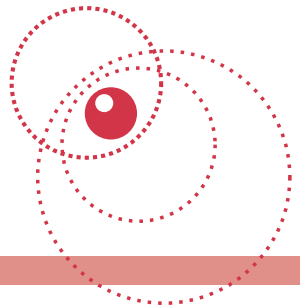


Structured Web Content Extraction

24 November 2015

TELECOM ParisTech

Plan

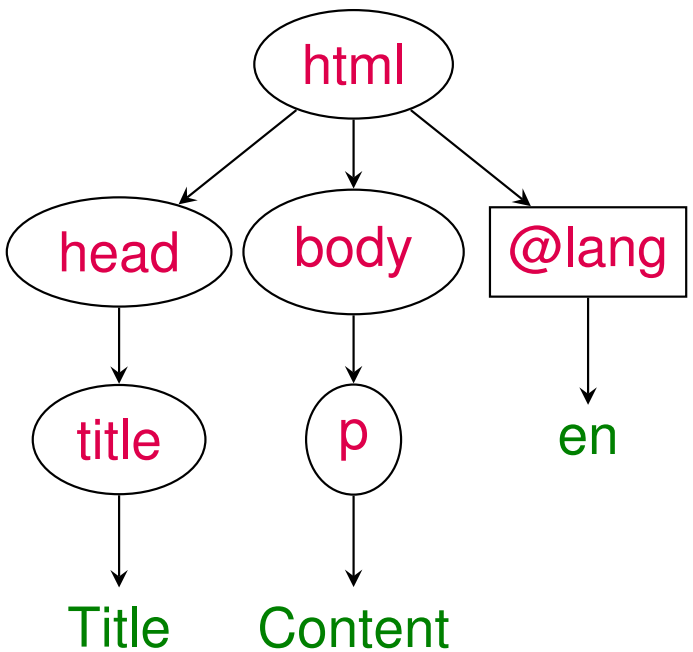
Plan

Document Object Model (DOM)

Tree representation of an HTML document, suitable for manipulation and extraction.

Example

```
<html lang="en">
<head><title>Title</title></head>
<body><p>Content</p></body>
</html>
```



Languages for extraction

- Based on serialization: regular expressions (see further)
- Based on DOM:

  DOM navigation expresses local navigation in the DOM, from a node to its parent, its children, its attribute, etc. Standard API [W3C] but variations.

  searching elements by tag names, identifiers, names, class names

  CSS selectors (see further)

  XPath (see further)

Plan

Regular Expressions

- Apply to the serialized representation, not to the DOM tree.
- Available in a wide range of host languages (including Python with the `re` package).
- The following characters are metacharacters.

```
? * + | ( ) ^ $ . [ ] { } " \
```

- Metacharacters have special meaning; they do not represent themselves.
- All other characters represent themselves.

Operators

r	One occurrence of r
$r?$	Zero or one occurrence of r
$r*$	Zero or more occurrences of r
$r+$	One or more occurrences of r
$r \mid s$	r or s
rs	r concatenated with s

r and s are regular expressions.

Grouping and extra symbols

- Parentheses are used for grouping.
- The expression

```
("+"|"-")?
```

  represents an optional plus or minus sign.
- If a regular expression begins with `^`, then it is matched only at the beginning of a line or string (depending on context).
- If a regular expression ends with `$`, then it is matched only at the end of a line or string (depending on context).
- The dot `.` matches any non-newline character.

Character groups

- Brackets `[ ]` match any single character listed within the brackets.
- For example,
  - `[abc]` matches `a` or `b` or `c`.
  - `[A-Za-z]` matches any letter.
- If the first character after `[` is `^`, then the brackets match any character *except* those listed.
  - `[^A-Za-z]` matches any nonletter.

Plan

Simple, multiple, universal selectors

Simple selector: tag name

Multiple selector: several selectors joined by commas

Universal selector: '`*`', selects everything

Examples

- `ul` selects unordered lists
- `h1,h2,h3,h4,h5,h6` selects all section titles
- `*` selects everything

Class selectors

Class selector: class name, prefixed with '.', as it appears in a `class` attribute of an HTML tag

Examples

- `.person` selects all tags with class `person`
- `p.comment` selects all `<p>` tags with class `comment`

Identifier selector

Identifier: as defined by the `id` attribute of an HTML tag. Similar to classes, but only one tag with a given `id` in the whole HTML document

Identifier selector: identifier name, prefixed with '#', as it appears in the `id` attribute of an HTML tag

Examples

- `#introduction` selects the tag with identifier `introduction`
- `p#introduction` selects the `<p>` tag with identifier `introduction`

Contextual selectors

Contextual selector: 2 selectors or more separated by spaces. *A B* selects *B*'s only if they are contained in *A*'s

Child selector: 2 selectors separated by >. $A > B$ selects *B*'s children of *A*'s

Next sibling selector: 2 selectors separated by +. $A + B$ selects *B*'s that are the next sibling of an *A*

Examples

- `h1 em` selects text in emphasis within a main title
- `ul ol, ol ul, ul ul, ol ol` selects nested lists

Pseudo-class

Pseudo-class: specify some external properties of a class

Examples

- `article > p:first-child` selects all paragraphs that are first-children of an `<article>`

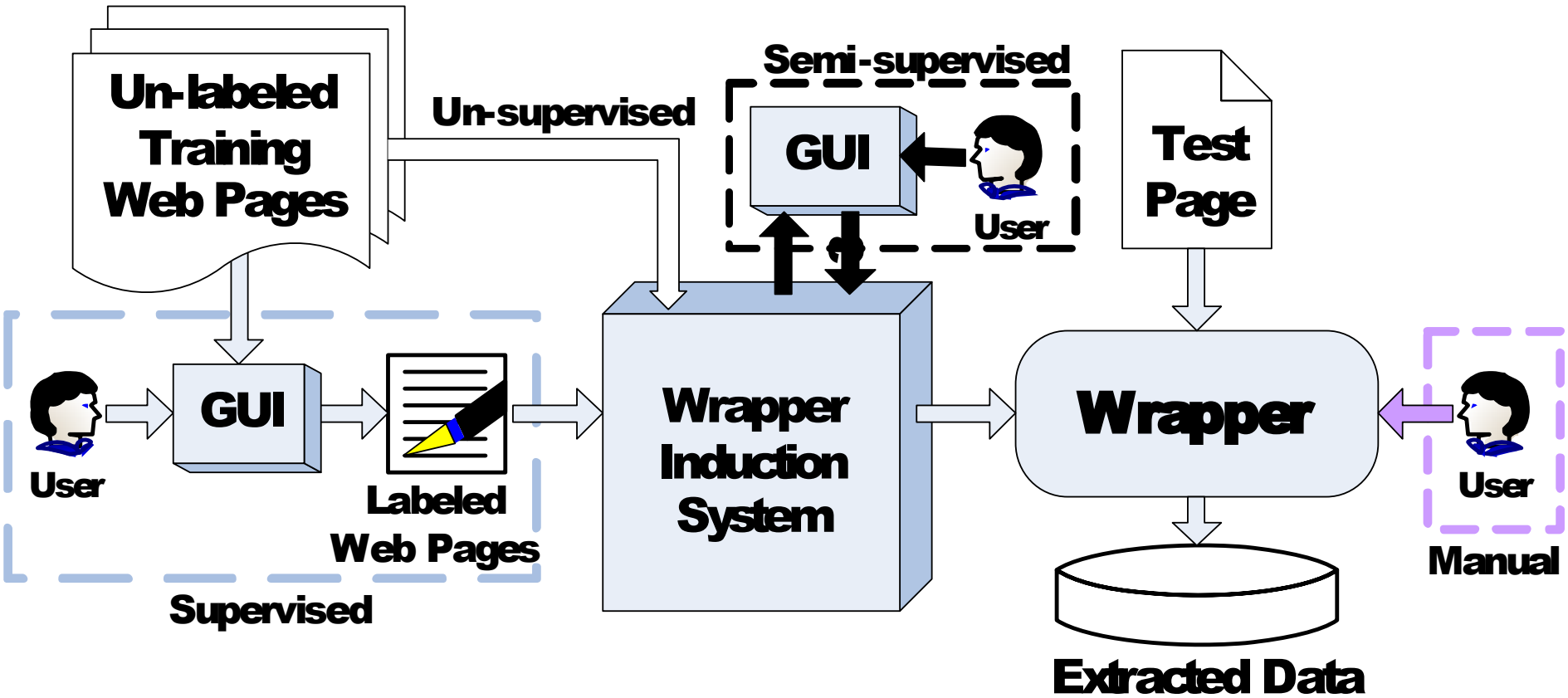
Plan

cf. separate set of slides

Manual Selection and Extraction Techniques

Wrapper induction

Principles [Chang et al., 2006]



Supervised, semi-supervised, and domain-based techniques

- Many academic approaches and systems
- No ready-to-use free software for supervised and semi-supervised extraction (as far as I know)
- Existing companies selling wrapper induction software: Lixto (semi-supervised), Wrapidity (domain-based)

Unsupervised techniques

- Exploiting data redundance within a page [Liu et al., 2004] or across pages [Crescenzi et al., 2001, Arasu and Garcia-Molina, 2003]
- RoadRunner: freely downloadable and existing demos at `http://www.dia.uniroma3.it/db/roadRunner/`

Bibliography I

Arvind Arasu and Hector Garcia-Molina. Extracting structured data from Web pages. pages 337–348, June 2003.

Chia-Hui Chang, Mohammed Kayed, Mohem Ramzy Girgis, and Khaled F. Shaalan. A survey of Web information extraction systems. 18(10):1411–1428, October 2006.

Valter Crescenzi, Giansalvatore Mecca, and Paolo Merialdo. RoadRunner: Towards Automatic Data Extraction from Large Web Sites. 2001.

Bing Liu, Robert L. Grossman, and Yanhong Zhai. Mining Web Pages for Data Records. *IEEE Intelligent Systems*, 19(6):49–55, 2004.

W3C. Document Object Model. `http://w3.org/DOM`.

Licence de droits d'usage

Contexte public } avec modifications

Par le téléchargement ou la consultation de ce document, l'utilisateur accepte la licence d'utilisation qui y est attachée, telle que détaillée dans les dispositions suivantes, et s'engage à la respecter intégralement.

La licence confère à l'utilisateur un droit d'usage sur le document consulté ou téléchargé, totalement ou en partie, dans les conditions définies ci-après et à l'exclusion expresse de toute utilisation commerciale.
Le droit d'usage défini par la licence autorise un usage à destination de tout public qui comprend :
– le droit de reproduire tout ou partie du document sur support informatique ou papier,
– le droit de diffuser tout ou partie du document au public sur support papier ou informatique, y compris par la mise à la disposition du public sur un réseau numérique,
– le droit de modifier la forme ou la présentation du document,
– le droit d'intégrer tout ou partie du document dans un document composite et de le diffuser dans ce nouveau document, à condition que :
  – L'auteur soit informé.

Les mentions relatives à la source du document et/ou à son auteur doivent être conservées dans leur intégralité.

Le droit d'usage défini par la licence est personnel et non exclusif.
Tout autre usage que ceux prévus par la licence est soumis à autorisation préalable et expresse de l'auteur : sitepedago@telecom-paristech.fr