

PS Web IE

<http://dbai.tuwien.ac.at/education/wie>

Intermediate hand-ins
Tuesday, April 12, 2005

0

„THE BIG PICTURE“

PROBLEM description

- What is the problem the paper wants to address?

Examples

- ...

SOLUTION approach

- What is the basic idea behind the solution of the paper?

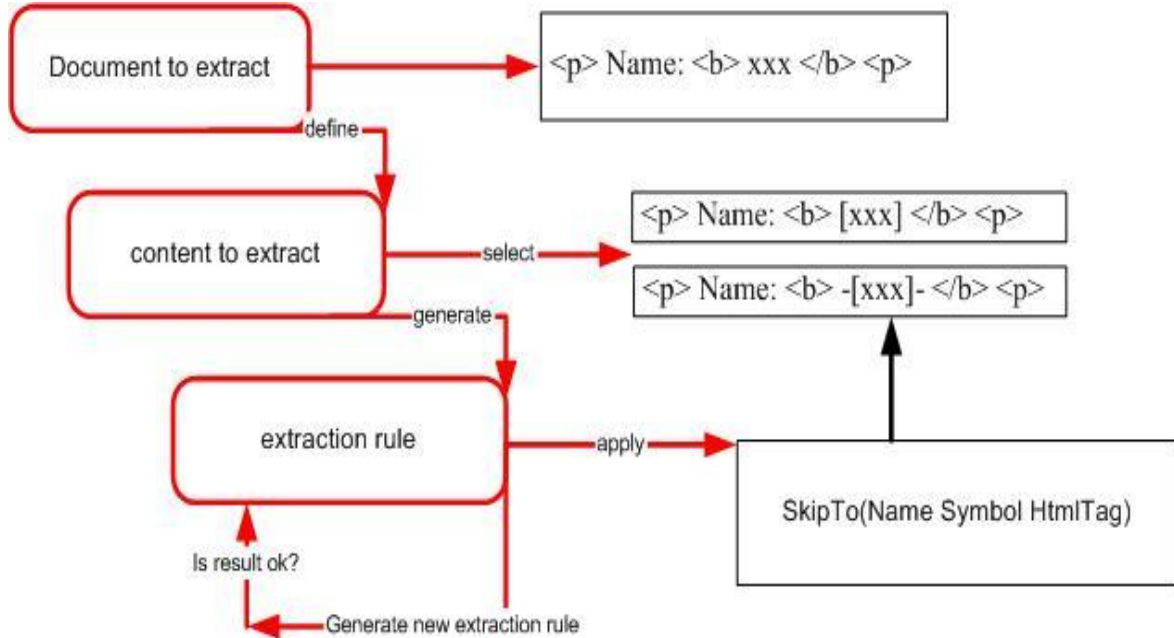
Examples

- ...

Stalker – A powerful algorithm

Stefan Schönig

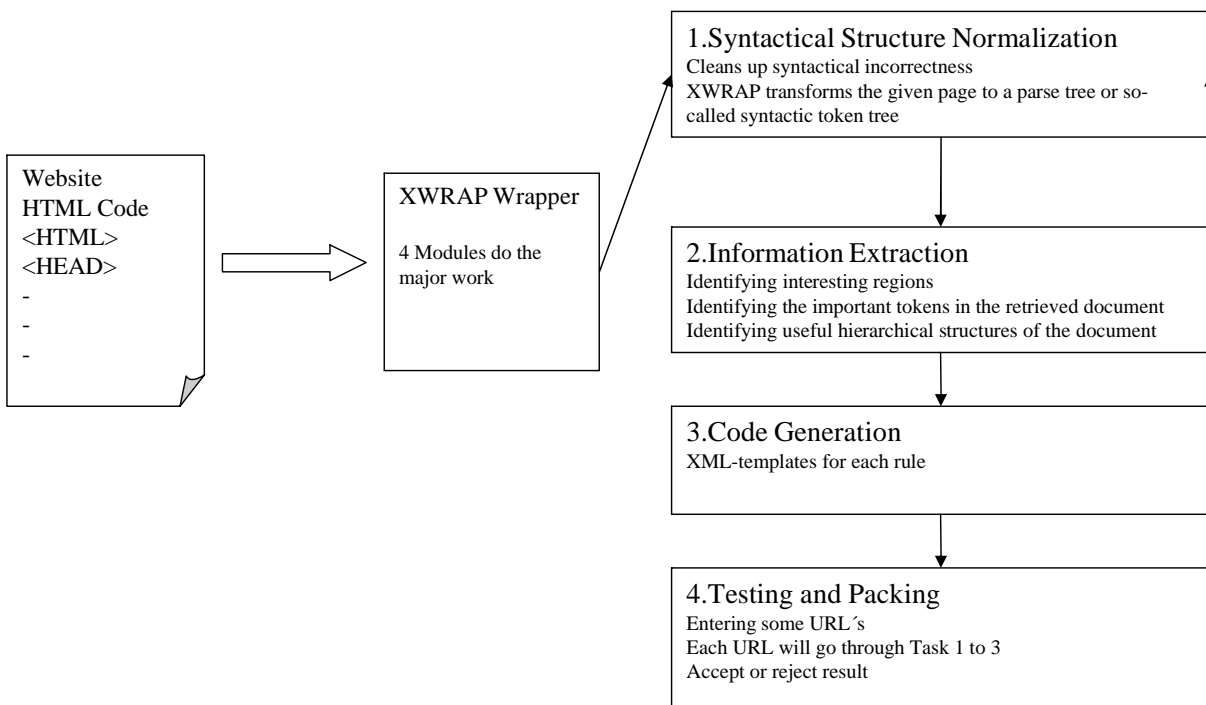
A Hierarchical Approach to Wrapper Induction



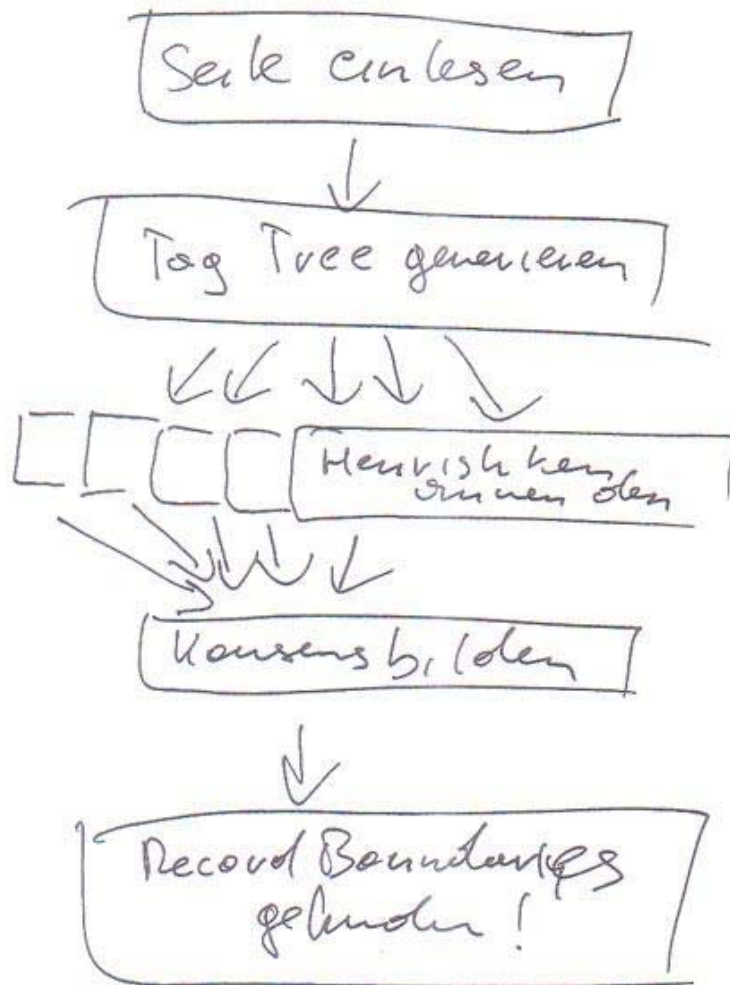
Xwrap Schema

Marco Schönig

XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources



#5



René Kiesler

Record-Boundary
Discovery in
Web Documents

4

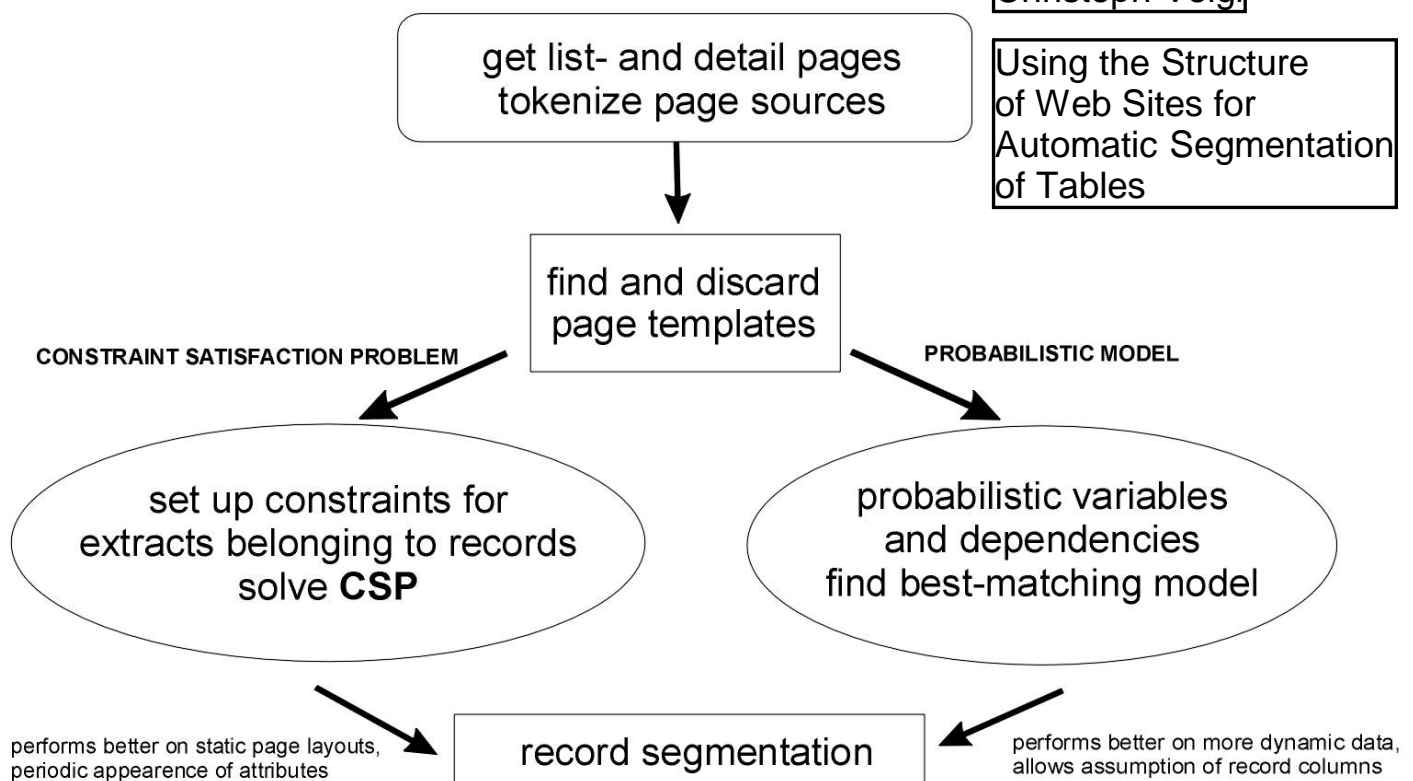
#6

Content-based table-extraction from dynamic web pages

CSPs and Hidden Markov Models are two possible approaches for automatic, domain-independent record extraction from the "hidden-web". The described techniques use redundancies and structure in list- and detail pages that are generated as results for standard web queries. The paper points out common preparation of input data, describes the application of the models and evaluates their performance.

Christoph Veigl

Using the Structure
of Web Sites for
Automatic Segmentation
of Tables



5

#7

Automatic Data Extraction from Lists and Tables in Web Sources (Lerman/Knoblock/Hinton)

Marian Schedenig
9725416/881

Example pages
(html)

unsupervised
Learning algorithm

template

Data page
(html)

extracted
data

html
Example pages

Tokenisation*

Page template

Tokens extracted from page code

1) clustering (using a set of delimiter tokens)
2) Autoclass

"Columns"

Dataset cells

ALERGIA-derived grammar induction algorithm

"Rows"

Dataset records / tuples of cells

still in token form → used as a "template"/wrapper for subsequent data extraction

*based on strings, which typically include html table and list tags, but could also be used for other file formats

6

#8

mdr - mining data records

Bohunsky Paul 0025058

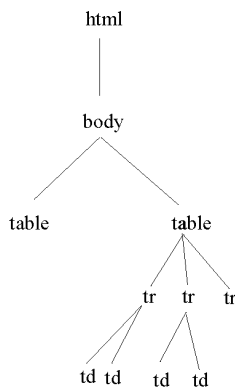
algorithm to find and identify data regions and data records*
only works in structured html documents which must be designed using <tags>

Mining Web Pages for Data Records

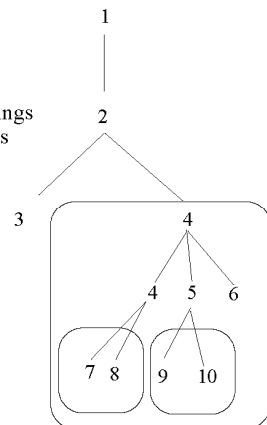
document.html

```
<html>
<body>
<table>
<tr><td></td></tr>
<tr><td></td></tr>
<tr><td></td></tr>
.
.
</table>
</body>
</html>
```

structured tag tree



string comparison of tag strings
to identify generalized nodes

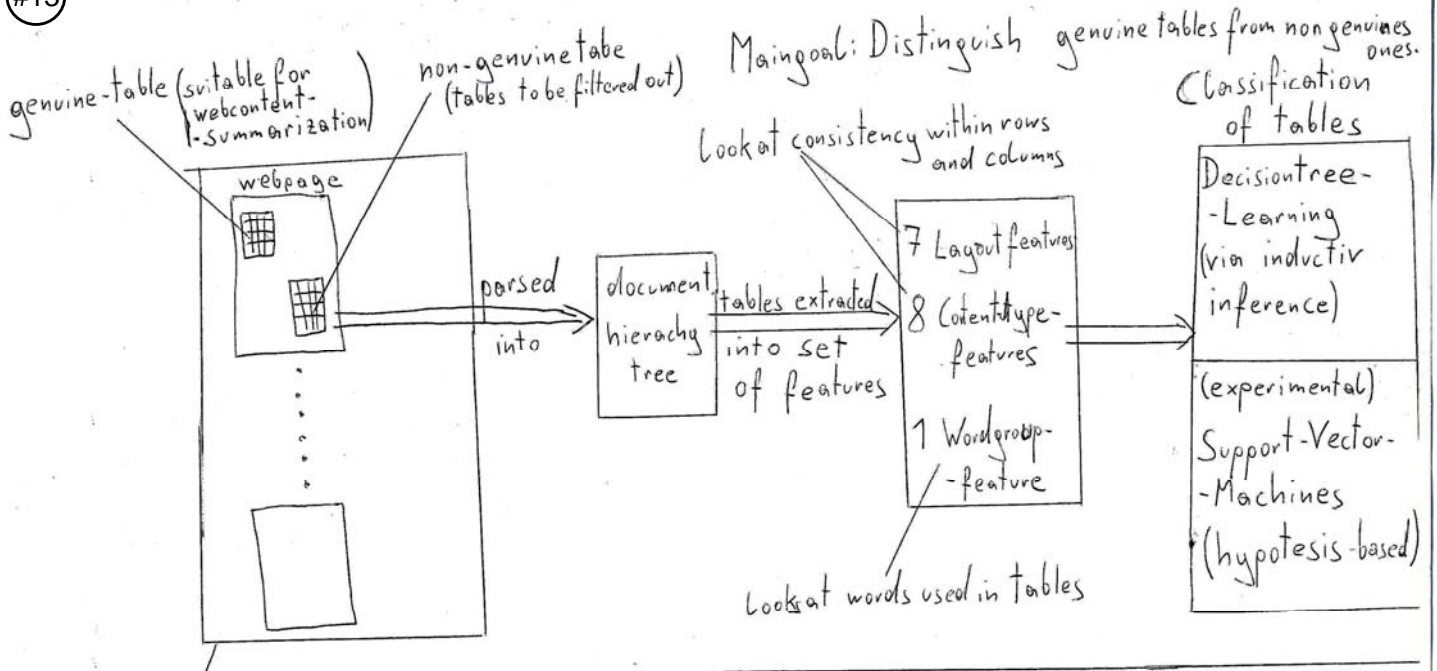


after generalized nodes are identified, data regions and their data records are determined by visiting each node recursively
this algorithm shows very good performance (almost perfect) on "faultless" webpages.

* data region: a group of data records in a contiguous region of a page
data record: group containing similar objects

7

#13



Database of webpages (training data)
 (tables get extra-atributes in the database)
 1) unique id
 2) Is it genuine? (yes/no)
 3) tabletitle

A Machine Learning Based Approach for Table Detection on the Web

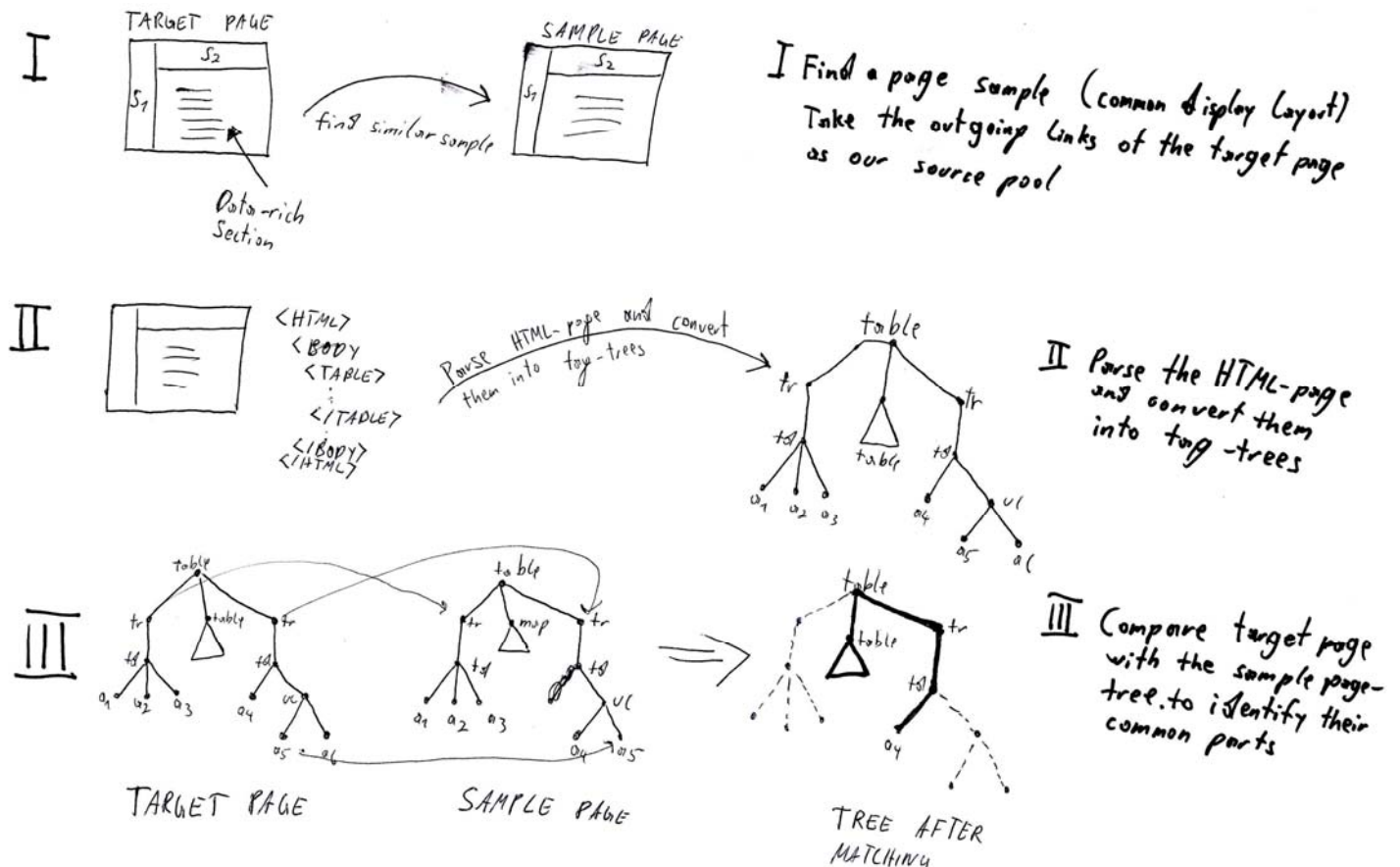
Gregor Pridun 9725153
 Max Arenas 9835111

10

#14

Data-rich Section Extraction from HTML pages

MAX ARENAS
 9835111

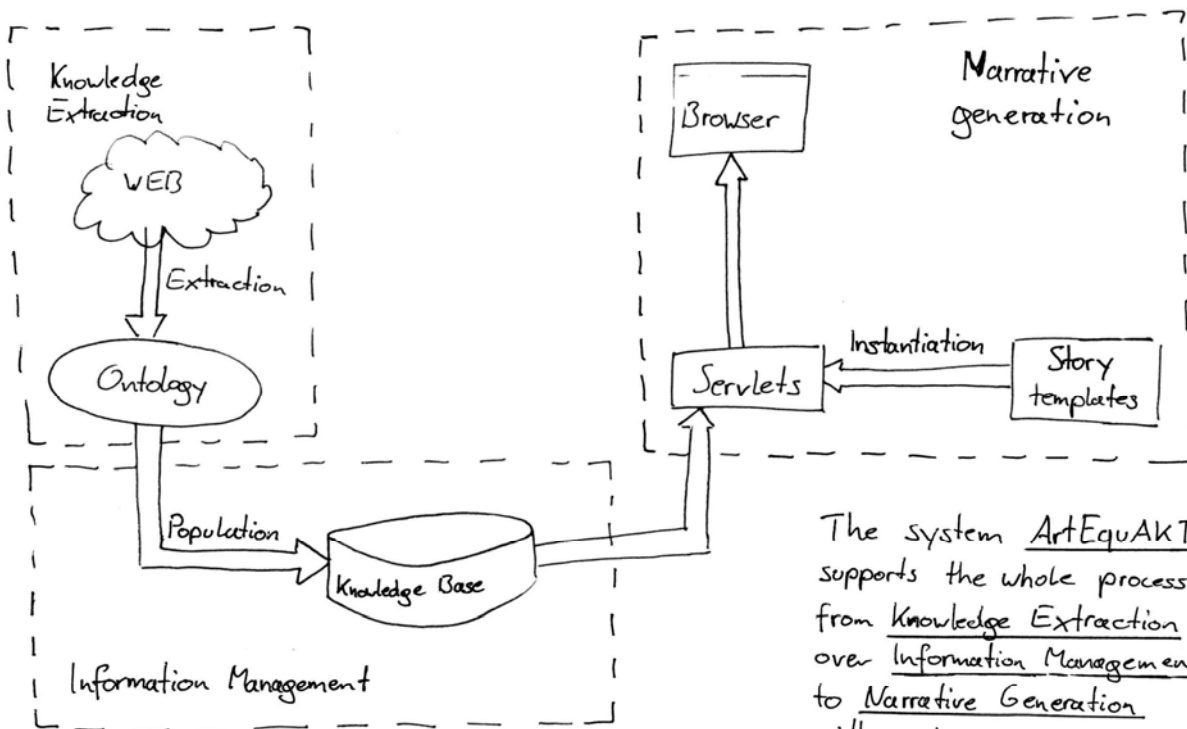


11

#15

Automatic Ontology based Knowledge Extraction from Web Documents

Stefan Bischof



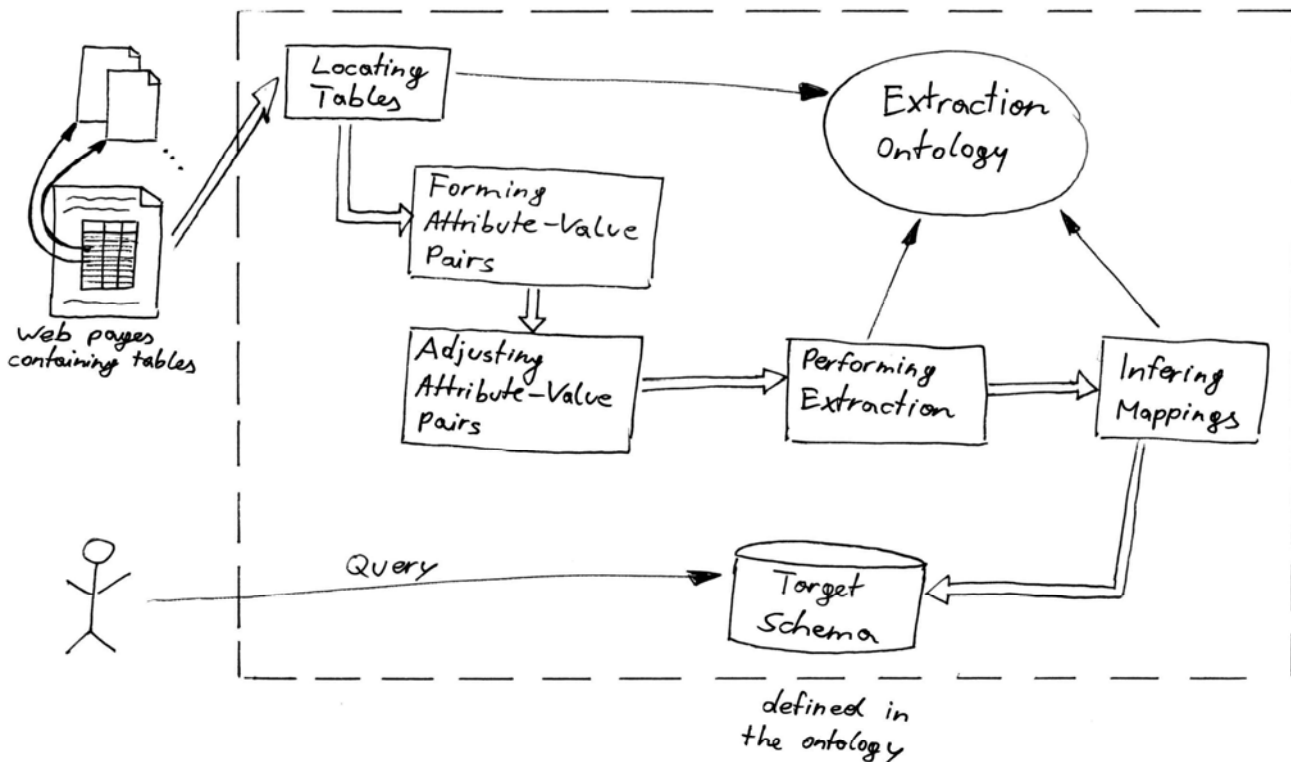
The system ArtEquAKT supports the whole process from Knowledge Extraction over Information Management to Narrative Generation with various programs.

12

#16

Automating the Extraction of Data from HTML Tables with Unknown Structure

Stefan Rümmele
0325665



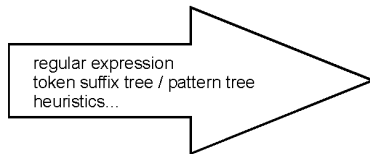
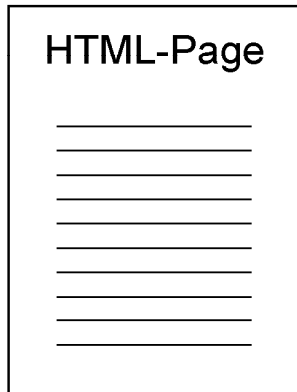
13

Data Extraction and Label Assignment for Web Databases

(Jiying Wang, Frederick H. Lochovsky)

to build a System that

extracts (automatically) text from a web-page into a table
assignes labels in a table



Title 1	Title 2	Title 3	...
Text	Text
...

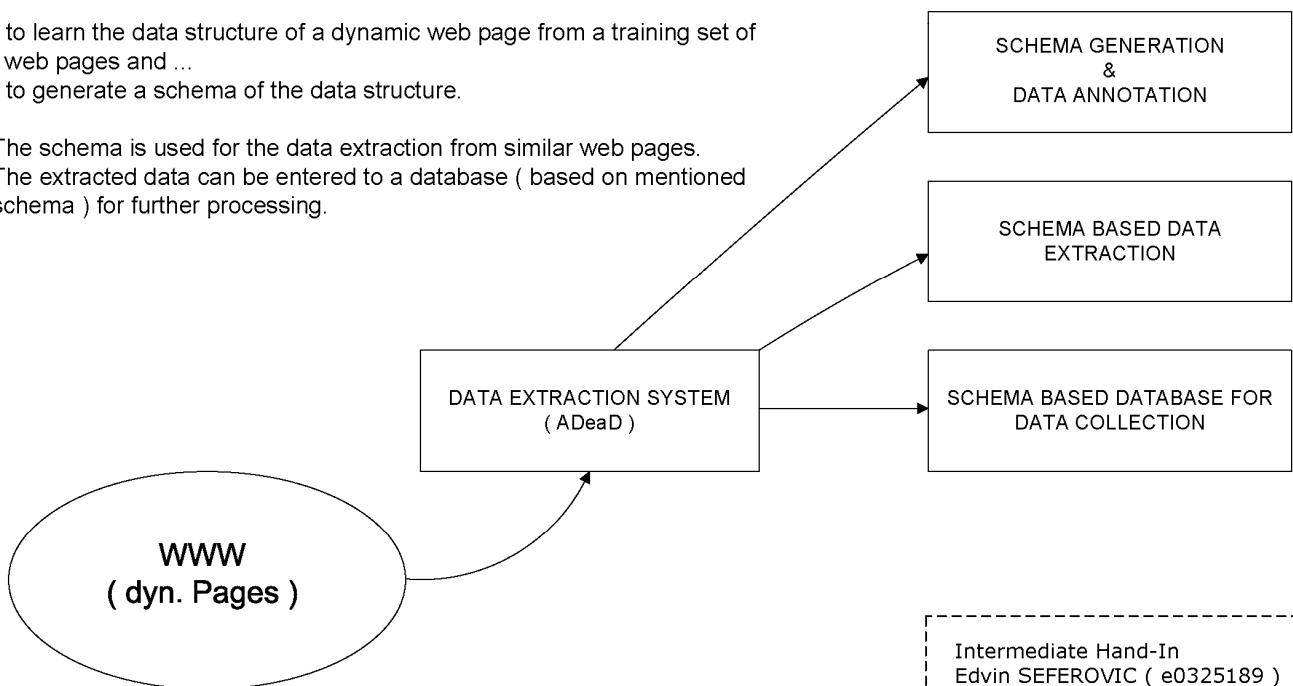
Leopold Redlingshofer, 0325929

Data Extraction and Annotation for Dynamic Web Pages

The aims of the ADeaD system are :

- to learn the data structure of a dynamic web page from a training set of web pages and ...
- to generate a schema of the data structure.

The schema is used for the data extraction from similar web pages.
The extracted data can be entered to a database (based on mentioned schema) for further processing.



Intermediate Hand-In
 Edvin SEFEROVIC (e0325189)
 PS - WIE SS2005
 Paper #20

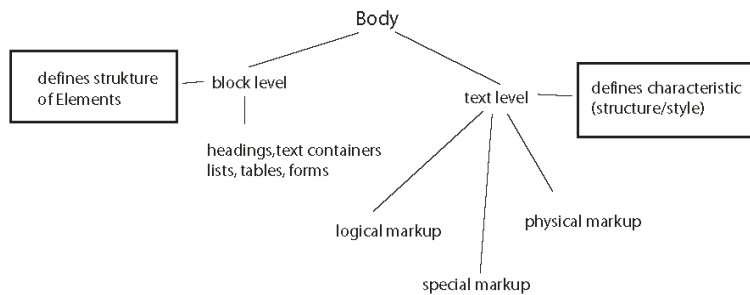
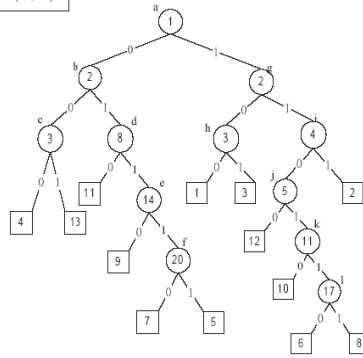
Applying Pattern Mining to Web Information Extraction

Hml(<H1>)Text(_Hml(<H1>)Hml()Hml()Text(_Hml()Text(_Hml()Text(_Hml()Text(_Hml())

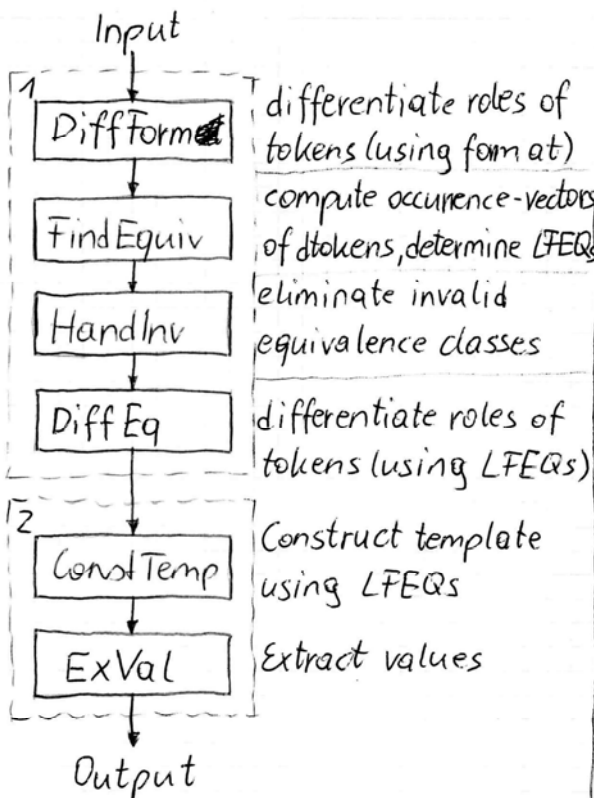
Hml()	000	Hml()	001
Hml()	010	Hml()	011
Hml(<H1>)	100	Hml(<H1>)	101
Text(_)	110		

0001101010000101100101100101100101100015

Indexing position:
 suffix 1 1001101010000101100101100101100101100015
 suffix 2 1101010000101100101100101100101100015
 suffix 3 1010000101100101100101100101100015
 suffix 4 0000101100101100101100101100101100015
 suffix 5 0101100101100101100101100101100015
 suffix 6 1100101100101100101100101100015
 suffix 7 0101100101100101100101100015
 suffix 8 1100101100101100101100015
 suffix 9 0101100101100101100015
 suffix 10 1100101100015
 suffix 11 0101100015
 suffix 12 1100015
 suffix 13 0015



Extracting Structured Data from Web Pages Tobias Dönz (0226-173)



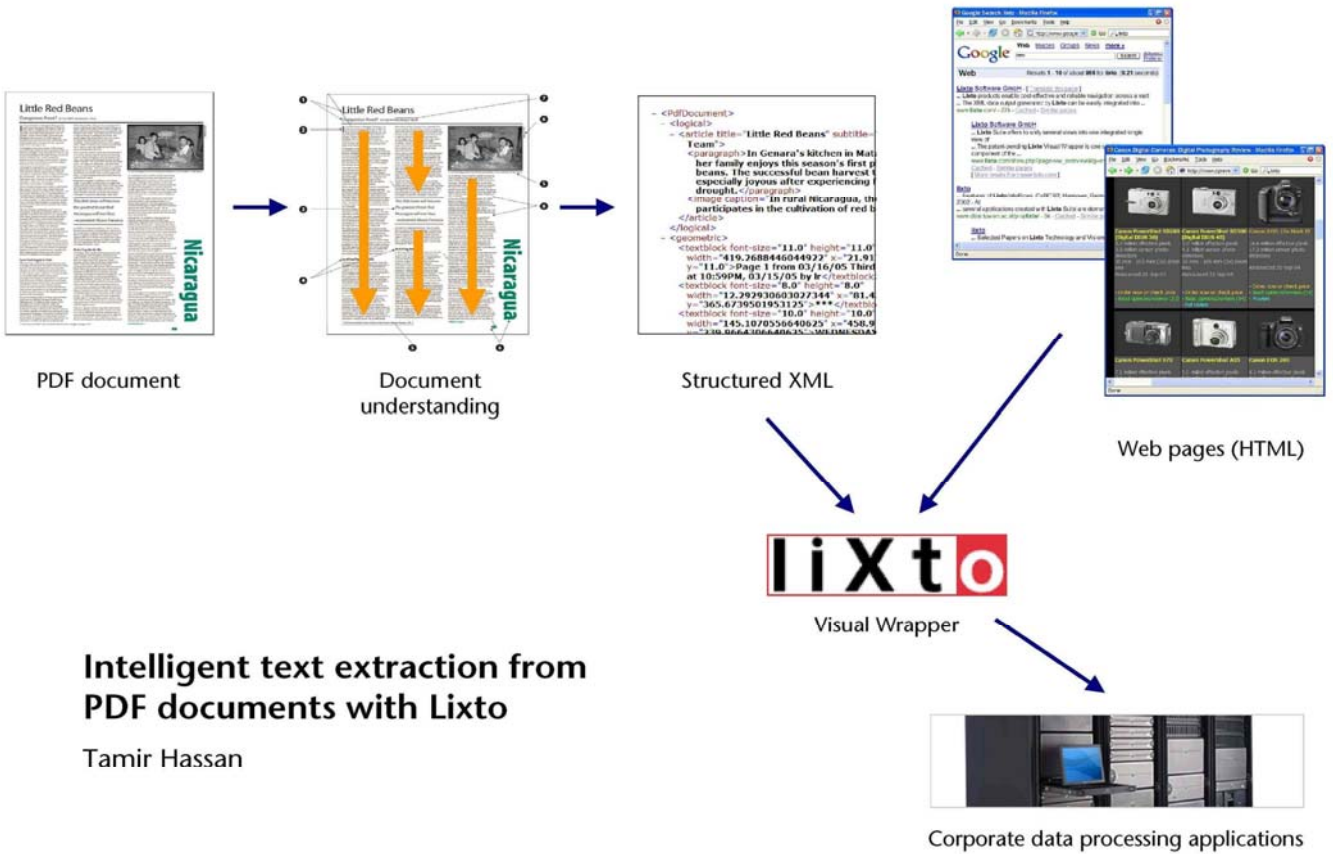
EXALG: Algorithm to solve the EXTRACT problem (which is to deduce the unknown template and values from a set of pages)

EXALG has two stages/modules:
 1) ECGM (Equivalence Class Generation M.)
 2) Analysis Module

Input: Set of pages (generated from a common template)
 Output: Template and set of values

Definition of several terms used in EXALG:

- LFEQ: Large and Frequently occurring Equivalence class
- Equivalence Class: Maximal set of tokens having the same occurrence-vector
- Occurrence-vector: vector containing the occurrences of a token in the input (set of pages)
- Role of a token: context in which it occurs
- Token: word or HTML tag; dtoken: differentiated t.



Informationsprozess

