Comparison of two unsupervised information extraction techniques
Extracting Structured Data from Web Pages (#22)
and
Applying Pattern Mining to Web Information Extraction(#21)

Tobias Dönz (0226173)
Jeremy Solarz (0330879)

PS Web Information Extraction, SS 2005
"Proseminar Wissenschaftliches Arbeiten"
at the
Database and Artificial Intelligence Group,
Institute of Information Systems,
Computer Science Department,
Vienna University of Technology

April 22, 2005
Similarities

- Unsupervised
- No user input, training sets or manually generated rules
- No annotation
- Cluster tokens that have the same occurrence-counts
<table>
<thead>
<tr>
<th><strong>ExAlg</strong></th>
<th><strong>PAT tree</strong></th>
</tr>
</thead>
<tbody>
<tr>
<td>algorithm</td>
<td>data structure</td>
</tr>
<tr>
<td>even for complex, nested pages</td>
<td>for web pages with flat structure</td>
</tr>
<tr>
<td>extracts data values</td>
<td>generates extraction rules</td>
</tr>
</tbody>
</table>
Use approach with PAT tree for "simple" extraction tasks

- Simple, fast
- High extraction rates

Use ExAlg for more complex pages with nested tables or lists

- Good extraction rates for that page type
- Tested with many different page collections
Thanks for your attention

Questions?