Content Extraction from Webpages Using Machine Learning
Master's Thesis

Hamza Yunis
Bauhaus Universität
26.01.2017

Supervised by:
Prof. Benno Stein
Dr. Andreas Jakoby

Advised by:
Johannes Kiesel
Motivation
FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

Updated 16:17 GMT (00:17 HKT) January 10, 2017

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

READ: FIFA's 48-team expansion plan explained
FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

○ Updated 1617 GMT (0017 HKT) January 10, 2017

World Cup: How would FIFA’s 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer’s world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

READ: FIFA’s 48-team expansion plan explained
FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

Updated 1617 GMT (0017 HKT) January 10, 2017

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan

New format set to start in 2026

Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48. The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

READ: FIFA's 48-team expansion plan explained
FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN
Updated 1617 GMT (0917 HKT) January 10, 2017

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

READ: FIFA's 48-team expansion plan explained
FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

FIFA, soccer’s world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

READ: FIFA’s 48-team expansion plan explained
FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

Updated 16:17 GMT (00:17 HKT) January 10, 2017

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

FIFA Council approves expansion plan
New format set to start in 2026
Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

READ: FIFA's 48-team expansion plan explained
FIFA gives go-ahead for expanded World Cup

By James Masters and Chris Murphy, CNN

Updated 1617 GMT (0017 HKT) January 10, 2017

World Cup: How would FIFA's 48-team plan work? 01:43

Story highlights

- FIFA Council approves expansion plan
- New format set to start in 2026
- Tournament to have 16 more teams

(CNN) — FIFA, soccer's world governing body, has approved a grand plan to revolutionize the World Cup by increasing the number of teams from 32 to 48.

The FIFA Council agreed unanimously to the move Tuesday, with the new format starting in 2026.

READ: FIFA's 48-team expansion plan explained
What is the Main Content?
What is the Main Content?

**Definition (i):** The main content is what the webpage is *supposed to communicate* according to the publisher.

Different users may have different interests in the webpage.
Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
What is the Main Content?

**Definition (i):** The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

**Definition (ii):** The main content is what makes the webpage interesting to the user.

**Definition (iii):** The main content of a webpage consists of information that cannot be found in other webpages. Usually used in template recognition.
What is the Main Content?

**Definition (i)**: The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

**Definition (ii)**: The main content is what makes the webpage *interesting in to the user*. 
**Definition (i):** The main content is what the webpage is *supposed to communicate* according to the publisher.

- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

**Definition (ii):** The main content is what makes the webpage *interesting in to the user*.

- Different users may have different interests in the webpage.
What is the Main Content?

Definition (i): The main content is what the webpage is *supposed to communicate* according to the publisher.
- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

Definition (ii): The main content is what makes the webpage *interesting in to the user*.
- Different users may have different interests in the webpage.

Definition (iii): The main content of a webpage consists of information that *cannot be found in other webpages*.
What is the Main Content?

**Definition (i)**: The main content is what the webpage is *supposed to communicate* according to the publisher.
- We cannot always tell what the webpage publisher wants to communicate.
- A single webpage may have different publishers, each wanting to communicate a different type of information.

**Definition (ii)**: The main content is what makes the webpage *interesting in to the user.*
- Different users may have different interests in the webpage.

**Definition (iii)**: The main content of a webpage consists of information that *cannot be found in other webpages.*
- Usually used in template recognition.
What is the Main Content?

FC Bayern München » Franck Ribéry

Franck Ribéry

Date of Birth: Apr 7, 1983 (Age 33)
Place of Birth: Boulogne-sur-Mer
Nationality: France
Height: 170 cm.
Weight: 72 Kg.
Position: Midfielder
Squad Number: 7
What is the Main Content?

FC Bayern München » Philipp Lahm

**Philipp Lahm**

- Date of Birth: Nov 11, 1983 (Age 33)
- Place of Birth: München
- Nationality: Germany
- Height: 170 cm.
- Weight: 62 Kq.
- Position: Defender
- Squad Number: 21

SQUAD

- Manuel Neuer (Goalkeeper)
What is the Main Content?

The image shows a webpage from GOAL, a sports news website, focusing on a player profile of Philipp Lahm from FC Bayern München. The profile includes details such as his date of birth (Nov 11, 1983, Age 33), place of birth (München), nationality (Germany), height (170 cm), weight (62 Kg), position (Defender), and squad number (21). The page also features links to different sections like Club Page, National Team Page, Player Profile, Player News, and Transfer Zone.
What is the Main Content?

The main content is the non-noisy content!
What is the Noisy Content?
What is the Noisy Content?

- Advertisements.
What is the Noisy Content?

- Advertisements.
- Navigation links.
What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.
What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.
- Legal information.
What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.
- Legal information.
- Irrelevant information.
What is the Noisy Content?

- Advertisements.
- Navigation links.
- Links to promoted webpages.
- Legal information.
- Irrelevant information.
- Input elements.
Types of HTML Elements
Types of HTML Elements

- Content elements.
- Inline semantic elements.
- Sectioning elements.

```html
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>

<div>
  <p>This is the <span class="important">first</span> paragraph.</p>
  <p>This is the <span class="important">second</span> paragraph.</p>
</div>
```
Types of HTML Elements

- **Content elements.**
- **Inline semantic elements.**
- **Sectioning elements.**

```
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>

<div>
  <p>This is the <span class="important">first</span> paragraph.</p>
  <p>This is the <span class="important">second</span> paragraph.</p>
</div>
```
Types of HTML Elements

- Content elements.
- **Inline semantic elements.**
- Sectioning elements.

```html
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>

<div>
  <p>This is the <span class="important">first</span> paragraph.</p>
  <p>This is the <span class="important">second</span> paragraph.</p>
</div>
```
Types of HTML Elements

- Content elements.
- Inline semantic elements.
- **Sectioning elements.**

```html
<ul>
  <li>List item 1.</li>
  <li>List item 2.</li>
</ul>

<div>
  <p>This is the <span class="important">first</span> paragraph.</p>
  <p>This is the <span class="important">second</span> paragraph.</p>
</div>
```
Types of HTML Elements

Elements to Be Classified

Paragraph elements:

Table cell elements:

List item elements:

Header elements:

Image elements:
Types of HTML Elements

Elements to Be Classified

- Paragraph elements: `<p>`.
Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
Types of HTML Elements

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
  - If they do not have content element descendants.
Types of HTML Elements

Elements to Be Classified

- **Paragraph elements**: `<p>`.
- `<div>` elements.
  - If they do not have content element descendants.
- **Table cell elements**: `<th>` and `<td>`.
Types of HTML Elements

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
  - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
  - If they do not have content element descendants.
Types of HTML Elements

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
  - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
  - If they do not have content element descendants.
- List item elements: `<li>`. 
Types of HTML Elements

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
  - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
  - If they do not have content element descendants.
- List item elements: `<li>`.
- Header elements: `<h1>`, `<h2>`, `<h3>`, `<h4>`, `<h5>`, and `<h6>`.
Types of HTML Elements

Elements to Be Classified

- Paragraph elements: `<p>`.
- `<div>` elements.
  - If they do not have content element descendants.
- Table cell elements: `<th>` and `<td>`.
  - If they do not have content element descendants.
- List item elements: `<li>`.
- Header elements: `<h1>`, `<h2>`, `<h3>`, `<h4>`, `<h5>`, and `<h6>`.
- Image elements: `<img>`.
Learning Workflow
Learning Workflow

HTML Documents
Learning Workflow

HTML Documents → Raw Features Extraction → CSV Documents
Learning Workflow

HTML Documents → Raw Features Extraction → CSV Documents → Concatenation → CSV Document
Learning Workflow

HTML Documents → Raw Features Extraction → CSV Documents → Concatenation → CSV Document

Feature Vectors
Learning Workflow

HTML Documents → Raw Features Extraction → CSV Documents → Concatenation → CSV Document

Feature Vectors → Derived Features Extraction → Enhanced Feature Vectors
Learning Workflow

HTML Documents → Raw Features Extraction → CSV Documents → Concatenation → CSV Document

Feature Vectors → Derived Features Extraction → Enhanced Feature Vectors → Learning → Classifier
Feature Engineering
Feature Engineering

```html
<div id="comments-section">
  <div class="comment">
    <div class="comment-header">
      <ul>
        <li>Author name.</li>
        <li>Comment title.</li>
      </ul>
    </div>
    <div class="comment-content">
      <p>The body of the comment</p>
    </div>
  </div>
  ...
</div>
```
Feature Engineering

<div id="comments-section">
  <div class="comment">
    <div class="comment-header">
      <ul>
        <li>Author name.</li>
        <li>Comment title.</li>
      </ul>
    </div>
    <div class="comment-content">
      <p>The body of the comment</p>
    </div>
  </div>
  ...
</div>
Feature Engineering

Raw features:

- `ancestor_names="div, div, div, ul"`
- `ancestor_classes="NO CLASSES, comment, comment-header, NO CLASSES"`
- `inner_text="Author name."`
Feature Engineering

<html>
  <div id="comments-section">
    <div class="comment">
      <div class="comment-header">
        <ul>
          <li>Author name.</li>
          <li>Comment title.</li>
        </ul>
      </div>
    </div>
    ...
  </div>
</html>

**Raw features:**
- ancestor_names="div, div, div, ul"
- ancestor_classes="NO_CLASSES, comment, comment-header, NO_CLASSES"
- inner_text="Author name."

**Derived features:**
- is_desc_comment="1"
- is_desc_cookies="0"
- is_desc_section="1"
- inner_text_length=2
Evaluation
Evaluation Workflow

HTML Documents
Evaluation Workflow

HTML Documents → Raw Features Extraction → CSV Documents → Concatenation → CSV Document

Feature Vectors → Derived Features Extraction → Enhanced Feature Vectors
Evaluation

Evaluation Workflow

- HTML Documents
  - Raw Features Extraction
  - CSV Documents
  - Concatenation
    - CSV Document

- Feature Vectors
  - Derived Features Extraction
  - Enhanced Feature Vectors
    - Evaluation Results

- Classifier
Confusion matrix:

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th>&quot;Noisy&quot;</th>
<th>&quot;Main&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Noisy&quot;</td>
<td></td>
<td>tn</td>
<td>fp</td>
</tr>
<tr>
<td>&quot;Main&quot;</td>
<td></td>
<td>fn</td>
<td>tp</td>
</tr>
</tbody>
</table>

Evaluation metrics:

Precision: \[ \text{precision} = \frac{tp}{tp + fp} \]

Recall: \[ \text{recall} = \frac{tp}{tp + fn} \]

\[ F_\beta = (1 + \beta^2) \cdot \frac{\text{precision} \cdot \text{recall}}{(\beta^2 \cdot \text{precision}) + \text{recall}} \]
Element-based results for textual content elements:

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th>&quot;Noisy&quot;</th>
<th>&quot;Main&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Noisy&quot;</td>
<td></td>
<td>4625</td>
<td>211</td>
</tr>
<tr>
<td>&quot;Main&quot;</td>
<td></td>
<td>277</td>
<td>1018</td>
</tr>
</tbody>
</table>

Precision = 0.828
Recall = 0.786
$F_1$ = 0.806
## Evaluation

Text-based results for textual content elements:

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th>&quot;Noisy&quot;</th>
<th>&quot;Main&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Noisy&quot;</td>
<td></td>
<td>496921</td>
<td>19618</td>
</tr>
<tr>
<td>&quot;Main&quot;</td>
<td></td>
<td>28654</td>
<td>163908</td>
</tr>
</tbody>
</table>

precision = 0.893  
recall = 0.851  
$F_1$ = 0.871
Element-based results for small and medium-size images ($\leq 40000\text{px}$):

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th>&quot;Noisy&quot;</th>
<th>&quot;Main&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Noisy&quot;</td>
<td>900</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>&quot;Main&quot;</td>
<td>97</td>
<td>25</td>
<td></td>
</tr>
</tbody>
</table>

- precision $= 0.833$
- recall $= 0.205$
- $F_1 = 0.328$
Element-based results for large images (> 40000px):

<table>
<thead>
<tr>
<th>Actual Class</th>
<th>Predicted Class</th>
<th>&quot;Noisy&quot;</th>
<th>&quot;Main&quot;</th>
</tr>
</thead>
<tbody>
<tr>
<td>&quot;Noisy&quot;</td>
<td>122</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>&quot;Main&quot;</td>
<td>10</td>
<td>29</td>
<td></td>
</tr>
</tbody>
</table>

precision = 0.828  
recall = 0.743  
$F_1$ = 0.783
Thank you for your attention!