

# A Gateway From HTML to XML

A Gateway From HTML to XML

by

Tao Fu & Mengchi Liu

Friedrich Dimmel, 0302230

# Overview

- The Gateway's purpose is, to transfer HTML web sites into an *XML structure*
- Heavily using *layout markups* as design patterns
- Algorithm looks at HTML page in user-view

# System overview—DOM

- HTML/ XML is a *tree* of element nodes with *attributes* and *text nodes*
- Gateway transfers *HTML DOM* into *T- DOM*
- In the end transformation from *T- DOM* into *XML- DOM*

# System overview—Web data

- meaningful values in source code is *alphanumeric*
- style attributes do not make sense for extracted data
- *href* and *alt* attributes are useful
- split content into *content blocks* and *topics*
- emphasized content is more important

# System overview—Extraction

- HTML document gets XML document
- *Node Classification*: splitting up HTML nodes into *records* and *delimiters*
- *Content Partition*: group consistent records into *content blocks*
- *Topic Aggregation*: find out *subjects* of records
- Transform T-DOM into XML DOM

# Content Partition

- recover information blocks as in HTML document's authors intention
- detecting *block boundaries*
- *explicit block boundaries*: hr, table, div, p, blockquote, pre, form, ul, ol
- *implicit boundaries*: delimiters intended as gaps between content blocks: br

# Topic Aggregation

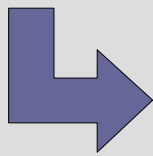
- clusters consistent records or blocks into specific topics
- normally a topic block is emphasized as *bold*, extended with a „:“ or content belonging to a specific topic is *intended*
- *static clustering*: headings from h1 to h6 for larger fonts
- *dynamic clustering*: calculate *similarity distance* with *Frequent Structure Mining*

# Table normalization

- *multi-dimensional tables* must be separated into *cells*, *rows* and *columns*
- cells merged with *colspan* or *rowspan* must be restored and content be copied
- headings and content will be *merged* into single rows
- headings are marked bold to differentiate them as *topics*

# Table normalization (cont.)

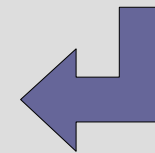
Faculty	Office	
	Room	Tel.
Will, Peter	HC1202	4012
Shaw, Erin	HC1225	3416
Frank, Martin		1677
Rogers, Craig	HC1240	2208



Faculty	Office	Office
Faculty	Room	Tel.
Will, Peter	HC1202	4012
Shaw, Erin	HC1225	3416
Frank, Martin	HC1225	1677
Rogers, Craig	HC1240	2208



Faculty	Office Room	Office Tel.
Will, Peter	HC1202	4012
Shaw, Erin	HC1225	3416
Frank, Martin	HC1225	1677
Rogers, Craig	HC1240	2208



<b>Faculty</b> Will, Peter <b>Office Room</b> HC1202 <b>Office Tel.</b> 4012
<b>Faculty</b> Shaw, Erin <b>Office Room</b> HC1225 <b>Office Tel.</b> 3416
<b>Faculty</b> Frank, Martin <b>Office Room</b> HC1225 <b>Office Tel.</b> 1677
<b>Faculty</b> Rogers, Craig <b>Office Room</b> HC1240 <b>Office Tel.</b> 2208

# Semantic Annotation

- algorithm uses two approaches for semantic annotation
- *annotation by topic nodes*, uses mappings in T-DOM
- *annotation by regular expressions*: language based, simplest Natural Language Processing

# Implementation

- algorithm is implemented as *web application*:  
<http://tao.my-net-space.net/h2x>
- three sub-systems: *Document Retrieving Module, Data Extraction Module, XSL Templates Factory Module*

# Implementation (cont.)

Kundendienst

Mail an den ORF

Impressum

Publikumsrat

- Inland** ...> ÖVP für schwarz-orange Option auch nach 2006
- ...> Bartenstein verteidigt EU-Dienstleistungsrichtlinie
- ...> "Profil"-Umfrage: SPÖ profitiert von FPÖ-Spaltung
- ...> 60 Jahre ÖVP: Volkspartei zieht zufriedene Bilanz

- Ausland** ...> Prager Regierungskrise verschärft sich
- ...> Nach Koalitionsbruch Rücktritt Berlusconis erwartet
- ...> Wahl im Baskenland: Test für Autonomiepläne
- ...> Gedenken an Befreiung von KZ Bergen-Belsen

# Implementation (cont.)

```
<link href="http://orf.at/images/news_startseite.gif">ORF.at als Startseite im B
<weblinks>
  <suche_ORF_at href="http://suche.orf.at"/>
</weblinks>
<weblinks>
  <radio_orf_at href="http://radio.orf.at"/>
  <tv_orf_at href="http://tv.orf.at"/>
  <confetti_orf_at href="http://confetti.orf.at"/>
  <sport_orf_at href="http://sport.orf.at"/>
  <oesterreich_orf_at href="http://oesterreich.orf.at"/>
  <wetter_orf_at href="http://wetter.orf.at"/>
</weblinks>
</webpage>
<webpage url="http://orf.at/ticker/editorial.html">
  <page_title>ORF ON News</page_title>
  <Kundendienst href="http://kundendienst.orf.at"/>
  <ORF href="http://kundendienst.orf.at/kontakte/www.html"/>
  <Impressum href="http://orf.at/impressum"/>
  <Publikumsrat href="http://publikumsrat.orf.at"/>
  <Kundendienst href="http://kundendienst.orf.at"/>
  <ORF href="http://kundendienst.orf.at/kontakte/www.html"/>
  <Impressum href="http://orf.at/impressum"/>
  <Publikumsrat href="http://publikumsrat.orf.at"/>
  <data>Ausland</data>
  <link href="http://orf.at/ticker/179153.html?tmp=17011">Regierungsbündnis verlie
  <link href="http://orf.at/ticker/179152.html?tmp=17916">Rätselraten um angeblich
  <link href="http://orf.at/ticker/179149.html?tmp=9122">Pro-Europäer Talat gewinn
```

# A Gateway From HTML to XML

Thank you for your patience!

Friedrich Dimmel, 0302230