Using the Structure of Web Sites for Automatic Segmentation of Tables
by Kiristina Lerman, Lise Getoor, Steven Minton and Craig Knoblock, 2004;
Paper #6 for PS Web-Extraction, Presentation: Christoph Veigl

Problems for layout-based segmentation techniques:

- variability in use of layout-tags
  `<td>, <tr>, .. multi-column-text, image-layout ..`
  `<br>, “~”, .. separate fields as well as items`
  -> *Pat-Trees* suffer from misinterpretations

- domain-dependence of *Web-wrappers* and many heuristics
- training examples have to be updated when site changes

GOALS:

- unsupervised extraction
- resistant to layout-changes
- domain independent and fully automatic processing

-> little need of human resources, adaption to site-changes
**IDEA:** Using structure in layout and content that is common to many “hidden-web”-sites generated by web-queries.

Web-Queries are generated by a de-facto convention:

- HTML-Form: user inputs the query

- overview of results is presented on a list-page that contains a description and a link to more specific information, the detail-page

- the detail page refers to exactly one item from the list page

This redundant information in the content of detail-pages could point to a possible record segmentation of items on the list page.

**Drawback:** will work only with web sites having list-/detail structure
An example for list-/detail structure:

Querying white-pages (superpages.com)

notice:

- List page
- Detail-pages
- Table
- Templates for the pages
Preparation and Segmentation-Implementations:

Logical Constrains:
\[ x_1 + x_2 = 1 \]

Probabilistic Dependencies:
\[ P(A|B) = 0.75 \]

results: TOKEN-TYPES
HTML, Punctuation, Alphabetic Numeric, Capitalized, Lowercased Bold, Italic

results:
Extracts \( E = \{E_1, E_2, \ldots, E_n\} \) from list page
Detail pages \( \{r_1, r_2, \ldots, r_k\} \)
Detail pages \( D_i \) on which \( E_i \) appears

<table>
<thead>
<tr>
<th>( D_1 )</th>
<th>( E_1 )</th>
<th>( E_2 )</th>
<th>( E_3 )</th>
<th>( E_4 )</th>
<th>( E_5 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>221 Wa</td>
<td>New</td>
<td>335-5555</td>
<td>John</td>
<td></td>
</tr>
<tr>
<td>Smith</td>
<td>Smith</td>
<td>John</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Approach I: CSP
Constraint Satisfaction

Approach II: HMM
Probabilistic Model

Logical Constrains:
\( x_1 + x_2 = 1 \)

Probabilistic Dependencies:
\( P(A|B) = 0.75 \)
To formulate a CSP for the record segmentation task, we need an “assignment-variable” $x_{ij}$: $x_{ij}$ is 1 $\iff$ $E_i$ is assigned to record $r_j$.

1. Uniqueness constraint:
Every extract belongs to exactly one record $r_j$ $\sum_j x_{ij} = 1$

<table>
<thead>
<tr>
<th>$E_1$</th>
<th>$E_2$</th>
<th>$E_3$</th>
<th>$E_4$</th>
<th>$E_5$</th>
<th>$E_6$</th>
<th>$E_7$</th>
<th>$E_8$</th>
<th>$E_9$</th>
<th>$E_{10}$</th>
<th>$E_{11}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>New</td>
<td>Smith</td>
<td>(740)</td>
<td>Smith</td>
<td>Wash</td>
<td>(740)</td>
<td>Smith</td>
<td>Findlay</td>
<td>George W.</td>
<td>(419)</td>
</tr>
<tr>
<td>Smith</td>
<td>221 Wa</td>
<td>Holland</td>
<td>335-5555</td>
<td>221R Wa</td>
<td>shington</td>
<td>335-5555</td>
<td>Smith</td>
<td>OH...</td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>ington</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>$D_i$</td>
<td>$r_{1,2}$</td>
<td>$r_1$</td>
<td>$r_1$</td>
<td>$r_{1,2}$</td>
<td>$r_2$</td>
<td>$r_2$</td>
<td>$r_{1,2}$</td>
<td>$r_3$</td>
<td>$r_3$</td>
<td>$r_3$</td>
</tr>
</tbody>
</table>

application for multiple extracts and for unique ones

$x_{11} + x_{12} = 1$  \hspace{1cm} $x_{21} = 1$
$x_{41} + x_{42} = 1$  \hspace{1cm} $x_{31} = 1$
$x_{51} + x_{52} = 1$  \hspace{1cm} $x_{62} = 1$
$x_{81} + x_{82} = 1$  \hspace{1cm} $x_{72} = 1$  ...
2. Consecutive constraint: 
only contiguous blocks of extracts can be assigned 
to the same record: \( x_{nj} = 0, \ k<n<i \Rightarrow x_{kj} + x_{ij} \leq 1 \)

3. Constraints from structural assumptions: 
extract \( E_i \) was not observed on detail page \( r_j \) \( \Rightarrow \) \( x_{ij} = 0 \)

4. Position constraints: 
\( E_i, E_k \) same position on the detail pages \( \Rightarrow \) assign to different records

WSAT(OIP) yields the segmentation:

<table>
<thead>
<tr>
<th>( E_1 )</th>
<th>( E_2 )</th>
<th>( E_3 )</th>
<th>( E_4 )</th>
<th>( E_5 )</th>
<th>( E_6 )</th>
<th>( E_7 )</th>
<th>( E_8 )</th>
<th>( E_9 )</th>
<th>( E_{10} )</th>
<th>( E_{11} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>John</td>
<td>221 Wa</td>
<td>New</td>
<td>(740)</td>
<td>John</td>
<td>221R Wa</td>
<td>Wash</td>
<td>(740)</td>
<td>George W.</td>
<td>Findlay,</td>
<td>(419)</td>
</tr>
<tr>
<td>Smith</td>
<td>ington...</td>
<td>Holland...</td>
<td>335-5555</td>
<td>Smith</td>
<td>ington...</td>
<td>ington...</td>
<td>335-5555</td>
<td>Smith</td>
<td>OH...</td>
<td>423-1212</td>
</tr>
<tr>
<td>( D_i )</td>
<td>r1,r2</td>
<td>r1</td>
<td>r1</td>
<td>r1,r2</td>
<td>r2</td>
<td>r2</td>
<td>r1,r2</td>
<td>r3</td>
<td>r3</td>
<td>r3</td>
</tr>
</tbody>
</table>

for first record (j=1):
\( x_{11} + x_{81} \leq 1 \)
\( x_{21} + x_{81} \leq 1 \)
\( x_{31} + x_{81} \leq 1 \)
\( x_{41} + x_{81} \leq 1 \) ...
The Probabilistic approach: Hidden Markov Models

- The Hidden Markov Model is a finite set of *states*
- Transitions among the states are governed by (time-invariant) *transition probabilities*
- The states generate *visible observations*, the states themselves are *hidden*
- Markov assumption: the state of the model depends only upon the previous n states

\[ \lambda=(S,M,A,B, \pi) \]

- \(S\) .. *states*
- \(M\) .. observation symbols
- \(A=aij\) .. state transition probabilities
- \(B=bj(k)\) .. observation probabilities
- \(\pi=\pi_i\) .. initial state distribution

The Learning Problem

Given a model and a sequence of observations, how should we adjust the model parameters in order to maximize the expectation of the observations?

-> *Baum-Welch, EM*
A Model for our segmentation task:

Observed Variables:
- \( T \)  token-types of extract \( E_i \)
- \( D \)  detail pages where \( E_i \) occurred

Unobserved Variables:
- \( R \)  record number of the extract
- \( C \)  column label of the extract
- \( S \)  true if \( E_i \) is the start of a new record, false otherwise
- \( \pi \)  table period (number of columns)

dependencies:
- \( P(T_i|C_i) \): token type depends on column label
- \( P(C_i|C_{i-1}) \): column label depends on previous column
- \( P(S_i|C_i) \): start of a new record depends on column label
- \( P(R_i|R_{i-1}, D_i, S_i) \): record number depends on previous record number, record-start and detail-page

bootstrapping:
- \( P(T_{ij}=\text{true}|C_i) = 1/|T| \)  (initial token type)
- \( P(R_{i}={}\text{ri}) = 1/|D_i| \)  (0 if there is no \( D_i \))
- \( P(S_i=\text{true}) \)  if \( D_{i-1} \cap D_i = 0 \)

compute “maximum a-posteriori probability” (MAP): \( \arg \max P(R,C|T,D) \)
Results:

<table>
<thead>
<tr>
<th>Wrapper</th>
<th>Probabilistic</th>
<th>CSP</th>
<th>notes</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Cor</td>
<td>InC</td>
<td>FN</td>
</tr>
<tr>
<td>Amazon</td>
<td>4</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>Books</td>
<td>2</td>
<td>5</td>
<td>3</td>
</tr>
<tr>
<td>BN</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Books</td>
<td>5</td>
<td>5</td>
<td>0</td>
</tr>
<tr>
<td>Allegheny</td>
<td>20</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>County</td>
<td>16</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Butler</td>
<td>15</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>County</td>
<td>12</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Lee</td>
<td>16</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>County</td>
<td>5</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Michigan</td>
<td>7</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Corrections</td>
<td>12</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Minnesota</td>
<td>11</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Corrections</td>
<td>17</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Ohio</td>
<td>8</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Corrections</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Canada</td>
<td>18</td>
<td>7</td>
<td>0</td>
</tr>
<tr>
<td>411</td>
<td>1</td>
<td>4</td>
<td>0</td>
</tr>
<tr>
<td>Sprint</td>
<td>17</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Canada</td>
<td>8</td>
<td>12</td>
<td>0</td>
</tr>
<tr>
<td>Yahoo</td>
<td>0</td>
<td>10</td>
<td>0</td>
</tr>
<tr>
<td>People</td>
<td>10</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Super</td>
<td>3</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>Pages</td>
<td>9</td>
<td>6</td>
<td>0</td>
</tr>
</tbody>
</table>

**Precision (P):**

\[ P = \frac{Cor}{Cor + Inc + FP} \]

**Recall (R):**

\[ R = \frac{Cor}{Cor + FN} \]

Notes:
- a. Page template problem
- b. Entire page used
- c. No solution found
- d. Relax constraints
Tests with local sites
Expected result: name, address, description. phone-number will not be segmented