Content-based Title Extraction from Web Page

Najlah Gali and Pasi Fränti
Motivation
Application: search engine
Summary Extraction

• **Title**  *Rosso restaurant*, “City pharmacy”
• **Keywords**  *restaurant, food, lunch, dinner*
• **Representative Image**
• **Short description**
  ma-pe: 16.00-22.00
  la: 12.00-22.00
  puh. 013 227 874
What we deal with
Content of Web Page
Hypertext Markup Language (HTML, XHTML)

Logo image
Title
Keywords
Text

Navigation bar
Images

Raspberry Pi 3 adds wi-fi and Bluetooth

The Raspberry Pi has become the most popular British computer yet made.

The title was formerly held by the Amstrad PCW which is believed to have sold a total of eight million units.

Sales of the Raspberry Pi will surpass that figure this month, said the Raspberry Pi project founder Eben Upton.

Top Stories
Oscars 2016: DiCaprio finally wins
1 hour ago
UN to expand Syria aid as truce holds
30 minutes ago
Pakistan hangs killer of state governor
4 hours ago

Features & Analysis
He said yes!
Eight women who proposed to their partners
Web Page Title
The title can be in three different places:

Title Tag
<title>Wentworth House Hotel Bath Hotels - Cheap Hotels in Bath, Somerset, UK</title>

Logo image

Web Page body
Segmenting title tags
Title and Meta Tags

• The obvious source
• But includes also additional information

  – <title> Joensuu Keskusta | Intersport - Sport to the people</title>

• Segmentation is needed!
  Joensuu Keskusta
  Intersport
  Sport to the people
Work flow

https://www.jdwetherspoon.com/pubs/all-pubs/england/london/the-coronet-holloway

1. Placement in title & meta tags
2. Popularity in header tags
3. Position in the web link

Title  The coronet
Content of Title and Meta tags

<title>The Coronet, Holloway | Our Pubs | J D Wetherspoon</title>

<meta name="keywords" content="The Coronet" />
Segmentation by delimiters

<title>Sydney Waterfront Restaurant | Restaurant Milsons Point - Aqua Dining</title>

<title>SIGNORELLI GASTRONOMIA - Pyrmont Italian Restaurant - EAT • DRINK • SHOP • COOKItalian Restaurant Pyrmont Sydney – Signorelli Gastronomia</title>

<title>Neutral Bay Club | Tennis, Bowls, Bistro & Functions | Sydney</title>

<title>The Coronet, Holloway | Our Pubs | J D Wetherspoon</title>

Pre-defined delimiter patterns

<table>
<thead>
<tr>
<th>space – space</th>
<th>space / space</th>
<th>space . space</th>
</tr>
</thead>
<tbody>
<tr>
<td>space : space</td>
<td>, space</td>
<td>space -</td>
</tr>
<tr>
<td>: space</td>
<td>space :</td>
<td>space</td>
</tr>
<tr>
<td>space &gt;</td>
<td>space «</td>
<td>space »</td>
</tr>
<tr>
<td>? ,</td>
<td>- ,</td>
<td>space ::</td>
</tr>
<tr>
<td>Space /</td>
<td>-</td>
<td></td>
</tr>
</tbody>
</table>
Candidate Segments

Candidates
• The Coronet
• Holloway
• Our Pubs
• J D Wetherspoon
Scoring the candidates
1. Position in Title and Meta Tags

- Appear first or last either in Title or Meta gets 0.1

```
<title>The Coronet, Holloway | Our Pubs | J D Wetherspoon</title>

Or

<meta name="keywords" content="The Coronet" />
```

**Candidates**

- The Coronet  0.1
- Holloway  0.0
- Our Pubs  0.0
- J D Wetherspoon  0.1
2. Popularity Among Header Tags

<h1 class="banner-inner__title">The Coronet</h1>
<h2 class="venue-finder__title-text">Find a pub or hotel</h2>
<h2 class="venue-finder__title-text">Our Pubs</h2>
<h2 class="venue-finder__title-text" ng-hide="isPubName">Check out your nearest pub or hotel</h2>
<h3 class="feature-panel__title">Discover our food menu</h3>
<h3 class="feature-panel__title">Our drinks selection</h3>
<h4 class="tab__title">Nearby J D Wetherspoons</h4>

**Candidates**

- The Coronet \(1 \times 6 = 6\)
- Holloway \(0\)
- Our Pubs \(1 \times 5 = 5\)
- J D Wetherspoons \(1 \times 3 = 3\)

Frequency: <h1>6</h1>  <h2>5</h2>  <h3>4</h3>  <h4>3</h4>  <h5>2</h5>  <h6>1</h6>

Weight: <h1>6</h1>  <h2>5</h2>  <h3>4</h3>  <h4>3</h4>  <h5>2</h5>  <h6>1</h6>
3. Position in Web Link

<table>
<thead>
<tr>
<th>Domain</th>
<th>Path</th>
<th>File name</th>
</tr>
</thead>
<tbody>
<tr>
<td><a href="https://www.jdwetherspoon.com/">https://www.jdwetherspoon.com/</a></td>
<td>pubs/all-pubs/england/london/</td>
<td>the-coronet-holloway</td>
</tr>
<tr>
<td>1</td>
<td>1.5</td>
<td>3</td>
</tr>
</tbody>
</table>

Dice similarity measure

**Candidates**

- The Coronet: $3 \times 0.70 = 2.10$
- Holloway: $3 \times 0.58 = 1.74$
- Our Pubs: $1.5 \times 0.00 = 0.00$
- J D Wetherspoon: $1 \times 1.00 = 1.00$
## Rank Segments

<table>
<thead>
<tr>
<th>Candidates</th>
<th>Position in tag</th>
<th>Popularity among header tags</th>
<th>Position in web link</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>The Coronet</td>
<td>0.1</td>
<td>6</td>
<td>2.10</td>
<td>2.10</td>
</tr>
<tr>
<td>Holloway</td>
<td>0.0</td>
<td>0</td>
<td>1.74</td>
<td>1.74</td>
</tr>
<tr>
<td>Our Pubs</td>
<td>0.0</td>
<td>5</td>
<td>0.00</td>
<td>0.00</td>
</tr>
<tr>
<td>J D Wetherspoon</td>
<td>0.1</td>
<td>3</td>
<td>1.00</td>
<td>1.00</td>
</tr>
</tbody>
</table>

### Normalizing

<table>
<thead>
<tr>
<th>Candidates</th>
<th>Position in tag</th>
<th>Popularity among header tags</th>
<th>Position in web link</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>The Coronet</strong></td>
<td>0.1</td>
<td>1.00</td>
<td>1.00</td>
</tr>
<tr>
<td>J D Wetherspoon</td>
<td>0.1</td>
<td>0.50</td>
<td>0.48</td>
</tr>
<tr>
<td>Holloway</td>
<td>0.0</td>
<td>0.00</td>
<td>0.83</td>
</tr>
<tr>
<td>Our Pubs</td>
<td>0.0</td>
<td>0.83</td>
<td>0.00</td>
</tr>
</tbody>
</table>
Experiments
Data Set

• Websites: 1245 http://cs.uef.fi/mopsi/titleextraction/data/ (in this paper)
  http://cs.uef.fi/mopsi/titler/ (extended)

• Categories: 8
  food & drink, home & garden, accommodation & hotels, shopping, arts & entertainment, hobbies & leisure, sport, and health & social care

• Sources: Google and Google maps

• Collected: 18–31.7.2014 + 19-23.4.2015

• Ground truth titles are manually extracted according to the specifications in [Hu et al. 2005]
Impact of criteria

**Criterion 1:** lowest impact (0.65)
- Generic words (*home, welcome*) often at the beginning
- Slogan or address often at the end.

**Criterion 2:** slightly higher impact (0.68)
- Heading tags not always used, not always correct title there.

**Criterion 3:** Best (0.84).

<table>
<thead>
<tr>
<th>Criteria</th>
<th>Average similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Position in tag</td>
<td>0.65</td>
</tr>
<tr>
<td>(2) Popularity among <em>hx</em> tags</td>
<td>0.68</td>
</tr>
<tr>
<td>(3) Position in web link</td>
<td><strong>0.84</strong></td>
</tr>
<tr>
<td>1 + 2</td>
<td>0.70</td>
</tr>
<tr>
<td>1 + 3</td>
<td><strong>0.85</strong></td>
</tr>
<tr>
<td>2 + 3</td>
<td>0.82</td>
</tr>
<tr>
<td>1 + 2 + 3</td>
<td><strong>0.84</strong></td>
</tr>
</tbody>
</table>
# Qualitative Analysis of TTA

<table>
<thead>
<tr>
<th>Title</th>
<th>Ground truth</th>
<th>Content of Title tag</th>
<th>Selected string</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correct</td>
<td>3 Weeds Hotel</td>
<td>3 Weeds Hotel</td>
<td>Unique Pub</td>
</tr>
<tr>
<td>Short</td>
<td>Irish Channel Restaurant &amp; Pub</td>
<td>Irish Channel - Restaurant &amp; Pub</td>
<td>500 H St NW DC (202) 216-0046</td>
</tr>
<tr>
<td>Long</td>
<td>Secret Garden Bed &amp; Breakfast</td>
<td>Secret Garden Bed &amp; Breakfast (formerly Whitegates Guest House), near Keynsham, Bristol: Rooms, Prices and Guest Information</td>
<td>Secret Garden Bed &amp; Breakfast (formerly Whitegates Guest House)</td>
</tr>
<tr>
<td>No title</td>
<td>Rio Pool</td>
<td>Hot Tubs, hot tub hire, swimming pools, Bristol, Gloucester</td>
<td>swimming pools</td>
</tr>
<tr>
<td>Incorrect</td>
<td>Slice and Dice</td>
<td>Home</td>
<td>Prepared Food</td>
</tr>
</tbody>
</table>

- Correct: 81%
- [CATEGORY NAME] (3 %)
- [CATEGORY NAME] (5 %)
- No title (6 %)
- Long (5 %)
## Comparative Results

<table>
<thead>
<tr>
<th>Method</th>
<th>Average similarity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Title tag (baseline)</td>
<td>0.62</td>
</tr>
<tr>
<td>TitleFinder (Mohammadzadeh et al. 2012)</td>
<td>0.52</td>
</tr>
<tr>
<td>TTA (proposed)</td>
<td>0.84</td>
</tr>
</tbody>
</table>
Comparative Results

The graph shows the number of websites with similarity scores ranging from 0.0 to 1.0 with the ground truth titles. The x-axis represents the similarity score, while the y-axis represents the number of websites. The graph compares three methods: TTA, TitleFinder, and Baseline. TTA shows the highest number of websites with high similarity scores, indicating its superior performance in title prediction compared to the other methods.
# Results with *Mopsi Services*

Annotated titles

<table>
<thead>
<tr>
<th>Method</th>
<th>Rouge-1</th>
<th></th>
<th></th>
<th>Jaccard</th>
<th>Dice</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F-score</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline (Title Tag)</td>
<td>0.71</td>
<td>0.33</td>
<td>0.41</td>
<td>0.44</td>
<td>0.54</td>
</tr>
<tr>
<td>TitleFinder (Moham.et al. 2012)</td>
<td>0.35</td>
<td>0.47</td>
<td>0.37</td>
<td>0.37</td>
<td>0.43</td>
</tr>
<tr>
<td>Styling (Changuel et al. 2009)</td>
<td>0.14</td>
<td>0.21</td>
<td>0.15</td>
<td>0.22</td>
<td>0.28</td>
</tr>
<tr>
<td>TTA (Gali and Fränti 2016)</td>
<td>0.52</td>
<td><strong>0.59</strong></td>
<td><strong>0.52</strong></td>
<td><strong>0.54</strong></td>
<td><strong>0.62</strong></td>
</tr>
</tbody>
</table>
WEBIST 2016 will be held in conjunction with CSEDU 2016, SMARTGREENS 2016, CLOSER 2016, VEHITS 2016, IoTBD 2016, and COMPLEXIS 2016.

Registration to WEBIST allows free access to the CSEDU, SMARTGREENS, CLOSER, VEHITS, IoTBD and COMPLEXIS conferences (as a non-speaker).

NEW REGISTRATIONS ARE NOW ONLY AVAILABLE AT THE CONFERENCE WELCOME DESK

Please visit the WEBIST 2017 website

The purpose of the 12th International Conference on Web Information Systems and Technologies (WEBIST) is to bring together researchers, engineers and practitioners interested in the technological advances and business applications of web-based information systems. The conference has five main tracks, covering different aspects of Web Information Systems, including Internet Technology, Web Interfaces and Applications, Society, e-Communities, e-Business, Web Intelligence and Mobile Information Systems.
The purpose of the 12th International Conference on Web Information Systems and Technologies (WEBIST) is to bring together researchers, engineers and practitioners interested in the technological advances and business applications of web-based information systems. The conference has five main tracks, covering different aspects of Web Information Systems, including Internet Technology, Web Interfaces and Applications, Society, e-Communities, e-Business, Web Intelligence and Mobile Information Systems.
What about logo images?

• ~89% of web pages have their title within a logo image
• Needs to detect logo image
• Apply OCR
• Challenging !!!
Conclusions

• TTA improves baseline
  
  62% → 84%

• Title tag works ok, but it needs to be processed
  
  <title>This still ok</title>

• Words in the page link have the highest impact
  
  http://www.webist.org/