# Towards a more fine grained analysis of scientific authorship: Predicting the number of authors using stylometric features

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

*{arexha, sklampfl, mkroell, rkern}@know-center.at*
Know-Center GmbH, Inffeldgasse 13, A-8010 Graz (Austria)

## Abstract

To bring bibliometrics and information retrieval closer together, we propose to add the concept of author attribution into the pre-processing of scientific publications. Presently, common bibliographic metrics often attribute the entire article to all the authors affecting author-specific retrieval processes. We envision a more fine-grained analysis of scientific authorship by attributing particular segments to authors. To realize this vision, we propose a new feature representation of scientific publications that captures the distribution of stylometric features. In a classification setting, we then seek to predict the number of authors of a scientific article. We evaluate our approach on a data set of ~ 6100 PubMed articles and achieve best results by applying random forests, i.e., 0.76 precision and 0.76 recall averaged over all classes.

## Introduction

The ongoing growth of the volume of scholarly publications poses significant challenges to both information retrieval processes in digital libraries as well as bibliometric techniques that analyse academic literature in a quantitative manner. Ideas from other fields such as computer linguistics have been incorporated into bibliometrics to improve and enhance the measuring and analysis processes.

To bring bibliometrics and information retrieval closer together, we propose to add the concept of author attribution into the pre-processing of the analysis of scientific publications. Yet, since common bibliographic metrics often attribute the entire article to all the authors, we introduce a reinterpretation of authorship attribution: to attribute particular segments of an article to individual authors allowing for a more fine-grained analysis of contribution and role. Information retrieval systems could then benefit from such authorship attribution in the following ways: Scholarly search engines could implement an author specific search which allows researchers to specifically look for text passages written by a particular author. This more precise passage-author attribution then allows the generation of researcher profiles. These profiles would reflect a researcher's contributions to different scientific fields in a more detailed manner. In addition, the profile might be valuable for predicting and thus understanding a researcher's role, for example, more actively involved (writing) vs. acting more like a mentor providing ideas and giving feedback (less involved in writing; reflected for example by author positioning).

As a first step in this direction we have recently applied text segmentation to identify potential author changes within the main text of a scientific article (Rexha et al., 2015). We have adopted a number of stylometric features to capture stylistic changes in the text, following the hypothesis that different authors manifest in different writing styles within the document. In this article we extend this work by applying a new feature representation of scientific documents that captures the distribution of stylometric features across the document and to predict the number of authors accordingly. The classification performance then represents so-to-say a quantification of the amount of information that is contained within the stylometry of a scientific article about the number of authors involved in writing it.

The text for the analysis is produced by a PDF processing pipeline, which analyses scientific articles and extracts, among other information, also the main text (Klampfl et al., 2014). As training data we have chosen a subset of PubMed research articles. This data set consists of a

wide variety of journals across different domains. We have selected an approximately equal number of research articles written by a certain number of authors, ranging from one to five.

This paper is structured as follows: First, we elaborate on existing work on authorship attribution techniques as well as the retrieval of higher level knowledge from scientific texts in general. Then, we describe our experimental setup, including the dataset and extracted stylometric features. Finally, we present our results and give an outlook for future work.

## Related Work

Over the past decades one can observe an ever growing amount of scientific output; much to the joy of research areas such as (i) Bibliometrics which applies statistics to measure scientific impact and (ii) Information Retrieval which applies natural language processing to make the valuable body of knowledge accessible. This interest in processing and exploiting scientific publications from different perspectives is reflected by venues such as the International Workshop on Bibliometric-enhanced Information Retrieval (cf. (Mayr et al.)), the International Workshop on Mining Scientific Publications[1] or Mining Scientific Papers: Computational Linguistics and Bibliometrics[2].

To be of value for both fields, scientific publications need to be semantically enriched. Adding semantics includes assigning instances to concepts which are organized and structured in dedicated ontologies. Entity and relation recognition thus represent a vital pre-processing step. To give an example, medical entity recognition (cf. (Abacha & Zweigenbaum, 2011)) seeks to extract instances from classes such as "Disease", "Symptom" or "Drug" to enrich the retrieval process. Research assistants such as BioRAT (cf. (Corney et al., 2004)) or FACTA (cf. (Tsuruoka et al., 2008)) then can offer an added value employing this type of semantic information.

Departing from a mere content-level, Liakata et al. (cf. (Liakata et al., 2012)) introduced a different approach by focusing on the discourse structure to characterize the knowledge conveyed within the text. For this purpose, the authors identified 11 core scientific concepts including "Motivation", "Result" or "Conclusion". Ravenscroft et al. (cf. (Ravenscroft et al., 2013)) present the Partridge system which automatically categorizes articles according to their types such as "Review" or "Case Study". In a similar manner, the TeamBeam (cf. (Kern et al., 2012)) algorithm extracts structured meta-data, such as the title, journal name and abstract, as well as information about the article's authors.

In this paper we introduce the concept of authorship attribution as an additional pre-processing step for subsequent retrieval procedures. Authorship attribution, in general, expresses a classification setting where from a set of candidate authors the author of a questioned article is to be selected (cf. Stamatatos (2009), Juola (2008)). This line of research can be traced back to the 19th century, when Mendenhall (1887) aimed to characterize the plays of Shakespeare. A century later (Mosteller & Wallace, 1964) used a Bayesian approach to analyse 'The Federalist Papers'. Since then, a line of research known as stylometry focused on defining features to quantify an author's writing style (Holmes, 1998). Bergsma, Post & Yarowsky (2012) used stylometric features to detect the gender of an author and to distinguish between native vs. non-native speakers and conference vs. workshop papers. In this paper, we use stylometric features to classify scientific papers according to the number of its authors.

---

[1] Conference: Proceedings of the 4th Workshop on Mining Scientific Publications. Co-located with the Joint Conference on Digital Libraries (JCDL), Knoxville, Tennessee, 2015.

[2] Conference: Proceedings of the First Workshop on Mining Scientific Papers: Computational Linguistics and Bibliometrics co-located with 15th International Society of Scientometrics and Informetrics Conference (ISSI), Istanbul, Turkey, 2015.

## Experimental Setup

### Dataset

For the evaluation we use a dataset composed of randomly selected documents from PubMed (http://www.ncbi.nlm.nih.gov/pubmed/), a free database created by the US National Library of Medicine holding full-text articles from the biomedical domain together with a standard XML mark-up that rigorously annotates the complete content of the published document, in particular the author metadata. The documents contained in this database are very diverse. In this work we limit ourselves to research articles only, but there is also a wide range of different article types, including book reviews and meeting reports.

For this evaluation we selected a subset of the PubMed dataset consisting of an approximately equal number of research articles written by a certain number of authors, ranging from one to five. For our evaluation, we chose 6144 research articles in total, across 563 different journals and publication entities. There were 983, 1192, 1391, 1418, and 1160 articles with one, two, three, four, and five authors, respectively.

### PDF Extraction

A prerequisite for the writing style analysis of scientific articles is the reliable extraction of their textual content. The portable document format (PDF), the most common format for scientific literature today, is optimised for presentation, but lacks structural information. As the raw character stream of the PDF is usually interrupted in mid-sentence by decorations or floating objects, extracting the main text of a scholarly article in the correct order requires the analysis of its document structure. To solve this task we build here upon our previous work (Kern et al. 2012, Klampfl et al., 2014), where we have developed a processing pipeline that analyses the structure a PDF document using a number of both supervised and unsupervised machine learning techniques and heuristics. It processes a given PDF file in a sequence of individual processing modules and outputs the extracted body text.

The first step builds upon the output of the Apache PDFBox library (http://pdfbox.apache.org) and uses unsupervised learning (clustering) to extract blocks of contiguous text from the raw PDF file and their column-wise reading order on each page. We consider these text blocks as the basic building blocks of a scientific article. In the next stage, these text blocks are categorized into different logical labels based on their role within the document: meta-data blocks, decorations, figure and table captions, main text, and section headings. This stage is implemented as a sequential pipeline of detectors each of which labels a specific type of block. Apart from the meta-data detectors they are completely model-free and unsupervised. For more details on each of these detectors the interested reader is referred to (Klampfl et al., 2014). In the final stage of our PDF extraction pipeline the main body text of a scientific article is extracted by concatenating blocks containing section headings and main text in the reading order. We resolve hyphenations at the end of lines and across blocks, columns, and pages. Furthermore, paragraphs that span more than one column or page are merged.

### Stylometric features and document representation

Capturing different writing styles within a document requires the extraction and analysis of suitable features. Topical features, such as word unigrams or other elements carrying semantic information, are helpful in identifying document segments which differ not only in the author, but also in the whole topic of the text. On the other hand, stylometric features reflect the author's writing style, rather than the topic, which typically does not change within a single scientific article, and generalizes across different domains.

To compare and classify different scientific articles based on the number of authors involved, we try to capture the distribution of stylometric features across a single document. We split

the document into continuous segments (here a segment corresponds to a sentence) and extract the stylometric features for each of those segments. We then view the document as a distribution of different stylometric features.

The literature suggests a broad amount of stylometric features (Mosteller & Wallace, 1964; Tweedie & Baayen, 2002; Stamatatos, 2009). Table 1 presents the list of features we extract for each segment. In addition, we calculate the minimum, maximum, average and variance for each of those features across every document.

**Table 1: List of stylometric features used in our text segmentation algorithm. Many of those features are defined in (Tweedie & Baayen, 2002).**

| feature name | Description |
|---|---|
| alpha-chars-ratio | the fraction of total characters in the paragraph which are letters |
| digit-chars-ratio | the fraction of total characters in the paragraph which are digits |
| upper-chars-ratio | the fraction of total characters in the paragraph which are upper-case |
| white-chars-ratio | the fraction of total characters in the paragraph which are whitespace characters |
| type-token-ratio | ratio between the size of the vocabulary (i.e., the number of *different* words) and the total number of words |
| hapax-legomena | the number of words occurring once |
| hapax-dislegomena | the number of words occurring twice |
| yules-k | a vocabulary richness measure defined by Yule |
| simpsons-d | a vocabulary richness measure defined by Simpson |
| brunets-w | a vocabulary richness measure defined by Brunet |
| sichels-s | a vocabulary richness measure defined by Sichel |
| honores-h | a vocabulary richness measure defined by Honore |
| average-word-length | average length of words in characters |
| average-sentence-char-length | average length of sentences in characters |
| average-sentence-word-length | average length of sentences in words |

## Evaluation

In order to evaluate whether the stylometric feature representation of scientific articles contains authorship information, we trained different classifiers in a supervised manner to predict the number of authors for each document. From the articles in the PubMed dataset, we extracted the stylometric features for each sentence of the document and represented the distribution of this features across the document as its maximum, minimum, average and variance. As a further preprocessing step we normalized these feature values to avoid dominating features in the learning process. For our experiments, we selected two classification algorithms: Logistic regression and Random Forest.

Table 2 and Table 3 report individual class results achieved by Logistic Regression and Random Forest algorithms. Comparing the two classification algorithms, we notice that the Random Forest outperforms the Logistic Regression algorithm by far.

As can be seen, both algorithms achieve the lowest performance in predicting the 5-authors class. We believe that this outcome might be due to two different aspects. The first aspect has to do with the amount of contribution from each author. The smaller the amount of text an author writes, the more difficult to distinguish it from the contribution of other authors. The second aspect relates to the actual writing contributions. We think that the larger the amount of writers the more likely is that some of them may not have contributed at all in the writing of the paper.

Another consideration that we can make relates to the 1-author class. The performance of both algorithms exceeds the results for the other classes. We believe that this is due to the

correctness of the data. The 1-author papers are less likely to have more authors than mentioned in the paper, making the data more representative for this class.

These experiments demonstrate that in the proposed stylometric feature space it is possible to a certain extent to discriminate between scientific articles with different numbers of authors.

**Table 2. Performance of classifying the number of authors of a scientific article using logistic regression on our dataset (10-fold cross-validation).**

| Class/Metric | Precision | Recall | F-Measure |
|---|---|---|---|
| Class 1-author | 0.533 | 0.482 | 0.506 |
| Class 2-authors | 0.330 | 0.301 | 0.315 |
| Class 3-authors | 0.369 | 0.522 | 0.432 |
| Class 4-authors | 0.235 | 0.432 | 0.394 |
| Class 5-authors | 0.235 | 0.105 | 0.145 |
| Average | 0.365 | 0.376 | 0.362 |

**Table 3. Performance of classifying the number of authors of a scientific article using random forests on our dataset (10-fold cross-validation).**

| Class/Metric | Precision | Recall | F-Measure |
|---|---|---|---|
| Class 1-author | 0.881 | 0.780 | 0.827 |
| Class 2-authors | 0.755 | 0.681 | 0.716 |
| Class 3-authors | 0.759 | 0.801 | 0.780 |
| Class 4-authors | 0.724 | 0.796 | 0.759 |
| Class 5-authors | 0.687 | 0.699 | 0.693 |
| Average | 0.759 | 0.755 | 0.755 |

## Conclusion

In this paper, we classified scientific articles according to their number of authors by using a set of stylometric features. We applied supervised learning to this setup and achieved best results with Random Forests. The classification results suggest that the stylometric feature space in fact captures variations in the writing style that we would expect from multiple contributing authors.

This work fosters our understanding towards a more fine-grained analysis of scientific authorship by attributing particular segments to authors. Information retrieval systems could benefit from this concept of authorship attribution, for instance, in course of author specific search.

## Acknowledgments

## References

Abacha, A. & Zweigenbaum, P. (2011). Medical entity recognition: a comparison of semantic and statistical methods. BioNLP 2011 Workshop. Association for Computational Linguistics.

Bergsma, S., Post, M., & Yarowsky, D. (2012). Stylometric analysis of scientific articles. Proceedings of the Conference of the North American Chapter of the ACL: Human Language Technologies.

Choi, F.Y. (2000). Advances in domain independent linear text segmentation. Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference. pp. 26-33.

Corney, D., Buxton, B., Langdon, W. & Jones, D. (2004). BioRAT: extracting biological information from full-length papers. Bioinformatics, 20.

Dias, G. & Alves, E. (2005). Unsupervised topic segmentation based on word co-occurrence and multi-word units for text summarization. Proceedings of the ELECTRA Workshop associated to 28th ACM SIGIR Conference, Salvador, Brazil. pp. 41-48.

Harpalani, M., Hart, M., Singh, S., Johnson, R. & Choi, Y. (2011). Language of vandalism: Improving wikipedia vandalism detection via stylometric analysis. Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies. pp. 83-88.

Hearst, M.A. (1997). TextTiling: Segmenting text into multi-paragraph subtopic passages. Computational linguistics 23(1), 33-64.

Holmes, D. (1998). The Evolution of Stylometry in Humanities Scholarship. Literary and Linguistic Computing, 13(3):111–117.

Juola, P. (2008). Authorship attribution. Foundations and Trends R in Information Retrieval, 1.

Kern, R., Jack, K., Hristakeva, M., & Granitzer, M. (2012). TeamBeam Meta-Data Extraction from Scientific Literature. D-Lib Magazine, 18(7), 1.

Klampfl, S., Granitzer, M., Jack, K. & Kern, R. (2014). Unsupervised document structure analysis of digital scientific articles. International Journal on Digital Libraries 14(3-4), 83-99.

Liakata, M., Saha, S., Dobnik, S., Batchelor, C. & Rebholz-Schuhmann, D. (2012). Automatic recognition of conceptualization zones in scientific articles and two life science applications. Bioinformatics 28 (7).

Mayr, P., Scharnhorst, A., Larsen, B., Schaer, P., & Mutschke, P. (2014). Bibliometric-enhanced information retrieval. In Advances in Information Retrieval (pp. 798-801). Springer International Publishing.

Mendenhall, T. (1887). The characteristic curves of composition. Science, ns-9(214S):237–246.

Mosteller, F. & Wallace, D. (1964). Inference and disputed authorship: The federalist. Addison-Wesley.

Peng, F. & McCallum, A. (2004). Accurate information extraction from research papers using conditional random fields. Proceedings of Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics, pages 329–336.

Ravenscroft, J., Liakata, M. & Clare, A. (2013). Partridge: An Effective System for the Automatic Classification of the Types of Academic Papers. AI-2013: The Thirty-third SGAI International Conference.

Rexha, A., Klampfl, S., Kröll, M. & Kern. R. (2015). Towards Authorship Attribution for Bibliometrics using Stylometric Features. In: Proc. of the Workshop Mining Scientific Papers: Computational Linguistics and Bibliometrics, 15th International Society of Scientometrics and Informetrics Conference (ISSI), Istanbul, Turkey, pp. 44-49. http://ceur-ws.org, 2015.

Stamatatos, E. (2009). A survey of modern authorship attribution methods. Journal of the American Society for Information Science and Technology, 60(3):538–556.

Tsuruoka, Y., Tsujii, J. & Ananiadou, S. (2008). FACTA: a text search engine for finding associated biomedical concepts. Bioinformatics 24(21).

Tweedie, F. & Baayen, H. (1998). How variable may a constant be? Measures of lexical richness in perspective. Computers and the Humanities. pp. 323-352.