Intermediate hand-ins
Tuesday, April 12, 2005

„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
• ...

Source: PS WIE
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

1. Syntactical Structure Normalization
   Cleans up syntactical incorrectness
   XWRAP transforms the given page to a parse tree or so-called syntactic token tree

2. Information Extraction
   Identifying interesting regions
   Identifying the important tokens in the retrieved document
   Identifying useful hierarchical structures of the document

3. Code Generation
   XML-templates for each rule

4. Testing and Packing
   Entering some URL’s
   Each URL will go through Task 1 to 3
   Accept or reject result
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.

get list- and detail pages
tokenize page sources

find and discard page templates

set up constraints for extracts belonging to records
solve CSP

probabilistic variables and dependencies
find best-matching model

performs better on static page layouts, periodic appearance of attributes

record segmentation

performs better on more dynamic data, allows assumption of record columns

René Kiesler
Record-Boundary Discovery in Web Documents

Christoph Veigl
Using the Structure of Web Sites for Automatic Segmentation of Tables
Mining Web Pages for Data Records

#7

Automatic Data Extraction from Lists and Tables in Web Sources
(Lecerf/Kollberg/Hidra)

Example pages (html)

Unsupervised learning algorithm

Page template

Tokens Extracted from page code

Y combing (using a set of distinctive tokens)

Columns

Dataset cells

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

#8

ndr - mining data records

algorithm to find and identify data regions and data records*

only works in structured html documents which must be designed using <tags>
A flexible learning system for koshering tables and lists in HTML documents
by William J. Cohen
Matthew Hunt
Lee J. Venven

NL2, a non-hidden learning system is applied on a website to extract data from tables and lists.

Various buildings that form part of NL2 explicit different characterizations of a document and in the process make data extraction more efficient and effective.

Extracted data from tables and lists.

Gateway from HTML to XML (Paper No. 10)
by Tao Fu & Mangzi Lu

Gateway

Transformed xml document

break apart

build together

<small>
<item>
<item>
<item>
<item>
</item>
</item>
</item>
</item>
</small>
### #13

**Main Goal:** Distinguish genuine tables from non-genuine ones.

**Classification of Tables**

- Decision-tree-Learning (via inductive inference)
- Support-Vector-Machines (hypothosis-based)

**Database of Webpages (Training Data)**

1. Unique id
2. Is it genuine? (yes/no)
3. TableTitle

---

**A Machine Learning Based Approach for Table Detection on the Web**

Gregor Pridun 9725753
Max Arends 9835411

---

### #14

**Data-rich Section Extraction from HTML Pages**

**I**

- Target Page
- Sample Page

**II**

- Parse HTML-page and convert tags into tag-trees

**III**

- Compare target page with the sample page-tree to identify their common parts
Automatic Ontology-based Knowledge Extraction from Web Documents

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

Automating the Extraction of Data from HTML Tables with Unknown Structure
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

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 SEPEROVIC ( e0325189 )
PS = WIE SS20D5
Paper #20
Applying Pattern Mining to Web Information Extraction

Jeremy Solarz

Extracting Structured Data from Web Pages

Tobias Dönz (0226 173)

Input

1. **Diff Format**
   - Differentiate roles of tokens (using format)

2. **Find Equiv**
   - Compute occurrence-vector of tokens, determine LFEQ
   - Eliminate invalid equivalence classes

3. **Hard Inv**
   - Differentiate roles of tokens (using LFEQs)

4. **Constr Temp**
   - Construct template using LFEQs

5. **Ex Val**
   - Extract values

**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) **ECG/M** (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; **token**: Differentiated t.
Intelligent text extraction from PDF documents with Lixto

Tamir Hassan