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

Document to extract
\(<p>\text{Name: } <b>\text{xxx} </b> <p>\)

content to extract
\(<p>\text{Name: } <b>\text{[xxx]} </b> <p>\)

\(<p>\text{Name: } <b>\text{-[xxx]-} </b> <p>\)

extraction rule
\(\text{SkipTo(Name Symbol HtmlTag)}\)

Is result ok?
Generate new extraction rule

Website HTML Code
\(<HTML> <HEAD> \- \- \- \)

XWRAP Wrapper
4 Modules do the major work

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

Xwrap Schema
Marco Schönig
XWRAP: An XML-enabled Wrapper Construction System for Web Information Sources
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

get list- and detail pages
tokenize page sources

find and discard
template 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
Mining Web Pages for Data Records

1. Data Extraction from Lists and Tables in Web Sources
   (Kaplan/Keen/Kain/Haralick)

Example pages

- Unsupervised learning algorithm

- Template

- Data page (html)

- Extracted data

Mining Web Pages

- HTML pages

- Tokens extracted from page code

- Columns

- Dataset cells

- Rows

- Dataset records / tuples of cells

- Still in token form → used as a template/wrapper for subsequent data extraction

#7

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

#8

Mdr - 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>

- Structured tag tree

- Document.html

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

- 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

Bohunsky Paul 0025058
A flexible learning System for Wrapping Tables and Lists in HTML Documents

by William H. Cohen
Matthew Hurst
Lee J. Venven

HLN, a wrapper learning system is applied on a website to extract data from tables & lists.

Extracted data from tables & lists.

A Gateway from HTML to XML (Paper No. 10)

by Toa Fu & mycket Liu

Gateway

break apart

build together

implemented scheme

transformed xml document
Database of webpages (training data) (tables get extra attributes in the database)
1) unique id
2) Is it genuine? (yes/no)
3) Table title

A Machine Learning Based Approach for Table Detection on the Web

Gregor Pridun 9725753
Max Arens 9835444

Data-rich Section Extraction from HTML pages

I Find a page sample (common display layout) Take the outgoing links of the target page

II Parse the HTML page and convert them into tag trees

III Compare target page with the sample page tree to identify their common parts
12

#15

Automatic Ontology-based Knowledge Extraction from Web Documents

Stefan Bischof

13

#16

Automating the Extraction of Data
from HTML Tables with Unknown Structure

Stefan Rümmel
0325665
Data Extraction and Label Assignment for Web Databases

(Jiyong Wang, Frederick H. Lochovsky)

to build a System that
extracts (automatically) text from a web-page into a table
assignes labels in a table

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
Jeremy Solarz

Applying Pattern Mining to Web Information Extraction

Extracting Structured Data from Web Pages

Tobias Dönz (0226 173)

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

ExAlgo 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 ExAlgo:
- 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.
Intelligent text extraction from PDF documents with Lixto

Tamir Hassan

Informationsprozess

Extraktion Aggregation Filterung Visualisierung

Newswrapper for Brokers

Martin Zeilinger