Towards Authorship Attribution for Bibliometrics 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
The overwhelming majority of scientific publications are authored by multiple persons; yet, bibliographic metrics are only assigned to individual articles as single entities. In this paper, we aim at a more fine-grained analysis of scientific authorship. We therefore adapt a text segmentation algorithm to identify potential author changes within the main text of a scientific article, which we obtain by using existing PDF extraction techniques. To capture stylistic changes in the text, we adopt a number of stylometric features. We evaluate our approach on a small subset of PubMed articles consisting of an approximately equal number of research articles written by a varying number of authors. Our results indicate that the more authors an article has the more potential author changes are identified. These results can be considered as an initial step towards a more detailed analysis of scientific authorship, thereby extending the repertoire of bibliometrics.

Conference Topic
Methods and techniques

Introduction
Bibliometrics has had to face the ever growing amount of scientific output in recent years -- a challenge as well as a great opportunity. Techniques from other fields such as computer linguistics have been taken over (i) to speed up measuring processes as well as to (ii) to introduce novel ideas. In this paper we propose authorship attribution as additional method for bibliometrics. So far, authorship of a scientific article has been attributed to the given authors in a more or less unchallenged way. The extent of authorship is in general defined by community standards, for instance, it is in many scientific domains assumed that the lead author did most of the (writing) work and the last author contributed ideas being the head of the group. Applying authorship attribution methods enables us to attribute particular segments of an article to individual authors thereby analysing scientific authorship on a more fine-grained level. We would like to get more insights into writing style habits of scientists, for instance: Is there a preferred partitioning amongst authors? Is there a relation to the author ordering? In addition, these methods may also have the potential to measure whether the distribution of credit within a community or a research group is just.
As a first step into this direction, we seek to identify author changes within text passages. We thus apply *TextSeqFault* (Kern & Granitzer, 2009), an algorithm for intrinsic plagiarism detection - a line of research exhibiting a closely related problem setting. The algorithm was originally developed to detect changes in topics in order to apply text segmentation. To be applicable for authorship attribution, we adapted the algorithm to catch writing style changes by taking into account stylometric features. To evaluate our approach, we created a small subset of PubMed research articles. This data set consists of an approximately equal number of research articles for certain number of authors, ranging from one to four. In our experiments we could show that there is a tendency of a correlation between the number of authors and the stylomentric differences within the text.

Background

Coined by Alan Pritchard in 1969, bibliometrics in general seeks to measure science by providing methods to explore, for example, the impact of a particular publication. Citation analysis (cf. Garfield (1972)) represents one common method being an expression for simply counting a scientific article's citations which can be regarded as indicator for an article's scientific impact.

To face the ever growing amount of written publications, there was an increased interest in automating these methods by including ideas and techniques from other domains such as computer linguistics and network analysis. To that end, linguistic resources such as the ACL Anthology Reference Corpus (Bird et al., 2008) were compiled for standardization as well as comparison purposes with respect to research problems including reference analysis (cf. (Peng & McCallum, 2004)), citation classification (cf. (Teufel, Siddharthan & Tidhar, 2006)) and generation of summaries (cf. (Elkiss et al., 2008)).

In this paper we introduce authorship attribution as an additional method for bibliometrics. Authorship attribution (cf. Stamatatos (2009), Juola (2008)) expresses a classification setting where from a set of candidate authors, the author of a questioned article is to be selected. 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) including (i) lexical features such as average word/sentence length and vocabulary richness, (ii) syntactical features such as frequency of function words and use of punctuation and (iii) structural features such as indentation. (Bergsma, Post & Yarowsky, 2012) used stylometric features to detect the gender, native speaker vs. non-native speaker and conference vs. workshop paper.

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 focus on 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 small 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 four. For our preliminary evaluation, we chose 10 research articles for each number of authors the *BMC Bioinformatics* journal – in total 40 articles.

PDF Extraction

A prerequisite for the analysis of the writing style 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 (Klampfl et al., 2014), where we have developed an unsupervised 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. Furthermore we resolve hyphenations at the end of lines and across blocks, columns, and pages.

Text Segmentation

Our intrinsic plagiarism detection algorithm is based on a sliding window approach, originally developed for text segmentation. Text segmentation is applied in order to reconstruct individual document borders of a single, long document that was constructed by concatenating multiple textual documents, e.g., transcripts of spoken text. The majority of techniques for text segmentation are designed to detect changes in topics (Choi, 2000; Dias & Alves, 2005). Our text segmentation algorithm (Kern & Granitzer, 2009), named *TextSeqFault*, is a derivative of the well-known *TextTiling* algorithm, proposed by Hearst (1997), and also falls into this category.
For each position within the document, preceding and succeeding consecutive sentences are combined into two adjacent sliding windows, which are then compared in a vector space. A dissimilarity measure calculates the relative difference between their inner similarity (the average pairwise similarity of sentences within the two windows) and their outer similarity (the average pairwise similarity of sentences across the two windows). This dissimilarity value is positive if the outer similarity is lower than the inner similarity, which indicates a potential topic change. The maximum value of 1 is reached if the outer similarity is zero, which is the case if the blocks correspond to orthogonal vectors. A topic change is reported when the dissimilarity exceeds a predefined threshold. As a similarity measure between two sentences we chose the common cosine similarity because of its simplicity and efficiency.

Stylometric Features

In the original *TextSeqFault* algorithm (Kern & Granitzer, 2009) the features used to detect a change in topic are directly derived from the words within the sentences, i.e., by building a vector space of unigrams. We adapted the algorithm for the domain of intrinsic plagiarism detection by using a different set of features. Instead of topical features, such as word unigrams or other elements carrying semantic information, we made use of stylometric features, as we expected that topical features will be limited to work in cases where not only the authorship, but also the whole topic of the text dramatically changes. These stylometric features were chosen to reflect the style of the author, rather than the topic, which typically does not change within a single scientific article. In literature a wide array of stylometric features have been proposed (Mosteller & Wallace, 1964; Tweedie & Baayen, 2002; Stamatatos, 2009). Stylometric features have also been put to use in a number of use cases, e.g. for author profiling (Koppel, Argamon & Shimoni, 2002) and vandalism detection (Harpalani et al., 2011). Table 1 shows the stylometric features used in our algorithm.

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	avarage length of sentences in words

Evaluation

In order to produce a preliminary evaluation we decided to have a visual landscape of the dissimilarity within documents. For each of the analysed documents we calculate a stylometric dissimilarity among two adjacent sliding windows containing thirty sentences each. To show the results of this step in a larger scale, we multiply them with a scaling factor of 10.000. Furthermore we have normalized the length of the documents, where each position in the chart represent the dissimilarity of the relative position in the document.

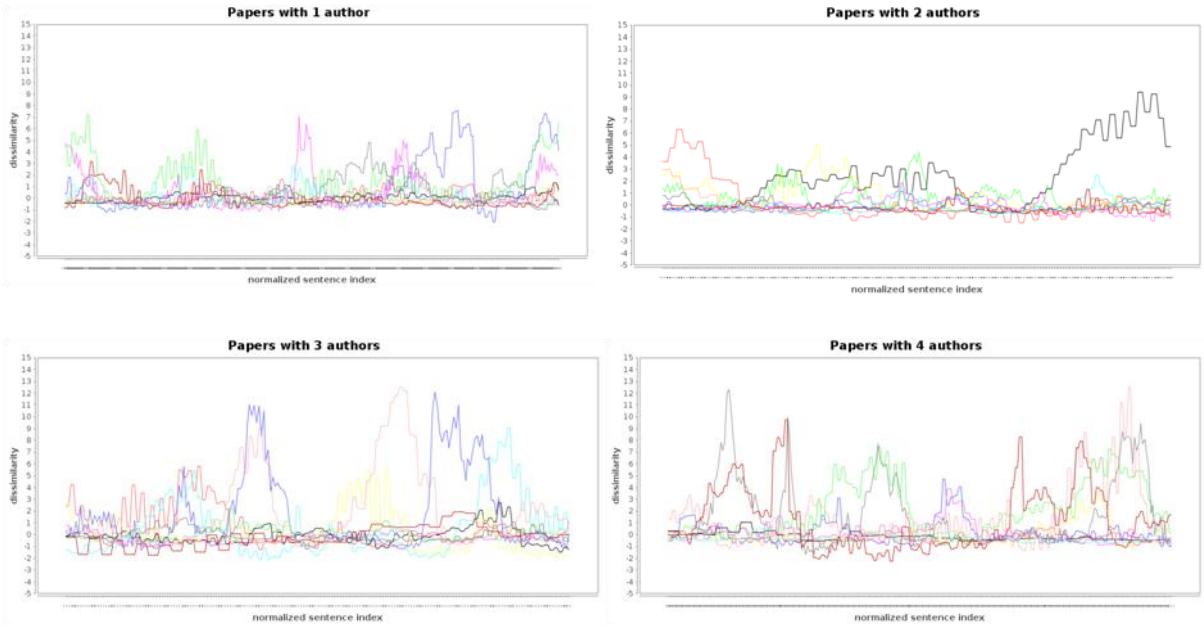

Figure 1: Landscape of the writing style dissimilarity for papers with different number of authors

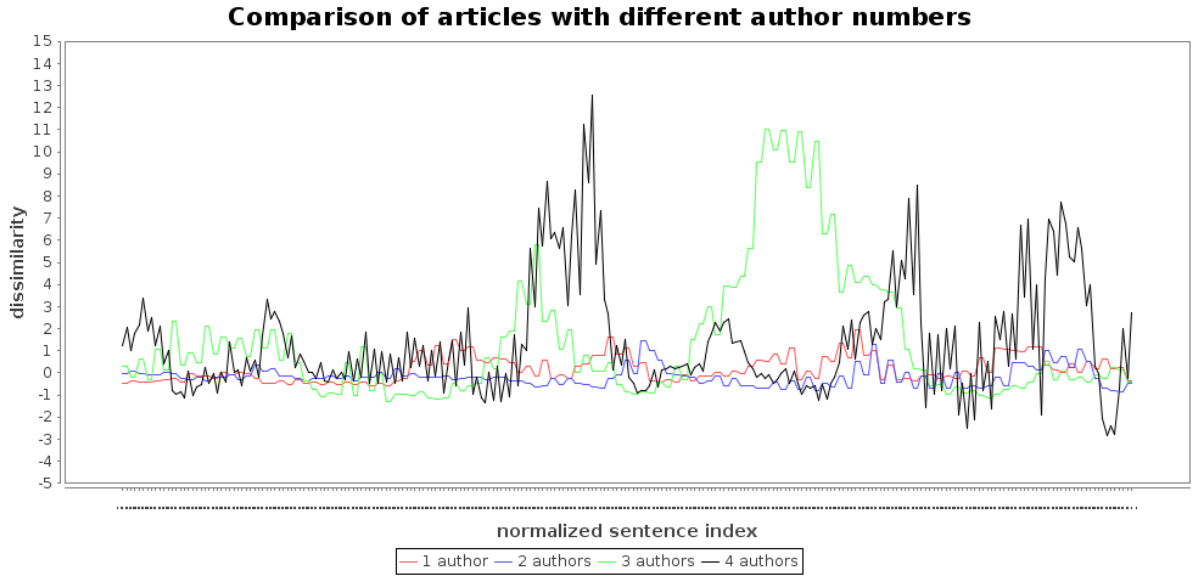

Figure 2: Comparison of writing style dissimilarity among papers with different number of authors

Below we show two types of charts that aim to illustrate the style change among papers within the same category(with same number of authors) as well as a comparison among articles with different numbers of authors which aims to show a correlation between the number of authors and the dissimilarity of the writing style.
As illustrated in the Figure 1, there is a tendency of higher changes of writing style with the growing number of authors. The number of high peaks (which represent a big change of the writing style) grows with the growing of the amount of the authors for the paper.
The inspection of the Figure 2 highlights the differences between papers written by different amount of authors. The papers with one and two authors tend to have a flat shape showing a small dissimilarity within the document. On the other hand the papers with three and four authors are inclined to have bigger and larger variations of writing style. In a closer look, also the document with four authors shows the tendency of higher number of large dissimilarity compared to the three authors paper.

Conclusion

In this paper, we proposed to add authorship attribution methods to the repertoire of bibliometrics thereby enabling a more fine-grained analysis of authorship. As a first step into this direction we presented an algorithm to segment scientific articles according to writing style changes. Our preliminary results corroborate the natural assumption that in most cases the more authors contribute the more author changes are identified. In future work, we will extend our evaluation to more articles across topics as well as across journals. In addition, we intend to learn classification models for individual authors capturing the respective writing style trying to associate each part to the individual author. This feature might be used to credit differently the contribution of each author to the paper.

Acknowledgments

This work is funded by the KIRAS program of the Austrian Research Promotion Agency (FFG) (project number 840824). The Know-Center is funded within the Austrian COMET Program under the auspices of the Austrian Ministry of Transport, Innovation and Technology, the Austrian Ministry of Economics and Labour and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

References

Bergsma, S., Post, M., & Yarowsky, D. (2012). Stylometric analysis of scientic articles. *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Bird, S., Dale, R., Dorr, B., Gibson, B., Joseph, M., Kan, M., Lee,D. Powley,B., Radev, D. & Fan Tan, Y. (2008). The ACL anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. *Proceedings of the 6th International Conference on Language Resources and Evaluation Conference (LREC08)*, pages 1755–1759.

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.

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.

Elkiss, A., Shen, S., Fader, A., Erkan, G., States, D. & Radev, D. (2008). Blind men and elephants: what do citation summaries tell us about a research article*? Journal of the American Society for Information Science and Technology*, 59(1):51–62.

Garfield, E. (1972). Citation analysis as a tool in journal evaluation. *Science* (178).

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

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.

Kern, R. & Granitzer, M. (2009). Efficient linear text segmentation based on information retrieval techniques. In: *Proceedings of the International Conference on Management of Emergent Digital EcoSystems*. p. 25.

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.

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.

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

Teufel, S. Siddharthan, A. & Tidhar, D. (2006). Automatic classification of citation function. In *Proceedings of the International Conference on Empirical Methods in Natural Language Processing*, pages 103–110.

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