2.0	CHRISTIAN
102.316	SE	Neuroinformatik	2005S	2.0	CHRISTIAN
104.105	VU	Theoretische Informatik 2	2004W	3.0	BAAZ
105.070	VO	Praktikum aus Operations Research 1 (für Wirtschaftsinformatiker)	2004W	2.0	FEICHTINGER
105.070	VO	Praktikum aus Operations Research 1 (für Wirtschaftsinformatiker)	2005S	2.0	FEICHTINGER
106.009	PR	Informatik-Praktikum II	2004W	10.0	WEINMÜLLER
106.009	PR	Informatik-Praktikum II	2005S	10.0	WEINMÜLLER
106.037	PR	Informatikpraktikum I	2004W	10.0	WEINMÜLLER
106.037	PR	Informatikpraktikum I	2005S	10.0	WEINMÜLLER
106.057	VO	Einführung in die Informatik für TM	2004W	2.0	AUZINGER
107.144	PR	Informatikpraktikum I	2004W	10.0	DUTTER
107.144	PR	Informatikpraktikum I	2005S	10.0	DUTTER
107.256	UE	Statistik und Wahrscheinlichkeitstheorie f. InformatikerInnen	2004W	1.0	KUSOLITSCH
107.285	VO	Statistik und Wahrscheinlichkeitstheorie f. InformatikerInnen	2004W	2.0	KUSOLITSCH
108.034	VU	Theoretische Informatik 2	2005S	3.0	BAAZ
108.035	VU	Theoretische Informatik 1	2004W	4.0	KUJICH
108.036	VO	Theoretische Informatik	2005S	2.0	KUJICH
108.037	UE	Theoretische Informatik, Übung	2005S	1.0	URBANER
108.038	VU	Theoretische Informatik 2	2005S	3.0	BAAZ
110.040	VO	Operations Research 1 (für Wirtschaftsinformatik)	2005S	2.0	HAUNSCHMIED
110.041	PR	Operations Research 1 (für Wirtschaftsinformatik)	2005S	2.0	MEHLMANN
117.017	VO	Geometrie für Informatiker	2004W	2.0	POTTMANN

EXAMPLE: TUWIS 2/2

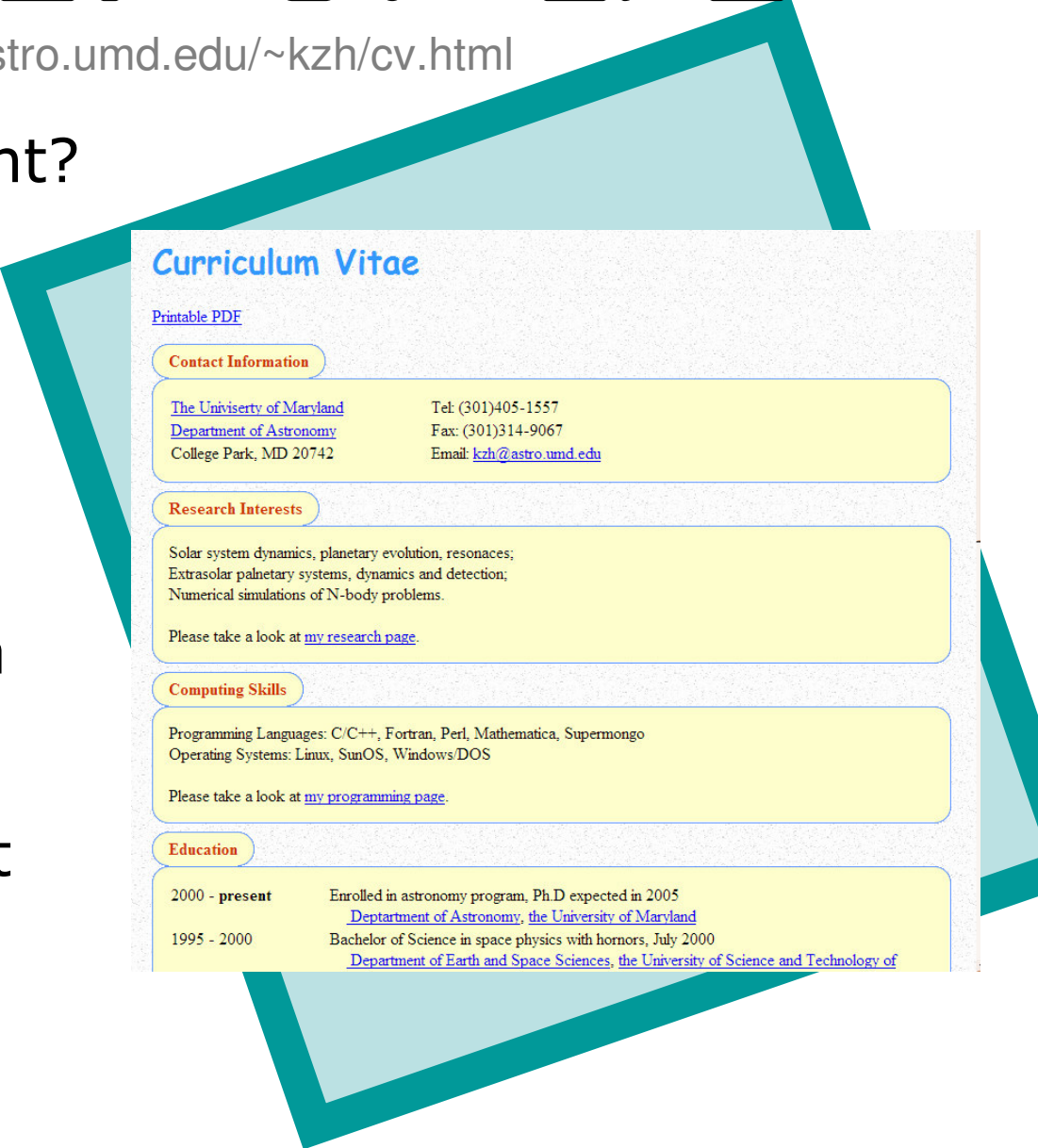
- large amount of TDs
- not-so-large amount of TRs
- only one big table
- a couple of meta-tags
- a bit of JavaScript
- perfect for our algorithm

```
<tr>
  <td>106.009</td>
  <td>PR</td>
  <td title="Status der
LVA: A wenn abgesagt,
sonst leer"></td>
  <td nowrap><a
href="http://tuwis.tuwien.
ac.at/zope/_ZopeId/60243
864A1y1VIWztxE/tpp/lv/lv
a_html?num=106009&sem
=2005S">Informatik-
Praktikum II</a></td>
  <td>2005S</td>
  <td>10.0</td>
  <td>WEINMÜLLER</td>
</tr>
```

EXAMPLE: CV 1/2

<http://www.astro.umd.edu/~kzh/cv.html>

- same structure, right?
- wrong!
- unstructured, no reoccurring patterns
- a bunch of tables, a bunch of divs
- heuristics would get confused



Curriculum Vitae

[Printable PDF](#)

Contact Information

The Univiserty of Maryland	Tel: (301)405-1557
Department of Astronomy	Fax: (301)314-9067
College Park, MD 20742	Email: kzh@astro.umd.edu

Research Interests

Solar system dynamics, planetary evolution, resonances;
Extrasolar palnetary systems, dynamics and detection;
Numerical simulations of N-body problems.

Please take a look at [my research page](#).

Computing Skills

Programming Languages: C/C++, Fortran, Perl, Mathematica, Supermongo
Operating Systems: Linux, SunOS, Windows/DOS

Please take a look at [my programming page](#).

Education

2000 - present	Enrolled in astronomy program, Ph.D expected in 2005 Department of Astronomy, the University of Maryland
1995 - 2000	Bachelor of Science in space physics with honors, July 2000 Department of Earth and Space Sciences, the University of Science and Technology of

EXAMPLE: CV 2/2

- 11 differently nested DIVs
- 33 (!) completely different tables
- almost no common structure
- a couple of BRs (maybe these would work?)
- not ideal for our algorithm

```
<table><tr><td>
<table><tr><td></td></tr>
>
<tr><td></td></tr>
<tr><td></td></tr>
</table></td>
<td><b>Computing
Skills</b></td>
<td><table>
<tr><td ></td></tr>
<tr><td></td></tr>
<tr><td></td></tr>
</table></td></tr>
</table>
```

NOW WHAT?

- find consensus
- present record separator
- we don't provide extraction to the database
- no crawling either