# FOREST: Focused Object Retrieval by Exploiting Significant Tag paths

Marilena Oita    Pierre Senellart

*WebDB*, May 31, 2015

# Problem Definition

Goal: extracting interesting content from Web pages
$\approx$ eliminating boilerplate

The problem has been viewed from different point of views...

- **Text Extraction?** extracting text, yes, but not not any kind of text (e.g., BOILERPIPE)
- **Information Extraction?** extracting "values" out of a common structure (e.g., wrapper induction)
- **Information Retrieval?** having some keywords, extract relevant data (not pages, but well-defined objects $\rightarrow$ the aim of FOREST)

# Problem Definition

Goal: extracting interesting content from Web pages
$\approx$ eliminating boilerplate

The problem has been viewed from different point of views...

- Text Extraction? extracting text, yes, but not not any kind of text (e.g., BOILERPIPE)
- Information Extraction? extracting "values" out of a common structure (e.g., wrapper induction)
- Information Retrieval? having some keywords, extract relevant data (not pages, but well-defined objects → the aim of FOREST)

# Problem Definition

Goal: extracting interesting content from Web pages
$\approx$ eliminating boilerplate

The problem has been viewed from different point of views...

- Text Extraction? extracting text, yes, but not not any kind of text (e.g., BOILERPIPE)
- Information Extraction? extracting "values" out of a common structure (e.g., wrapper induction)
- Information Retrieval? having some keywords, extract relevant data (not pages, but well-defined objects $\rightarrow$ the aim of FOREST)

# Problem Definition

Goal: extracting interesting content from Web pages
$\approx$ eliminating boilerplate

The problem has been viewed from different point of views...

- Text Extraction? extracting text, yes, but not not any kind of text (e.g., BOILERPIPE)
- Information Extraction? extracting “values” out of a common structure (e.g., wrapper induction)
- Information Retrieval? having some keywords, extract relevant data (not pages, but well-defined objects $\rightarrow$ the aim of FOREST)

# Problem Definition

Goal: extracting interesting content from Web pages
$\approx$ eliminating boilerplate

The problem has been viewed from different point of views…

- Text Extraction? extracting text, yes, but not not any kind of text (e.g., BOILERPIPE)
- Information Extraction? extracting "values" out of a common structure (e.g., wrapper induction)
- Information Retrieval? having some keywords, extract relevant data (not pages, but well-defined objects $\rightarrow$ the aim of FOREST)

# Problem Definition

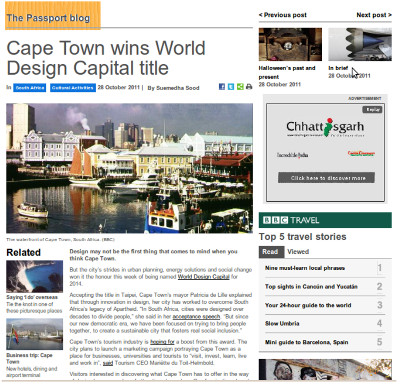
Goal: extracting interesting content from Web pages
$\approx$ eliminating boilerplate

The problem has been viewed from different point of views...

- Text Extraction? extracting text, yes, but not not any kind of text (e.g., BOILERPIPE)
- Information Extraction? extracting "values" out of a common structure (e.g., wrapper induction)
- Information Retrieval? having some keywords, extract relevant data (not pages, but well-defined objects $\rightarrow$ the aim of FOREST)

# Use Case: Dynamic Pages

## BBC *travel* blog news pages

# Wrapper Induction

## (extracting data from structured sources)

unsupervised wrapper induction: given a set of objects, infer a wrapper procedure (by grammar rules/ XPath) for extracting the data values of these objects

Typical: deep Web response pages $\rightarrow$ records (e.g., Amazon books)
Here: blogs, news, social media $\rightarrow$ objects (e.g., posts, events, tweets)

# Outline

Context

Keyword-based Relevance

Experiments

Conclusions

# Distinguishing the Main Content

## using linguistic clues

A Day of the Dead offering in the Nunkini cemetery, in Campeche, Mexico. (Jeffrey Becom/LPI)

### Related

**Worldwide weird: Halloween to the extreme**

Taking normally odd rituals one step further

**Oktoberfest: Then and now**

It has not changed that much over time

**On the last night of the autumn harvest, the world changes from the sunny warmth of summer to the cold dark of winter, the land from fertile to barren. The ancient Celts believed this transition gave supernatural forces a chance to break through into the world of the living, and their evil mischief to flourish.**

They came to celebrate the night leading into winter as Samhain (meaning "summer's end"), the festival widely considered to be the precursor of Halloween. On Samhain night, the Celts believed, the spirits of people who had died in the past year would walk among the living, so. villagers put out food and sweets to pacify these spirits – a ritual that may have preceded trick-or-treating. (There is no hard evidence, however, that Samhain was indeed a festival of the dead, points out historian Nicholas Rogers, in his book Halloween: From Pagan Ritual to Party Night.)

Although Halloween has pagan origins, its name is derived from the Christian holiday "All Hallows Eve", or the evening before All Saints' Day (1 November). The holiday itself was adapted by Christians who hoped to stamp out paganism, and over the years, some of the darker aspects of Halloween have been replaced by more light-hearted, family-friendly festivities. But Halloween's ties with the scary and supernatural still hold strong today, in celebrations all over the world.

**Ireland**

In Ireland, arguably the holiday's birthplace, Halloween is still greeted

"Halloween, past and present"

# Keywords

- Model a Web page as its DOM (Document Object Level) tree
- Distinguish significant content nodes
  - textual DOM leaf nodes
  - at least one keyword
- Keywords automatically acquired:
  - *Tf-Idf* analysis
    - IN: set of sample Web pages
    - OUT: top-$k$ tf-idf weighted terms
  - basic text preprocessing on feed item metadata to identify top-$k$ feed keywords

# Keywords

- Model a Web page as its DOM (Document Object Level) tree
- Distinguish significant content nodes
  - textual DOM leaf nodes
  - at least one keyword
- Keywords automatically acquired:
  - *Tf-Idf* analysis
    - IN: set of sample Web pages
    - OUT: top-$k$ tf-idf weighted terms
  - basic text preprocessing on feed item metadata to identify top-$k$ feed keywords

# Keywords

- Model a Web page as its DOM (Document Object Level) tree
- Distinguish significant content nodes
  - textual DOM leaf nodes
  - at least one keyword
- Keywords automatically acquired:
  - *Tf-Idf* analysis
    - IN: set of sample Web pages
    - OUT: top-$k$ tf-idf weighted terms
  - basic text preprocessing on feed item metadata to identify top-$k$ feed keywords

# Keywords

- Model a Web page as its DOM (Document Object Level) tree
- Distinguish significant content nodes
  - textual DOM leaf nodes
  - at least one keyword
- Keywords automatically acquired:
  - *Tf-Idf* analysis
    - IN: set of sample Web pages
    - OUT: top-$k$ tf-idf weighted terms
  - basic text preprocessing on feed item metadata to identify top-$k$ feed keywords

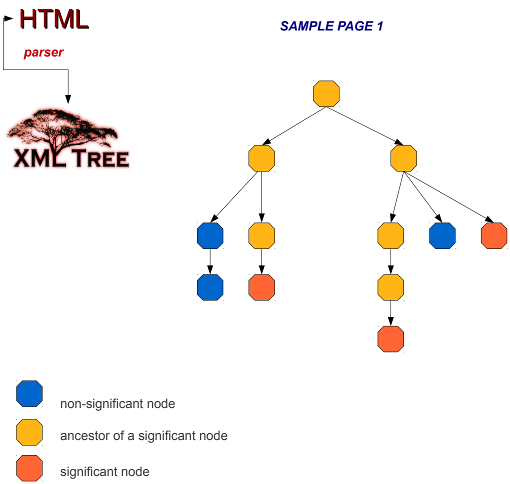
# Keywords

- Model a Web page as its DOM (Document Object Level) tree
- Distinguish significant content nodes
  - textual DOM leaf nodes
  - at least one keyword
- Keywords automatically acquired:
  - *Tf-Idf* analysis
    - IN: set of sample Web pages
    - OUT: top-$k$ tf-idf weighted terms
  - basic text preprocessing on feed item metadata to identify top-$k$ feed keywords

# Standard Feed Metadata

```
−<rss version="2.0">
  −<channel>
      <title>WORLDMag.com</title>
      <link>http://www.worldmag.com</link>
    +<description></description>
      <pubDate>Sun, 14 Aug 2011 03:20:18 GMT</pubDate>
      <language>en</language>
    +<item></item>
    +<item></item>
    −<item>
        <title>Driver wanted</title>
        <link>http://www.worldmag.com/webextra/18476</link>
        <guid isPermaLink="true">http://www.worldmag.com/webextra/18476</guid>
        <pubDate>Thu, 11 Aug 2011 11:34:01 GMT</pubDate>
        <dc:creator>Joel Hannahs</dc:creator>
      −<description>
          The 'Values Bus' rolls through Iowa in search of a leader on key issues
        </description>
      </item>
    +<item></item>
```

# Example: Sample Page I



# DOM Element Identification

## DOM element identifier

- tagName
- $\langle$attributes$\rangle$
- node index of depth-first search traversal: dfs

## Structural pattern $\rightarrow sp_i, i \in 1 : n$

- an *XPath expression*
- describes an identifier of a DOM node which is significant
- example:

```
//div[@id='wrapper' and @class='article']
//p[@dfs=24]
```

# DOM Element Identification

DOM element identifier

- tagName
- ⟨attributes⟩
- node index of depth-first search traversal: dfs

Structural pattern $\rightarrow sp_i, i \in 1:n$

- an *XPath expression*
- describes an identifier of a DOM node which is significant
- example:

```
//div[@id='wrapper' and @class='article']
//p[@dfs=24]
```

# DOM Element Identification

## DOM element identifier

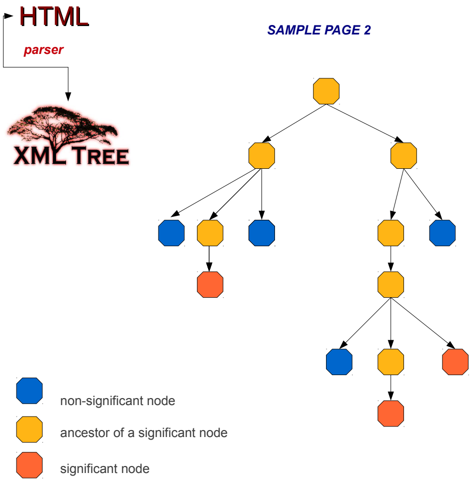
- tagName
- $\langle$attributes$\rangle$
- node index of depth-first search traversal: dfs

## Structural pattern $\rightarrow sp_i, i \in 1 : n$

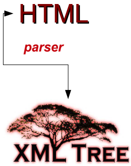
- an *XPath expression*
- describes an identifier of a DOM node which is significant
- example:

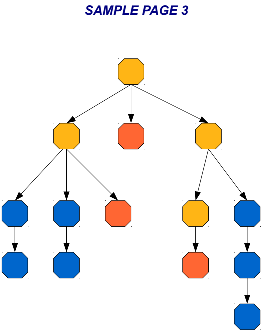
```
//div[@id='wrapper' and @class='article']
                //p[@dfs=24]
```

# Sample Page II



# Sample Page III



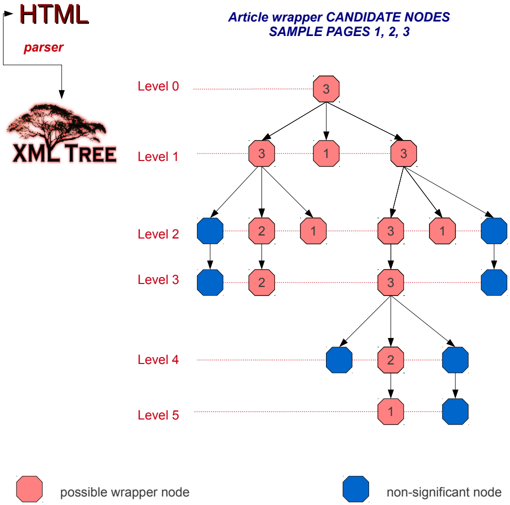




non-significant node

ancestor of a significant node

significant node

# Aggregating across Pages: Candidate Patterns



# Ranking Significant DOM nodes

## Keyword density

- $x$= nb. keywords
- $y$= nb. non-significant terms
- $N$= total number of terms

$\rightarrow \frac{x}{N}$

## Statistical Corrections:

- $N$ can be small for some nodes
  - *Jeffrey's add-half* estimator $f = \frac{x+1/2}{N+1}$
- $N$ potentially large set $\rightarrow$ margin of error
  - confidence interval of 1
  - one standard deviation (margin of error at 70%) $\sqrt{\frac{f(1-f)}{N}}$

# Ranking Significant DOM nodes

## Keyword density

- $x$= nb. keywords
- $y$= nb. non-significant terms
- $N$= total number of terms

$\rightarrow \frac{x}{N}$

## Statistical Corrections:

- $N$ can be small for some nodes
  - *Jeffrey's add-half* estimator $f = \frac{x+1/2}{N+1}$
- $N$ potentially large set $\rightarrow$ margin of error
  - confidence interval of 1
  - one standard deviation (margin of error at 70%) $\sqrt{\frac{f(1-f)}{N}}$

# Ranking Significant DOM nodes

## Keyword density

- $x =$ nb. keywords
- $y =$ nb. non-significant terms
- $N =$ total number of terms

$\rightarrow \frac{x}{N}$

## Statistical Corrections:

- $N$ can be small for some nodes
  - *Jeffrey's add-half* estimator $f = \frac{x+1/2}{N+1}$
- $N$ potentially large set $\rightarrow$ margin of error
  - confidence interval of 1
  - one standard deviation (margin of error at 70%) $\sqrt{\frac{f(1-f)}{N}}$



TELECOM ParisTech

# Ranking Significant DOM nodes

## Keyword density

- $x=$ nb. keywords
- $y=$ nb. non-significant terms
- $N=$ total number of terms

$\rightarrow \frac{x}{N}$

## Statistical Corrections:

- $N$ can be small for some nodes
  - *Jeffrey's add-half* estimator $f = \frac{x+1/2}{N+1}$
- $N$ potentially large set $\rightarrow$ margin of error
  - confidence interval of 1
  - one standard deviation (margin of error at 70%) $\sqrt{\frac{f(1-f)}{N}}$

# Statistical Keyword Density

$$f \pm \sqrt{\frac{f(1-f)}{N}} = \frac{x+1/2}{N+1} \pm \frac{1}{N+1}\sqrt{\frac{(x+1/2)\times(y+1/2)}{N}}$$

$$J = \max\left(0, \frac{1}{N+1}\left(x+1/2-\sqrt{\frac{(x+1/2)\times(y+1/2)}{N}}\right)\right)$$

# Unexpectedness

@DOM node level

- $x=$ keywords
- $y=$ non-significant terms

@Web page level $\rightarrow X, Y$
global context

unexpected content: simpler to describe than to generate

*generation* complexity $C_w = (x + y) \log (X + Y)$
*description* complexity $C = x \log X + y \log Y$

$$U = C_w - C$$

# Unexpectedness

@DOM node level

- $x=$ keywords
- $y=$ non-significant terms

@Web page level $\rightarrow X, Y$
global context

unexpected content: simpler to describe than to generate

*generation* complexity $C_w = (x+y)\log(X+Y)$
*description* complexity $C = x\log X + y\log Y$

$$U = C_w - C$$

# Unexpectedness

@DOM node level

- $x$= keywords
- $y$= non-significant terms

@Web page level $\rightarrow X, Y$
global context

unexpected content: simpler to describe than to generate

*generation* complexity $C_w = (x + y) \log(X + Y)$
*description* complexity $C = x \log X + y \log Y$

$$U = C_w - C$$

# Node Content Informativeness Metric

- measures the interest of a structural pattern $sp_i$
- in a document $d_k$

$$I(sp_i, d_k) = J(sp_i, d_k) \times U(sp_i, d_k)$$

I: informativeness

J: statistical semantic density

U: unexpectedness

# Ranking of Structural Patterns

$$R(sp_i) = \sum_{k=0}^{m} I(sp_i, d_k) \times \mathrm{level}(sp_i) \times \mathrm{nbOcc}_i$$

- Allows ranking generic XPath expressions
  - `//div[@class='wrapper' and (@dfs='27' or @dfs='31')]`
- Final output: subtrees extracted from DOM trees of Web pages


TELECOM ParisTech

# Ranking of Structural Patterns

$$R(sp_i) = \sum_{k=0}^{m} I(sp_i, d_k) \times \text{level}(sp_i) \times \text{nbOcc}_i$$

- Allows ranking generic XPath expressions
  - `//div[@class='wrapper' and (@dfs='27' or @dfs='31')]`
- Final output: subtrees extracted from DOM trees of Web pages

# Outline

Context

Keyword-based Relevance

Experiments

Conclusions

# FOREST Methods

- FOREST$_{\text{info}}$: keywords acquired through tf-idf analysis
- FOREST$_{\text{feed}}$: keywords acquired through feed meta-information

# 3 Baselines with Alternative Design Choices

- $\text{FOREST}_{\text{Cov}}$
  - Same framework as FOREST
  - But score of a pattern is just (normalized) tf-idf weighting of the corresponding DOM nodes
- AbsElems
  - Same framework as $\text{FOREST}_{\text{info}}$
  - Consider only patterns returning significant leaves, not those that are ancestors of significant leaves
- AbsPaths
  - Same as AbsElems
  - Elements are identified in a pattern by their root-to-leaf path

# 3 Baselines from the Literature

- CETR (WWW, 2010)
  - clustering technique based on a tag ratio per line of the HTML file
  - relies on the fact that text is denser in the main content
- BOILERPIPE (WSDM, 2010)
  - machine learning to determine rules to classify text as content/not-content
  - relies on shallow text features
- DESCRIPTION
  - Main test is just content of the description metadata within a Web feed

# Datasets

- CYAN dataset: dataset used for evaluation of CETR (only 9 different Web sites)
- RED dataset
  - publicly available
  - http://dbweb.enst.fr/software/
  - feed-based (search4Rss); crawl of Web pages referred through feed items
  - manual annotation of the corpus
    - 90 Web sites and 1006 Web pages
    - gold standard: fulltext + metadata (title, author, categories etc.)

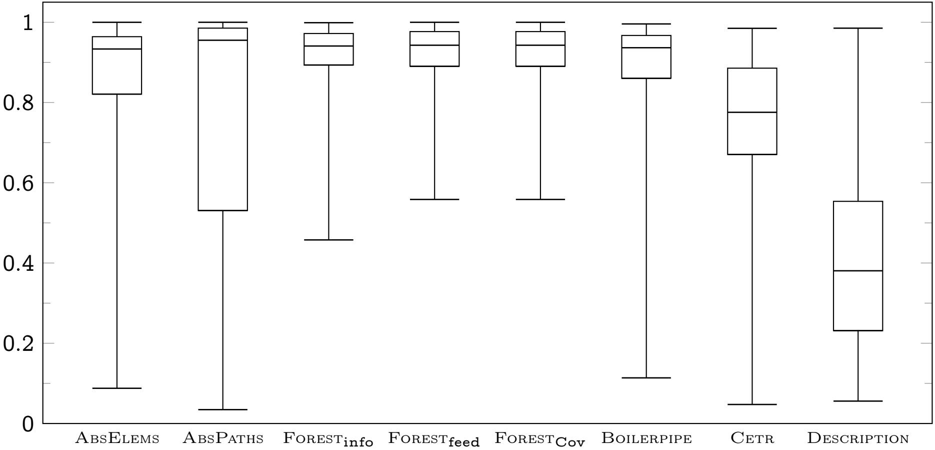
# Datasets

- CYAN dataset: dataset used for evaluation of CETR (only 9 different Web sites)
- RED dataset
  - publicly available
  - `http://dbweb.enst.fr/software/`
  - feed-based (search4Rss); crawl of Web pages referred through feed items
  - manual annotation of the corpus
    - 90 Web sites and 1006 Web pages
    - gold standard: fulltext + metadata (title, author, categories etc.)

# Results (I): Precision, Recall

| | CYAN | | RED | |
|---|---|---|---|---|
| | Prec. | Rec. | Prec. | Rec. |
| AbsElems | 65 | 68 | 87 | 93 |
| AbsPaths | 58 | 64 | 72 | 74 |
| $\text{Forest}_{\text{info}}$ | 87 | 99 | 88 | 98 |
| $\text{Forest}_{\text{feed}}$ | | | 86 | 98 |
| $\text{Forest}_{\text{Cov}}$ | 78 | 89 | 88 | 98 |
| Boilerpipe | 94 | 97 | 89 | 90 |
| CETR | 65 | 95 | 67 | 93 |
| Description | | | 92 | 31 |

# Results (II): Box Chart of $F_1$ Measure on RED




9th and 91th percentile (whiskers), first and third quartile (box) and median (horizontal rule)

# Outline

Context

Keyword-based Relevance

Experiments

Conclusions

# Contributions

- A fully automatic, robust, effective algorithm for article extraction from dynamically generated Web pages
- Original use of statistical and information-theory–based relevance measures
- Versatile algorithm: has been applied to deep Web object extraction as well (VLDS 2012)
- Requires a source of keywords, which can be external (feed metadata) or internal (informative words on the page itself)
- Extensive and freely available dataset

Merci !

# Contributions

- A fully automatic, robust, effective algorithm for article extraction from dynamically generated Web pages
- Original use of statistical and information-theory–based relevance measures
- Versatile algorithm: has been applied to deep Web object extraction as well (VLDS 2012)
- Requires a source of keywords, which can be external (feed metadata) or internal (informative words on the page itself)
- Extensive and freely available dataset

Merci !