Researcher Homepage Identification and Name Extraction
Application of Machine Learning with Multiple Views

Marc Ingold

Albert-Ludwigs University of Freiburg
Introduction

- **Topic:** The implementation and assessment of a machine learning approach for the information extraction from web pages.
- **Motivation:** Automated means of gaining insights from the web; an enormous collection of semi- and unstructured data.
- **Method:** Supervised Machine Learning - Binary Classification
The Problem

Typical Researcher Homepage
The Problem

Homepage Identification as Supervised Machine Learning Task

Supervised Learning

For a collection of data points \( \{(x_i, y_i)\}_{i=1}^N \), learn a function \( h : x \rightarrow y \), which predicts the label \( y_{N+1} \) for a new datapoint \( x_{N+1} \).

- Number of datapoints \( N \), which were collected in the past
- \( x_i \in \mathbb{R}^D \)
- \( y_i \in \{true, false\} \)
- \( h(x_{N+1}) = P(y_{N+1}|x_{N+1}) \)

Boedecker et al. (2017)
Questions?
The Main Tasks

1. Obtain suitable web page data
2. Identify researcher homepages
   - Develop two prediction models using disjoint feature sets
   - Bag of words approach
3. Extract the researchers name from the page
   - Extract all person names from the homepage
   - Identify the correct person name
   - Augmenting heuristic with machine learning features
The Approach

Common Crawl

- Non profit organization that crawls the web on a monthly basis
- Crawl data is stored in Amazon Web Services as part of their Public Datasets Program
- Approx. 300 index files per crawl. (∼1.5 TiB uncompressed)
- Crawl of August 2019: 260 TiB (uncompressed), 2.95 billion web pages

https://commoncrawl.org/
URL Based Features - URL Surface Patterns

<table>
<thead>
<tr>
<th></th>
<th>URL</th>
<th>Surface Patterns</th>
</tr>
</thead>
<tbody>
<tr>
<td>2</td>
<td><a href="http://abi.inf.uni-tuebingen.de/People/krueger">http://abi.inf.uni-tuebingen.de/People/krueger</a></td>
<td>people, nondict</td>
</tr>
<tr>
<td>3</td>
<td><a href="http://people.ucas.ac.cn/~zhangxiaopeng?language=en">http://people.ucas.ac.cn/~zhangxiaopeng?language=en</a></td>
<td>tildenondict, querykeylanguage, queryvaluenondict</td>
</tr>
</tbody>
</table>

Surface Patterns:
- numeric, alphanumeric, hyphenated, underscored, long term
- nondict: No proper English word or not in the term dictionary
- tildenondict: Researcher name prefixed by ~
- querykey, queryvalue: prefix to URL query terms

---

\(^1\)Gollapalli et al. (2015)
The Approach

Homepage Identification

URL Based Features

- Natural language specific preprocessing applied
- Uni- and bigrams
- Vectorized via Term Frequency Inverse Document Frequency (Tfidf)

<table>
<thead>
<tr>
<th>Url_Id</th>
<th>tildenondict</th>
<th>numeric</th>
<th>querykeyid</th>
<th>...</th>
<th>news</th>
<th>people</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>0.723131</td>
<td>0</td>
<td>0</td>
<td>...</td>
<td>0</td>
<td>0.160545</td>
<td>1</td>
</tr>
<tr>
<td>1</td>
<td>0</td>
<td>0.983265</td>
<td>0.324515</td>
<td>...</td>
<td>0</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Each row represents a web page and is a sparse vector of 8386 features.
Page Content Based Features

- Text from title and h1 tag prefixed with identifier
- Concatenated with the rest of the page text content
- Numeric Features:
  - Num. tables
  - Num. external links
  - Num. internal links
  - Num. images
  - Num. person names in title / h1 tag

Features after preprocessing: 20006 Tfidf vectorized uni-, bi- and tri-grams
Random Forest and linear models with Stochastic Gradient Descent learning were compared.

Best URL based model:
- Random Forest
- Default parameters except for number of trees (1000)

Best Page Content based model:
- Support Vector Machine with modified huber loss function

Combined model prediction:
\[ P_{combined}(y|x) = P_{url}(y|x) \times P_{page}(y|x) \]
Questions?
Training Data

Training data downloaded from July, August and September Crawl of 2018.

<table>
<thead>
<tr>
<th>Source</th>
<th>No. Homepages</th>
</tr>
</thead>
<tbody>
<tr>
<td>World Wide Knowledge Base(^2) - 4 Universities Dataset</td>
<td>52</td>
</tr>
<tr>
<td>Computer Science Bibliography(^3)</td>
<td>14 473</td>
</tr>
<tr>
<td>Manual Labelling (Freiburg, Munich, Stanford, Media Faculty of the MIT)</td>
<td>2130</td>
</tr>
<tr>
<td>After filtering the html data</td>
<td>13 670</td>
</tr>
</tbody>
</table>

Undersampling was applied to account for imbalanced classes

\(^2\)https://cs.cmu.edu/afs/cs.cmu.edu/project/theo-20/www/data/
\(^3\)https://dblp.org/
Evaluation
Sampling

Validation- and Test-Data

Validation Data:
- Cross Validation as part of GridSearch
- Hold out dataset

Test Data:
- 1500 web pages were manually labelled
- Universities: Caltech, Princeton, York, Stuttgart, Hamburg, Applied Science Upper Austria
- Contained 86 homepages
Metrics

**Precision:** Quality of the predictions made by the model. How good are the predictions of the model.

\[
\frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Positive}}
\]

**Recall:** Measure for the coverage of the model. How well is the model suited to predict the label.

\[
\frac{\sum \text{True Positive}}{\sum \text{True Positive} + \sum \text{False Negative}}
\]

**F1 Score:** Measure of models performance, where precision and recall contribute evenly.

\[
2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}
\]
## Classification Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Label</th>
<th>Validation Data</th>
<th></th>
<th>Test Data</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Precision</td>
<td>Recall</td>
<td>F1 Score</td>
<td>Precision</td>
</tr>
<tr>
<td>Page</td>
<td>0</td>
<td>0.97</td>
<td>0.94</td>
<td>0.96</td>
<td>1</td>
</tr>
<tr>
<td>Content</td>
<td>1</td>
<td>0.94</td>
<td>0.97</td>
<td>0.96</td>
<td>0.12</td>
</tr>
<tr>
<td>Url</td>
<td>0</td>
<td>0.91</td>
<td>0.91</td>
<td>0.91</td>
<td>0.99</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>0.92</td>
<td>0.92</td>
<td>0.92</td>
<td>0.18</td>
</tr>
<tr>
<td>Combined</td>
<td>0</td>
<td>0.95</td>
<td>1</td>
<td>0.97</td>
<td>0.98</td>
</tr>
<tr>
<td></td>
<td>1</td>
<td>1</td>
<td>0.94</td>
<td>0.97</td>
<td>0.68</td>
</tr>
</tbody>
</table>
Summary

- 68% F1 Score achieved in the homepage identification task
- 94% F1 Score achieved in the person identification task
- Simple machine learning algorithms and features well suited for the web page classification
- Great benefit from using two disjoint feature sets under suboptimal condition
- Convention of writing the researcher name in the title tag is widely held
- Person identification heuristic could be improved with machine learning features


A0: Person Identification - Sampling

- Names extracted with Stanford NE Recognizer, merged and manually labelled

- Training data:
  - Sample taken from the homepage identification training dataset
  - Extracted and labelled 36123 person names from 1705 homepages

- Test data:
  - Sample taken from the homepage identification test dataset
  - Extracted and labelled 2106 person names from 83 homepages
A1: Person Identification - The Method

<table>
<thead>
<tr>
<th>Url_Id</th>
<th>Name</th>
<th>In_Title</th>
<th>In_h1</th>
<th>In_h2</th>
<th>Count</th>
<th>Count_Third</th>
<th>Count_Half</th>
<th>No_Parts</th>
<th>Label</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>Name1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>5</td>
<td>2</td>
<td></td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>0</td>
<td>Name2</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>10</td>
<td>3</td>
<td>6</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>0</td>
<td>Name3</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>6</td>
<td>0</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1</td>
<td>Name1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>16</td>
<td>14</td>
<td>16</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>

Z-score standardization of numeric features: \( \frac{x_i - \mu}{\sigma} \)

Machine Learning Algorithm:
- Random Forest (250 trees)
## A2: Person Identification Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Validation Data</th>
<th>Test Data</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Precision</td>
<td>Recall</td>
</tr>
<tr>
<td>Heuristic</td>
<td>0.93</td>
<td>0.84</td>
</tr>
<tr>
<td>4 Features</td>
<td>0.93</td>
<td>0.92</td>
</tr>
<tr>
<td>All Features</td>
<td>0.96</td>
<td>0.91</td>
</tr>
</tbody>
</table>

![Importance Chart](chart.png)

Legend:
- Appeared in Title
- Appeared in Header1
- Count
- Count in first half Feature
- Count in first third
- Appeared in Header2
- Parts
Appendix

B: Development Overview
C: Prediction Probabilities by Model and Web Page Type

(A) Non Homepages
(B) Homepages
(C) Combined model
D: Natural Language Preprocessing

- **Tokenization**: Splitting sequences of characters into useful semantic units.
- **Lower case**
- **Stopword / Punctuation removal**
- **Stemming / Lemmatization**: Reduce terms to a common base form. (Word Stem / Lemma)
- **Term Frequency Inverse Document Frequency (Tfidf)**:

  \[ tfidf(t, d) = tf(t, d) \cdot \log \frac{N}{\sum_{D:t \in D} 1}, \]

  for term \( t \), documents \( d \in D \), number of documents \( N \).
E: URL Based Features - 20 Most Frequent Terms by Label
F: URL Based Features - Feature Importance
Appendix

G: Page Based Features - 30 Most Frequent Terms by Label
Appendix

H: Page Based Features - Feature Importance
J: URL Based Model: Common Errors / Improvements

| URL                                                                 | $P_{url}(y = 1|x)$ | $P_{page}(y = 1|x)$ | Error Type            |
|----------------------------------------------------------------------|---------------------|----------------------|-----------------------|
| (1) http://www-users.cs.york.ac.uk/~susan/sf/dani/PS_019.htm        | .73                 | .14                  | false positive        |
| tildenondict, nondict, nondict, underscoredword                      |                     |                      |                       |
| (2) https://www.ifm.uni-hamburg.de/en/datenschutz.html              | 1                   | .79                  | false negative        |
| en, nondict                                                         |                     |                      |                       |
| (3) https://www.york.ac.uk/economics/our-people/staff-profiles/john-hutton/ | .37                 | .88                  |                       |
| economics, hyphenatedword, hyphenatedword, hyphenatedword           |                     |                      |                       |
| (4) http://carvermead.caltech.edu/research.html                      | .58                 | .74                  |                       |
| research                                                             |                     |                      |                       |

Improvements:

- (1) Add features representing the beginning and end of the URL.
- (2) Handle non-english terms
- (3) Include meaning of hyphenated terms
Appendix

K: Page Content Based Model: Improvements

Index Name
Arvind, K.

Co-authors
Akish, B.; Anand, S.V.; Bharath, P.; Chakraborty, N.; Mahapatra, D. Roy

Publication Titles
2009: Coupled electro-mechanical response of an electroactive polymer cantilever structure and its application in energy harvesting

Seiteninfo: Impressum | Last Change 1. Mai 2010 by Volkmar Vill und Ron Zenczykowski
Blättern: 

- Topic Modelling
- Substantially expand stopword lists

Gollapalli et al. (2011)
Future Work

Improvements:

- Homepage Identification:
  - Training data sampling
  - Individual model feature engineering and feature selection

- Person Identification:
  - Name extraction and name merging procedures at the preprocessing for the person identification

- Overall approach:
  - Co-training\textsuperscript{5}
  - Improvements to the probability estimates produced by tree based models\textsuperscript{6}

\textsuperscript{5}Gollapalli et al. (2015)
\textsuperscript{6}Tanha et al. (2017)