

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 ..

, “~”, .. 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:

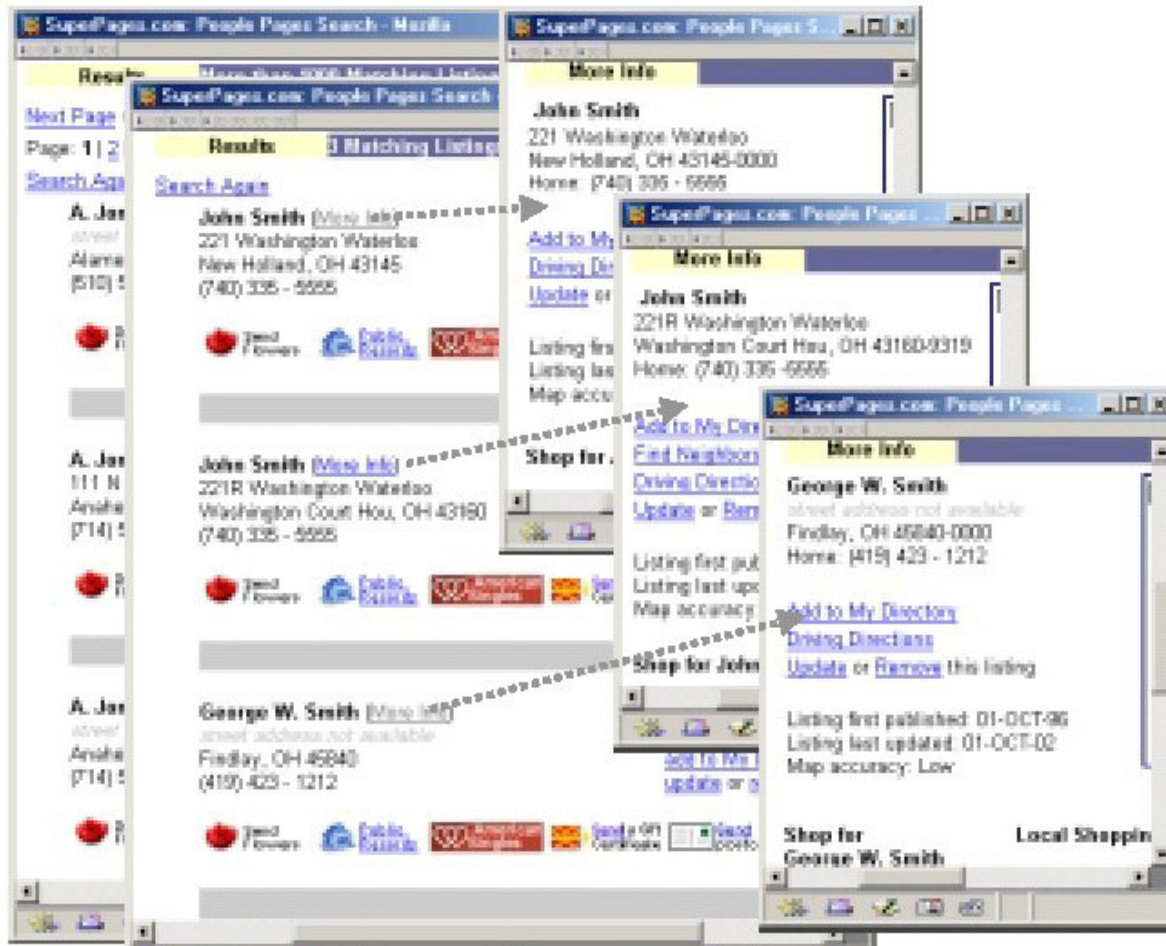
- HTML-Form: user inputs the query
- overview of results is presented on a **list-page** that contains a description and a link to more specific information, the **detail-page**
- the detail page refers to exactly one item from the list page

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)

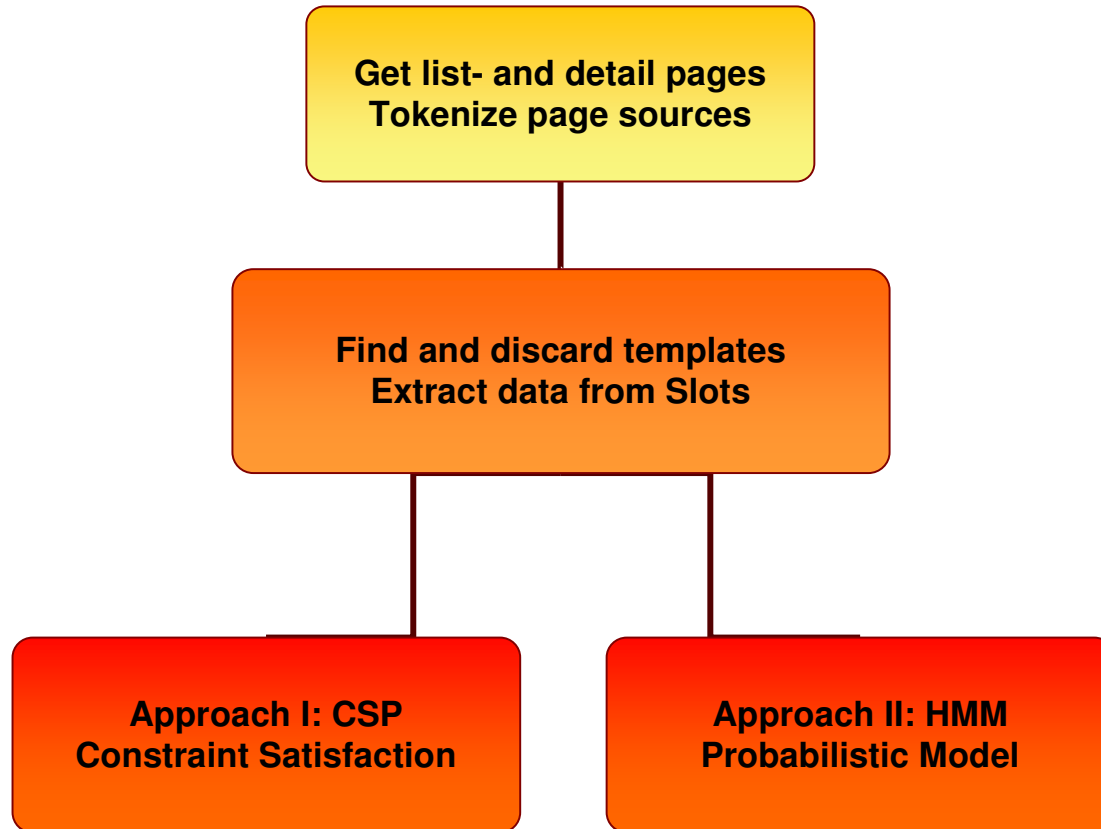


notice:

- List page
- Detail-pages
- Table

- Templates for the pages

Preparation and Segmentation-Implementations:



results: TOKEN-TYPES

HTML, Punctuation, Alphabetic
 Numeric, Capitalized, Lowercased
 Bold, Italic

results:

Extracts $E = \{E_1, E_2, \dots, E_n\}$ from list page
 Detail pages $\{r_1, r_2, \dots, r_k\}$
 Detail pages D_i on which E_i appears

	E_1	E_2	E_3	E_4	E_5
	John Smith	221 Wa ington...	New Holland...	(740) 335-5555	John Smith
D_i	r_1, r_2	r_1	r_1	r_1, r_2	r_1, r_2

Logical Constrains:
 $x_1 + x_2 = 1$

Probabilistic Dependencies:
 $P(A|B) = 0,75$

The CSP Approach: Constraint Satisfaction Problems

- logical expressions over a set of variables
- values of variables can be 0 or 1
- solved CSP: all constraints are satisfied at the same time

To formulate a CSP for the record segmentation task, we need an “assignment-variable” x_{ij} : x_{ij} is 1 \Leftrightarrow E_i is assigned to record r_j

1. Uniqueness constraint:

Every extract belongs to exactly one record r_j

$$\sum_j x_{ij} = 1$$

	E_1	E_2	E_3	E_4	E_5	E_6	E_7	E_8	E_9	E_{10}	E_{11}
	John Smith	221 Wa ington...	New Holland...	(740) 335-5555	John Smith	221R Wa shington...	Wash ington...	(740) 335-5555	George W. Smith	Findlay, OH...	(419) 423-1212
D_i	$r1, r2$	$r1$	$r1$	$r1, r2$	$r1, r2$	$r2$	$r2$	$r1, r2$	$r3$	$r3$	$r3$

application for multiple extracts

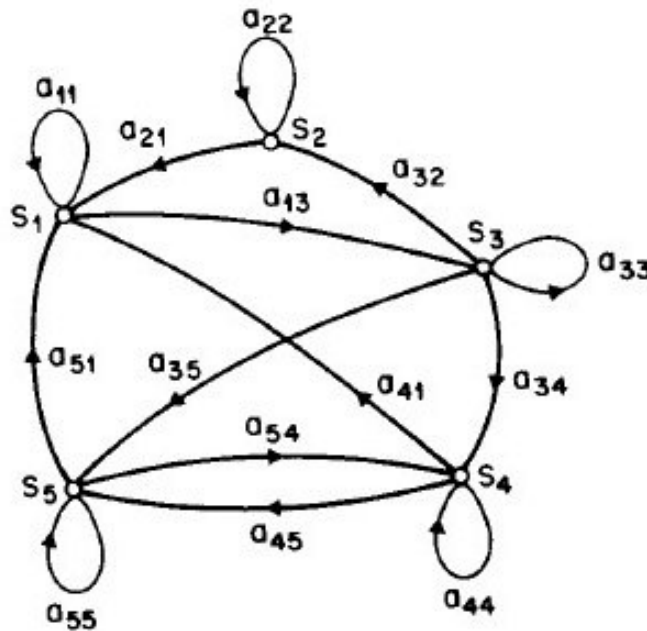
$$\begin{aligned} x_{11} + x_{12} &= 1 \\ x_{41} + x_{42} &= 1 \\ x_{51} + x_{52} &= 1 \\ x_{81} + x_{82} &= 1 \end{aligned}$$

and for unique ones

$$\begin{aligned} x_{21} &= 1 \\ x_{31} &= 1 \\ x_{62} &= 1 \\ x_{72} &= 1 \quad \dots \end{aligned}$$

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

$A = a_{ij}$.. state transition probabilities

$B = b_j(k)$.. observation probabilities

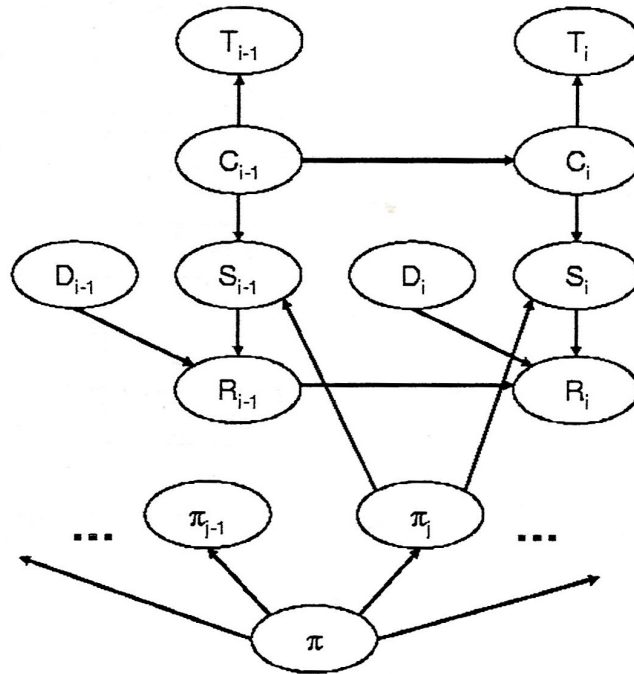
$\pi = \pi_i$.. initial state distribution

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
- π table period (number of columns)

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

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)
- $P(R_i = r_i) = 1 / |D_i|$ (0 if there is no D_i)
- $P(S_i = \text{true})$ if $D_{i-1} \cap D_i = 0$

Results :

Wrapper	Probabilistic				CSP				notes
	Cor	InC	FN	FP	Cor	InC	FN	FP	
Amazon Books	4 2	6 5	0 3	1 4	0 0	0 0	10 10	0 0	a, b
BN Books	5 5	5 5	0 0	0 0	2 0	0 0	8 10	0 0	a, b, c, d
Allegheny County	20 16	0 4	0 0	0 0	20 20	0 0	0 0	0 0	
Butler County	15 12	0 0	0 0	0 0	15 12	0 0	0 0	0 0	
Lee County	16 5	0 0	0 0	0 0	16 5	0 0	0 0	0 0	
Michigan Corrections	7 12	0 4	0 0	0 0	4 2	3 8	0 6	0 0	c, d
Minnesota Corrections	11 17	0 2	0 0	0 0	4 8	7 9	0 0	0 2	a, b, c, d
Ohio Corrections	8 10	2 0	0 0	0 0	10 10	0 0	0 0	0 0	
Canada 411	18 1	7 4	0 0	0 0	25 1	0 4	0 0	0 0	c, d
Sprint Canada	17 8	3 12	0 0	0 0	20 20	0 0	0 0	0 0	
Yahoo People	0 10	10 0	0 0	0 0	5 10	5 0	0 0	0 0	a, b, c, d b
Super Pages	3 9	0 6	0 0	0 0	3 15	0 0	0 0	0 0	a, b
<i>Precision</i>	0.74				0.85				
<i>Recall</i>	0.99				0.84				
<i>F</i>	0.85				0.84				

Notes

- a. Page template problem; b. Entire page used; c. No solution found;
- d. Relax constraints

Cor: correctly segmented
InC: incorrectly segmented
FN: unsegmented
FP: non records

Precision (P):
 $P = \text{Cor} / (\text{Cor} + \text{Incor} + \text{FP})$

Recall (R):
 $R = \text{Cor} / (\text{Cor} + \text{FN})$

Tests with local sites

The screenshot shows the HEROLD.at website interface. At the top left is the logo "HEROLD.at". Below it is a navigation bar with links: "Bookmark anlegen", "English", "Download Toolbar", "HEROLD.at", and "Hilfe". A secondary navigation bar includes "Home Page", "Gelbe Seiten", "TelefonBuch", "Ihr Bezirk", "B2B Info", "KundenZone", "ShoppingZone", "Über HEROLD", and "Services".

The main content area is titled "Suchmodus: Teilnehmer". It features a search form with the following fields:

- Name**: Input field containing "Veigl".
- Vorname**: Empty input field.
- Zusatz**: Input field with placeholder text "z.B. Ärzte, Restaurant,...".
- Region**: Dropdown menu set to "Wien".
- Ort, PLZ**: Input field.
- Straße**: Input field.
- Nr.**: Input field.

A "Suchen" button is located below the form, along with icons for a trash can and a question mark.

To the right of the search form, there is a "sponsored by" section with logos for "MANPOWER AUSTRIA" and "tele".

Below the search form, a section titled "Ihre letzten Suchen" lists four previous searches:

- 1 [Veigl/ohne Einschränkung/](#)
- 2 [Veigl/Vorarlberg/](#)
- 3 [Sucher/Vorarlberg/](#)
- 4 [veigl/ohne Einschränkung/](#)

The "Datenstand" is noted as "KW 14 / 2005".

At the bottom of the search results, there are three informational links:

- [TelefonBuch Eintrag](#) - Für einen Neueintrag oder eine Änderung informieren Sie sich hier.
- [Werben in HEROLD.at](#) - Informieren Sie sich über Ihre Werbemöglichkeiten.
- [Datenquellen](#) - Wer stellt die Daten des HEROLD TelefonBuch zur Verfügung.

The footer of the page includes the text: "Database Engine: ©1995-2002 [Altova GmbH](#)".

The bottom of the browser window shows a status bar with "Internet" and a small globe icon.

HEROLD.at

Bookmark anlegen Download Toolbar

Home Page Gelbe Seiten TelefonBuch Ihr

Ergebnisse: nach Alphabet Neue Suche

Suche nach: Veigl
Region: Wien

1-10 von 66

Redaktionelle Einträge

[Veigl Adolf /](#)
Inge, Elektrotechn,
1190 Wien (W), Nußbergg 23

[Veigl Alexander,](#)
Büroorganisation,
1130 Wien (W), Auhofstr 58

[Veigl Alfred /](#)
Anna, BBea,
1210 Wien (W), Bessemerstr 10-16, Stg 1

[Veigl Amalia,](#)
1160 Wien (W), Paletzg 17, Stg 3

[Veigl Andreas,](#)
1030 Wien (W), Kleing 20

[Veigl Annemarie,](#)
1210 Wien (W), Herzmanovsky-Orlando-G 13, Stg 37

[Veigl Charlotte,](#)
1030 Wien (W), Droryg 8, Stg 1

Fertig

1-10 von 66

Wien (W)

Veigl Alexander,
Büroorganisation,
13 Auhofstr 58, 1130 Wien (W)
01 / 797 79...-0
01 / 797 79 02
0676 / 430 27 57
0676 / 430 27 60

1-10 von 66

Redaktionelle Einträge

[Veigl Adolf /](#)
Inge, Elektrotechn,
1190 Wien (W), Nußbergg 23

[Veigl Alexander,](#)
Büroorganisation,
1130 Wien (W), Auhofstr 58

[Veigl Alfred /](#)
Anna, BBea,
1210 Wien (W), Bessemerstr 10-16, Stg 1

Wien (W)

Veigl Alfred /
Anna, BBea,
21 Bessemerstr 10-16, Stg 1, 1210 Wien (W)
01 / 256 16 01

1-10 von 66

Redaktionelle Einträge

[Veigl Adolf /](#)
Inge, Elektrotechn,
1190 Wien (W), Nußbergg 23

[Veigl Alexander,](#)
Büroorganisation,
1130 Wien (W), Auhofstr 58

[Veigl Alfred /](#)
Anna, BBea,
1210 Wien (W), Bessemerstr 10-16, Stg 1

Wien (W)

Veigl Christoph,
12 Vivenotg 51, 1120 Wien (W)
01 / 817 64 23
0676 / 780 76 80

Expected result: name, address, description. phone-number will not be segmented

The screenshot shows the Conrad Austria website interface. At the top left is the logo for CONRAD ÖSTERREICH. The search bar contains 'avr'. Below the search bar is a navigation menu with categories like 'Akkus & Batterien', 'Elektronik', and 'Computer / Büro'. The search results section shows 'Ihr Suchergebnis (Übersicht)' with 37 products in 5 categories. A list of products is displayed under 'Elektronik (35 Treffer)', including ATMEGA8535-16PC, MIKR.CONTR. ATMEGA161-8PC=162-16PI DIL40, and ATMEGA161-TQFP44. A detailed view of the AT90S1200-12SC processor is shown on the right, with a table of specifications.

AT 90 S 1200-12 SC=I PROZESSOR
 Artikel-Nr.: 152951 - HK

Marke	Zusatzinfo
ATMEL	Mengenrabatt
Typ	AT90S1200-12SC
Gehäuse	SOIC20
FLASH (KB)	1
EEPROM (Bytes)	64
RAM (Bytes)	-
I/O Pins	15
SPI	-
UART	-
TWI	-
Hardware Multiplier	-
8-bit Timer	1
16-bit Timer	-
10-bit A/D Channels	-

Expected result: art.-num, art.-name, price will be segmented. little item-infos