Extracting Representative Image from Web page

Najlaa Gali, Andrei Tabarcea and Pasi Fränti
Motivation: summarize search result

Calculating distance
Structure of location-based search
First we traveled, tasted all kinds of flavors and enjoyed the world. Then we brought our experiences together in the kitchen and opened a restaurant. Boulevard Social is our third restaurant and it opened its doors in May 2012. The kitchen and bar on the beautiful boulevard offers an exciting and innovative combination of traditional flavors from Greece, Lebanon, Tunisia, Morocco and other Mediterranean countries. The simple bistro décor used in the restaurant has been designed to enhance the restaurant’s unique flavors we have created with love.

Come try our flavors or pop in for a drink!

Tomi Björck and Matti Wikberg
Overall extraction process

Web page

Web page link

Extract images

Analyze

Categorize

Rank

Representative image

Images found
What to extract

Three sources: html, CSS, JS

<link rel="stylesheet" type="text/css" media="screen" href="http://www.ompelimot.com/css/rakenne.css" title="Oletus" />

Representative image

rankenne.css

#ylaosa {
  height: 150px;
  background: url("../images/2.png") no-repeat scroll 0px 0px #EEE6C8;
  border-bottom: 2px solid #FFF;
  width: 694px;
  margin: 0px auto;
}
### Image features used

<table>
<thead>
<tr>
<th>src</th>
<th><a href="http://www.ravintolakreeta.fi///images/banner.jpg">http://www.ravintolakreeta.fi///images/banner.jpg</a></th>
</tr>
</thead>
<tbody>
<tr>
<td>alt</td>
<td>--</td>
</tr>
<tr>
<td>title</td>
<td>--</td>
</tr>
<tr>
<td>from</td>
<td>css</td>
</tr>
<tr>
<td>format</td>
<td>jpg</td>
</tr>
<tr>
<td>width</td>
<td>945</td>
</tr>
<tr>
<td>height</td>
<td>202</td>
</tr>
<tr>
<td>size</td>
<td>190,890 px</td>
</tr>
<tr>
<td>aspect ratio</td>
<td>4.67</td>
</tr>
<tr>
<td>parent tag</td>
<td>&lt;div&gt;</td>
</tr>
<tr>
<td>class</td>
<td>header</td>
</tr>
</tbody>
</table>
Image features used

alt: Ravintola Martina Joensuu
title: --
from: html
format: jpg
width: 920
height: 313
size: 287.96 px
aspect ratio: 2.94
parent tag: <div>
class: header_fiilis
class of parent: content clearfix
Image categories

- Logo
- Representative image
- Banner
- Advertisement

Formatting
Category 1: Representative images
Images that are directly related to the content
Category 2: Logos
Images of logo of the company or institution

Criteria: Image link, class or id attribute of the <img> or its parent element contains text logo

<br>

<img id="logo" alt="" src="http://www.martina.fi/sites/martina.fi/themes/martinaomega/logo.png" />

<br>

<div class="logo">
    <img src="http://www.pizzaspecial.fi/web_ulkoasut/ypj4_joen_pizza/images/footer.jpg">
</div>

<br>

</div>

</div>
Category 3: Banners
Wide or tall images usually used as logo of the service

Criteria:
• **link, class or id** contains: *banner, header, footer, button*
• High aspect ratio (> 1.8)
• Not classified as advertisement, formatting or logo
Category 4: Advertisement
Images that advertise products from other websites

Criterion:
• **Link, class** or **id** contains text: *free, now, buy, join, adserver, click, affiliate, adv, hits, counter*
• [Considered adding well known adv. server but not used]
Category 5: Formatting and icons
Images used as backgrounds, decorators or icons

Criteria:
• Link, class or id contains text: background, bg, sprite, template
• Height or width is smaller than 100 px
## Summary of rules

<table>
<thead>
<tr>
<th>Category</th>
<th>Features</th>
<th>Keywords</th>
</tr>
</thead>
<tbody>
<tr>
<td>Representative</td>
<td>Not in other category</td>
<td></td>
</tr>
<tr>
<td>Logo</td>
<td></td>
<td>logo</td>
</tr>
<tr>
<td>Banner</td>
<td>Ratio &gt; 1.8</td>
<td>Banner, header, Footer, button</td>
</tr>
<tr>
<td>Advertisement</td>
<td></td>
<td>Free, adserver, now, buy, join, click, affiliate, adv, hits, counter</td>
</tr>
<tr>
<td>Formatting and Icons</td>
<td>Width &lt; 100 px</td>
<td>Background, bg, spirit, templates</td>
</tr>
<tr>
<td></td>
<td>Height &lt; 100 px</td>
<td></td>
</tr>
</tbody>
</table>
Decision tree for categorization

- **Image**
  - **Logo?**
    - Yes: **Logo category**
    - No: **Adv.?**
      - Yes: **Advertisement category**
      - No: **Format?**
        - Yes: **Formatting category**
        - No: **Banner?**
          - Yes: **Banner category**
          - No: **Representative category**
Scoring images

http://ptiszai.com/imageext/

<table>
<thead>
<tr>
<th>Rule</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Image size $\geq 10,000$ px</td>
<td>1</td>
</tr>
<tr>
<td>Aspect ratio $\leq 1.8$</td>
<td>1</td>
</tr>
<tr>
<td>Image <strong>alt</strong> or <strong>title</strong> set a value</td>
<td>1</td>
</tr>
<tr>
<td>Keywords of <strong>alt</strong> or <strong>title</strong> appear also in <strong>&lt;title&gt;</strong> tag</td>
<td>1</td>
</tr>
<tr>
<td>Keywords of <strong>alt</strong> or <strong>title</strong> appear also in <strong>&lt;h1&gt;</strong> tag</td>
<td>1</td>
</tr>
<tr>
<td>Keywords of <strong>image path</strong> also in <strong>&lt;title&gt;</strong> or <strong>&lt;h1&gt;</strong> tags</td>
<td>1</td>
</tr>
<tr>
<td>The image is in the sub-tree of <strong>&lt;h1&gt;</strong> or <strong>&lt;h2&gt;</strong> tags</td>
<td>1</td>
</tr>
<tr>
<td>Format = <strong>jpg</strong></td>
<td>1</td>
</tr>
<tr>
<td>Format = <strong>svg</strong>, <strong>png</strong> or <strong>gif</strong></td>
<td>0.5</td>
</tr>
</tbody>
</table>
Mopsi Weblma dataset

http://cs.uef.fi/mopsi/img/

Summary of data collected:

• Websites: 1002
• Images: 2363
• Per page: Min=1, Average=2.36, Max=154

Collection details:

• Who: 117 volunteers
• When: September 2014
• What: Pages of own choice or Mopsi search
• How: Select 1-3 most representative images
• Issues: Some level of subjectivity unavoidable
## Overall results

<table>
<thead>
<tr>
<th>Platform</th>
<th>Accuracy</th>
<th>Extracted Images</th>
</tr>
</thead>
<tbody>
<tr>
<td>WebIma</td>
<td>64%</td>
<td>99%</td>
</tr>
<tr>
<td>Google+</td>
<td>48%</td>
<td>92%</td>
</tr>
<tr>
<td>Facebook</td>
<td>39%</td>
<td>90%</td>
</tr>
</tbody>
</table>
## GOOD cases
Subset for which WebIma gives 100% accuracy

<table>
<thead>
<tr>
<th></th>
<th>Set 1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ground truth (%)</td>
</tr>
<tr>
<td>Representative</td>
<td>63</td>
</tr>
<tr>
<td>Logo</td>
<td>37</td>
</tr>
<tr>
<td>Banner</td>
<td>0</td>
</tr>
<tr>
<td>Advertisement</td>
<td>0</td>
</tr>
<tr>
<td>Formatting</td>
<td>0</td>
</tr>
</tbody>
</table>
# BAD cases

Subset for which WebIma gives 0% accuracy

<table>
<thead>
<tr>
<th></th>
<th>Ground truth (%)</th>
<th>WebIma (%)</th>
<th>Google+ (%)</th>
<th>Facebook (%)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Representative</strong></td>
<td>33</td>
<td>83</td>
<td>27</td>
<td>40</td>
</tr>
<tr>
<td><strong>Logo</strong></td>
<td>30</td>
<td>7</td>
<td>30</td>
<td>7</td>
</tr>
<tr>
<td><strong>Banner</strong></td>
<td>37</td>
<td>3</td>
<td>27</td>
<td>33</td>
</tr>
<tr>
<td><strong>Advertisement</strong></td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td><strong>Formatting</strong></td>
<td>0</td>
<td>7</td>
<td>13</td>
<td>17</td>
</tr>
</tbody>
</table>
Good enough?

Subjective Ground truth

WebIma

Facebook

Google+

ERCIS
Conclusions

• **Lightweight** method suitable for real time applications

• **Unsupervised**: No training, no user feedback needed

• Finds correct image **64%** of the cases. Outperforms Google+ (48%) and Facebook (39%)

• In use in MOPSI: *Search* and *Service upgrade*
Thank you!