Researcher Homepage Identification and Name Extraction

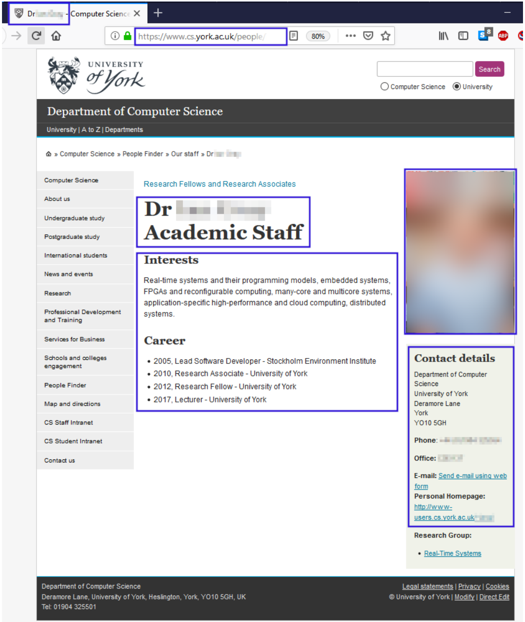
Application of Machine Learning with Multiple Views

Marc Ingold

Albert-Ludwigs University of Freiburg

Introduction

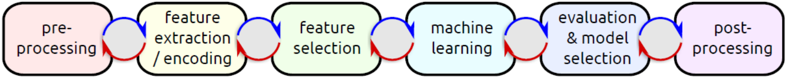
- Topic: The implementation and assessment of a machine learning approach for the information extraction from web pages.
- Motivation: Automated means of gaining insights from the web; an enormous collection of semi- and unstructured data.
- Method: Supervised Machine Learning - Binary Classification


Dr - Computer Science
https://www.cs.york.ac.uk/people/
80%
UNIVERSITY of York
Search
Computer Science
University
Department of Computer Science
University | A to Z | Departments
» Computer Science » People Finder » Our staff » Dr
Computer Science
About us
Undergraduate study
Postgraduate study
International students
News and events
Research
Professional Development and Training
Services for Business
Schools and colleges engagement
People Finder
Map and directions
CS Staff Intranet
CS Student Intranet
Contact us
Research Fellows and Research Associates
Dr
Academic Staff
Interests
Real-time systems and their programming models, embedded systems, FPGAs and reconfigurable computing, many-core and multicore systems, application-specific high-performance and cloud computing, distributed systems.
Career
• 2005, Lead Software Developer - Stockholm Environment Institute
• 2010, Research Associate - University of York
• 2012, Research Fellow - University of York
• 2017, Lecturer - University of York
Contact details
Department of Computer Science
University of York
Deramore Lane
York
YO10 5GH
Phone:
Office:
E-mail: Send e-mail using web form
Personal Homepage:
http://www-users.cs.york.ac.uk
Research Group:
• Real-Time Systems
Department of Computer Science
Deramore Lane, University of York, Heslington, York, YO10 5GH, UK
Tel: 01904 325501
Legal statements | Privacy | Cookies
© University of York | Modify | Direct Edit

Supervised Learning

For a collection of data points $\langle (x_i, y_i) \rangle_{i=1}^{N}$, 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

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. ($\sim$ 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[1]

(1)	https://www.inf.uni-hamburg.de/en/inst/ab/hci/news/rse15.html nondict, nondict, nondict, nondict, news, alphanumeric
(2)	http://abi.inf.uni-tuebingen.de/People/krueger people, nondict
(3)	http://people.ucas.ac.cn/∼zhangxiaopeng?language=en tildenondict, querykeylanguage, queryvaluenondict

Surface Patterns:

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

[1] Gollapalli et al. (2015)

URL Based Features

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

Url_Id	tildenondict	numeric	querykeyid	...	news	people	Label
0	0.723131	0	0	...	0	0.160545	1
1	0	0.983265	0.324515	...	0	0	0
...	...	...	...	...	...	...	

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

Machine Learning Models

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:

$$P_{combined}(y|x) = P_{url}(y|x) * P_{page}(y|x)$$

Questions?

Training Data

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

Source	No. Homepages
World Wide Knowledge Base[2] - 4 Universities Dataset	52
Computer Science Bibliography[3]	14 473
Manual Labelling (Freiburg, Munich, Stanford, Media Faculty of the MIT)	2130
After filtering the html data	13 670

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/

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.

$$\frac{\sum \mathit{True\ Positive}}{\sum \mathit{True\ Positive} + \sum \mathit{False\ Positive}}$$

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

$$\frac{\sum \mathit{True\ Positive}}{\sum \mathit{True\ Positive} + \sum \mathit{False\ Negative}}$$

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

$$2 \cdot \frac{\mathit{Precision} \cdot \mathit{Recall}}{\mathit{Precision} + \mathit{Recall}}$$

Classification Results

Model	Label	Validation Data Precision	Validation Data Recall	Validation Data F1 Score	Test Data Precision	Test Data Recall	Test Data F1 Score
Page Content	0	0.97	0.94	0.96	1	0.58	0.73
	1	0.94	0.97	0.96	0.12	0.97	0.22
Url	0	0.91	0.91	0.91	0.99	0.77	0.86
	1	0.92	0.92	0.92	0.18	0.84	0.29
Combined	0	0.95	1	0.97	0.98	0.98	0.98
	1	1	0.94	0.97	0.68	0.69	0.68

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

Bibliography I

Boedecker, J., Hutter, F., and Tangermann, M. (2017). Machine learning. Albert Ludwigs University Freiburg Lecture.

Criminisi, A. and Shotton, J. (2013). *Decision forests for computer vision and medical image analysis*. Springer Science & Business Media.

Friedman, J., Hastie, T., and Tibshirani, R. (2001). *The elements of statistical learning*, volume 1. Springer series in statistics New York.

Gollapalli, S. D., Caragea, C., Mitra, P., and Giles, C. L. (2015). Improving researcher homepage classification with unlabeled data. *ACM Transactions on the Web*, 9(4):1–32.

Bibliography II

Gollapalli, S. D., Giles, C. L., Mitra, P., and Caragea, C. (2011). On identifying academic homepages for digital libraries. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 123–132. ACM.

Tanha, J., van Someren, M., and Afsarmanesh, H. (2017). Semi-supervised self-training for decision tree classifiers. *International Journal of Machine Learning and Cybernetics*, 8(1):355–370.

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

Url_Id	Name	In_Title	In_h1	In_h2	Count	Count_Third	Count_Half	No_Parts	Label
0	Name1	0	0	1	5	2	2	1	0
0	Name2	1	1	0	10	3	6	2	1
0	Name3	0	0	0	1	1	1	6	0
...	...	...	...	...	...	...	...	...	...
1	Name1	1	0	0	16	14	16	2	1
...	...	...	...	...	...	...	...	...	...

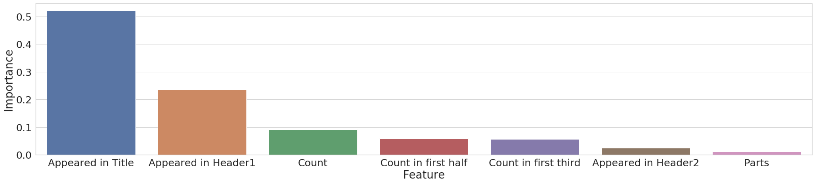
Z-score standardization of numeric features : $\frac{x_i - \mu}{\sigma}$

Machine Learning Algorithm:

- Random Forest (250 trees)

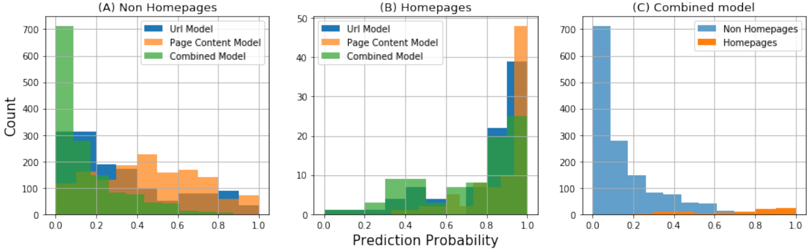
A2: Person Identification Results

	Validation Data			Test Data		
Model	Precision	Recall	F1 Score	Precision	Recall	F1 Score
Heuristic	0.93	0.84	0.88	0.95	0.92	0.93
4 Features	0.93	0.92	0.92	0.95	0.93	0.94
All Features	0.96	0.91	0.94	0.95	0.93	0.94



B: Development Overview

C: Prediction Probabilities by Model and Web Page Type



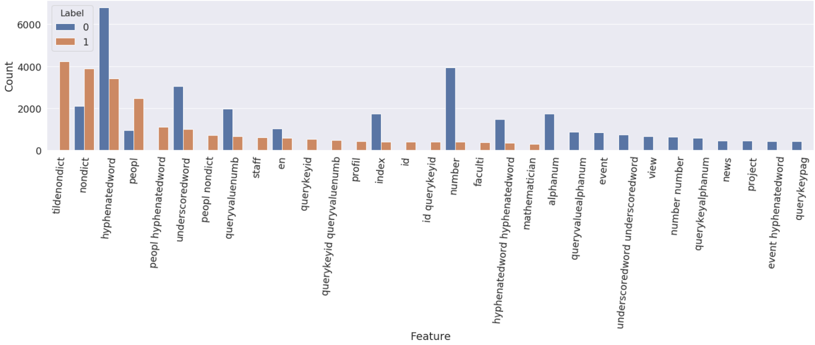
D: Natural Language Preprocessing

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

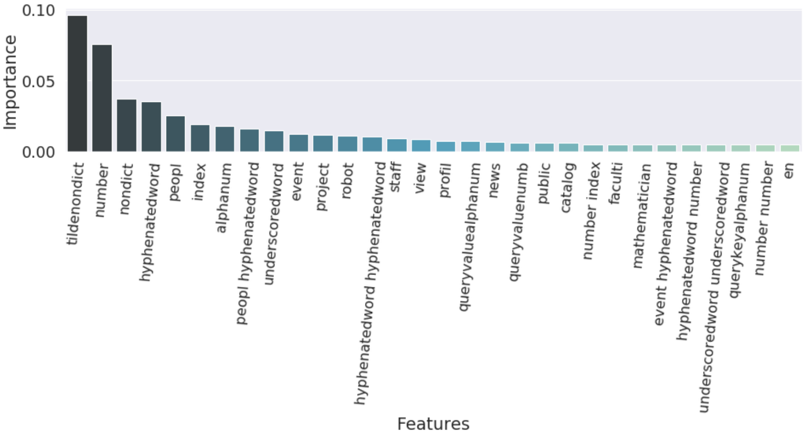
$$tfidf(t,d) = tf(t,d) \cdot \log \frac{N}{\sum_{D:t \in D} 1},$$

for term t, documents $d \in D$, number of documents N.

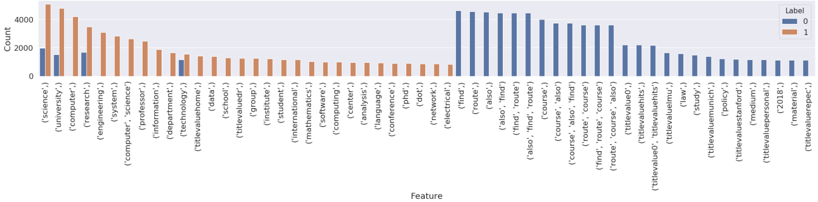
E: URL Based Features - 20 Most Frequent Terms by Label



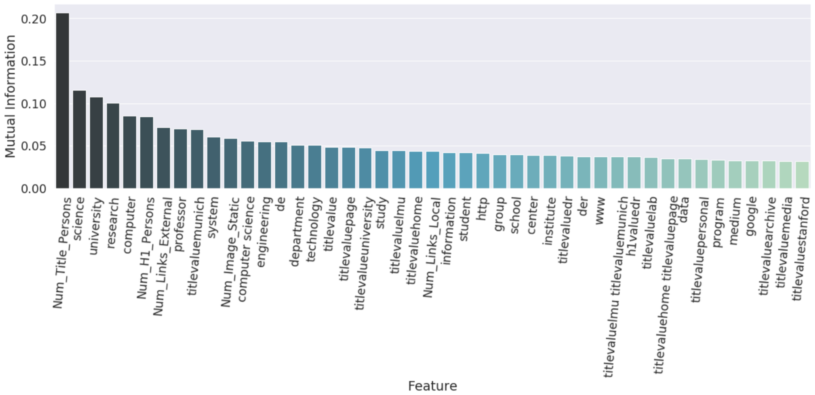
F: URL Based Features - Feature Importance



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



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 tildenondict, nondict, nondict, underscoredword	.73	.14	false positive
(2)	https://www.ifm.uni-hamburg.de/en/datenschutz.html en, nondict	1	.79	false positive
(3)	https://www.york.ac.uk/economics/our-people/staff-profiles/john-hutton/ economics, hyphenatedword, hyphenatedword, hyphenatedword	.37	.88	false negative
(4)	http://carvermead.caltech.edu/research.html research	.58	.74	false negative

Improvements:

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

K: Page Content Based Model : Improvements

Index Name

Arvind, K.

Co-authors

Ahish, 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 [4]
- Substantially expand stopword lists

[4] 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[5]
  - Improvements to the probability estimates produced by tree based models[6]

[5] Gollapalli et al. (2015)
[6] Tanha et al. (2017)