Towards a More Fine Grained Analysis of Scientific Authorship

Predicting the Number of Authors Using Stylometric Features

Andi Rexha, Stefan Klampfl, Mark Kröll, **Roman Kern**

Know-Center - Research Center for Data-Driven Business and Big Data Analytics

BIR 2016

Introduction

Setting of our Work

Intrinsic Plagiarism Detection

Detect a change in writing style within a document

- A change in the style can be seen as indicator for a change in authorship
- For single author documents this might indicate some form of "lifted" text

Authorship Attribution

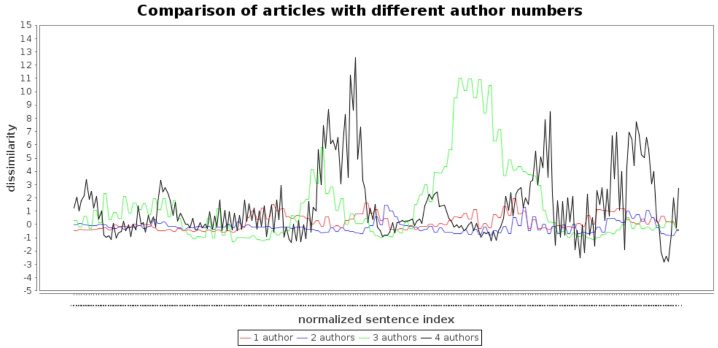
Attribute certain writing styles to different authors

- Usually documents are assumed to be authored by a single person
- The authors and reference documents are known beforehand

Past Work

In previous work we:

- We applied techniques from intrinsic plagiarism detection on multi-author papers
- More changes in writing style for papers with more authors



Motivation

Main Research Question

To which extend is it possible to predict for a given paper the correct **number of authors**?

Goal

- More fine grained analysis of the individual author contribution & roles
- Improved researchers' profiles
- Tune parameters for authorship profiling & intrinsic plagiarism detection

Limitations

There are a number of limitations/caveats with this approach:

- Writing papers is a highly dynamic, individual and complex process
  - Instead of a rigid mapping of sections to single authors
  - Native vs non-native speakers
  - E.g. even single sentences might be written by multiple authors
- Many authors may contribute little to the written part
  - E.g. mentors, engineers, co-workers, …
  - Would not show up
- Highly sensitive issue (plagiarism, varying degree of contributions, …)
- Hard task even for humans

Method

Approach & Setting

Algorithmic Approach

- Supervised machine learning
  - Requires labelled training data-set
- Classification algorithms
  - Comparison of various algorithms
- Features extracted from the papers
- Papers are directly taken as PDF documents

Pre-Processing Pipeline

PDF Extraction

- PDF is the most common file format for scientific articles
  - … but it does not contain any structural information
  - Therefore the PDF needs to be parsed and analysed
- Result of PDF Extraction
  - Main text of the article as plain text

Limitation of PDF Extraction

- PDF extraction is an complex task with many steps
  - Each step may introduce noise and errors

PDF Extraction Pipeline

1. PDF as starting point, parsing with Apache PDFBox
   - Characters -> Words -> Lines -> Blocks
2. Detection of:
   - Decoration (e.g. page numbers)
   - Captions, tables & images
   - Headings & main text
   - Reference section
3. Reverse engineer the structure
   - Sort by reading-order
   - Automatic table of contents from the headings
   - Analyse tables
4. Classification
   - Meta-data: title, journal name, authors, …
   - References & citations (within the text, reference section)
5. Post-processing
   - De-hyphenation
   - Merging & splitting of paragraphs

Features

- Focus on stylometric features
  - Instead of content related features, e.g. unigrams
- Applied on various granularity levels
  - Sentence, paragraph, document
- Aggregated for a single set of features for each article
  - Min, max, mean and variance for each feature
  - Each instance (paper) consists of 60 numeric features

Type of Features

Table 1: Overview of the stylometric features used, grouped by their type

Type of Feature	**Examples**
Character-based statistics	Ratio of upper/lower case characters
Vocabulary usage	Relation size of vocabulary to number of words
	Hapax legomena
Averages	Average length of words
	Average length of sentence

Classification Algorithms

Focus on two well-know classification algorithms

- Logistic regression
  - Commonly used for textual classification tasks
- Random forests
  - Popular in many different domains (e.g. image analysis)

Dataset

Dataset Overview

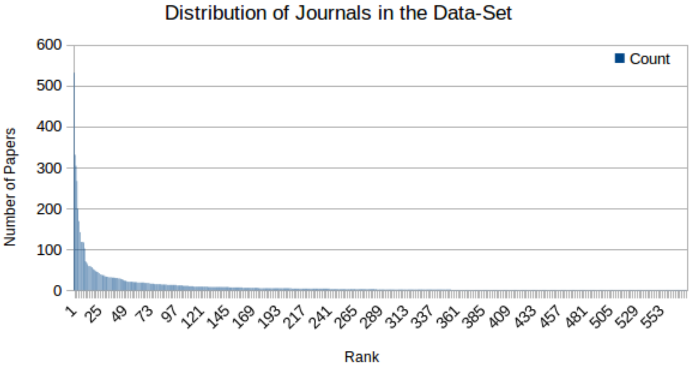
- Papers with a varying degree of authors
  - Ranging from 1 to 5 authors
- Based on PubMed
  - Mostly papers from the bio-medical domain
- Selected 6144 papers on random
  - Only labelled as *research_article*
  - Yielded 563 different journals

Top Journals

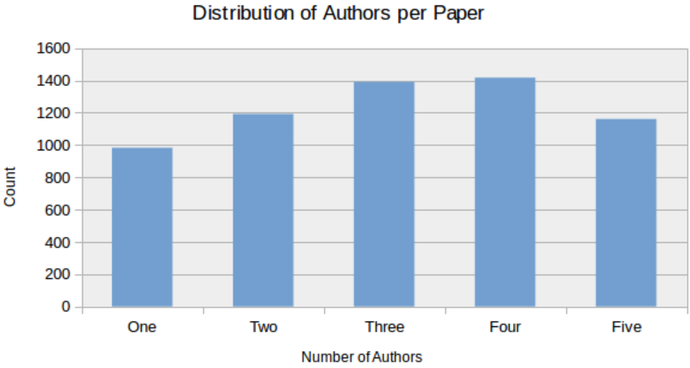
Table 2: Top 5 journals in the data-set

Name	**Count**
Environmental Health Perspectives	531
Nucleic Acids Research	331
PLoS ONE	304
The Yale Journal of Biology and Medicine	267
BMC Bioinformatics	199

Distribution of Journals



Distribution of Authors



Results

Baselines

Point of reference

- Random guessing
  - 20% chance for correct decision
- Most frequent
  - Always pick *4 authors*
  - F_1 of 0.09

Logistic Regression

Table 3: Results for logistic regression for 10-fold cross-validation

Metric	Value
Precision	0.365
Recall	0.376
F_1	**0.362**

Random Forest

Table 4: Results for the random forest algorithm for 10-fold cross-validation

Metric	Value
Precision	0.759
Recall	0.755
F_1	**0.755**

Confusion Matrix

84%	2%	5%	4%	2%
2%	76%	7%	3%	10%
6%	7%	76%	6%	6%
6%	4%	7%	78%	9%
2%	10%	5%	8%	73%

Confusion matrix for the random forest algorithm

- Best performance for single author
- Worst performance for 5 authors

Discussion

- *Surprisingly* good results
  - The features are able to discriminate between the classes
    - ★ I.e. papers are different (regardless of true number of writers)
    - ★ But no linear relationships
  - The pre-processing pipeline keeps the main features intact
- 1 author papers appear different to others
  - I.e. easier to discriminate against multi-author papers

Thank You!