Extending Scientific Literature Search by Including the Author's Writing Style

Andi Rexha, Mark Kröll, Hermann Ziak, and Roman Kern

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

Abstract

Our work is motivated by the idea to extend the retrieval of related scientific literature to cases, where the relatedness also incorporates the writing style of individual scientific authors. Therefore we conducted a pilot study to answer the question whether humans can identity authorship once the topological clues have been removed. As first result, we found out that this task is challenging, even for humans. We also found some agreement between the annotators. To gain a better understanding how humans tackle such a problem, we conducted an exploratory data analysis. Here, we compared the decisions against a number of topological and stylometric features. The outcome of our work should help to improve automatic authorship identification algorithms and to shape potential follow-up studies.

Introduction

The retrieval of related literature represents a day-to-day activity for research personnel of any kind, be it a PhD student hunting down the latest publications in context of her work or a R&D professional compiling a state-of-the-art for a prospective area of interest. In previous work, we introduced the idea of extending the retrieval process by including authorship attribution. We suggested implementing an author specific search which allows researchers to specifically look for text passages written by a particular author. In (Rexha et al., 2015), we applied text segmentation to identify potential author changes within the main text of a scientific article. Rexha et al., 2016 presented 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.

In this follow-up work, we seek to incorporate writing style information into the retrieval process. Writing style would then represent an additional, novel search dimension which is orthogonal to standard search dimensions such as content (based on terms and keywords) and relevance (based on bibliometric features like citation and impact). The author's writing style contains valuable information; besides reflecting an author's personality, the writing style also directly relates to the readability of textual content. As an example for this, we can think about authors preferring short over long sentences, or those favoring different vocabulary or phrases than others.

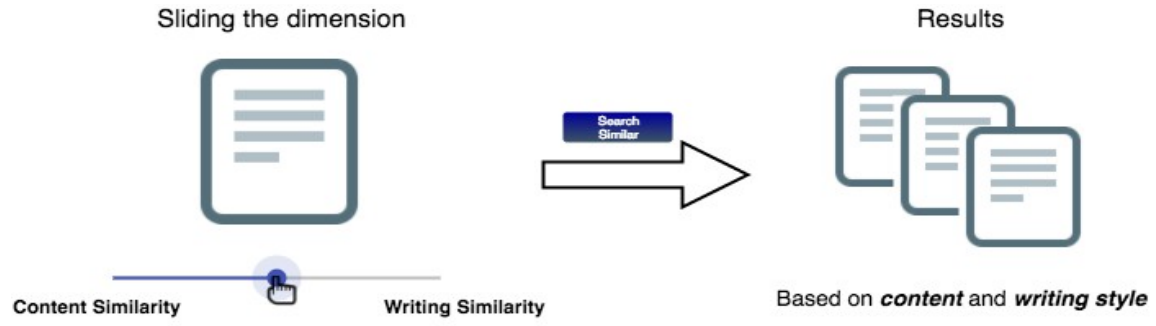


Figure 1: The user can give weight to her preferred search dimension using a slider. In this case, the user can fine tune with respect to similarity based on content and similarity based on writing style.

A PhD student, for instance, who intends to improve her writing could benefit from the aforementioned search functionality. She could search for similarly written papers with the one she likes, and learn writing patterns out of them. Hence, we envision new retrieval systems providing a slider to give more weight to the preferred dimension of search. Figure 1 shows a possible search scenario, where the user may choose how much emphasis should be put on the topological information(which we consider the semantic of the text) and the writing style.

As a first step towards building such a system, we conducted a pilot study in order to understand whether humans are capable of distinguishing between writing styles without having topological information. Though we found some agreement between the annotators, our findings reveal that this task is a challenging one, even for humans. To learn more about this problem and its difficulties, we conducted an exploratory data analysis where we statistically compared the decisions against a number of topological and stylometric features. We believe our findings to be valuable – not only for integrating writing style information into the retrieval process – but also to improve the automatic attribution of authorship. In addition, we make our dataset publicly available[1].

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.

Both fields benefit from adding semantics to scientific publications. This includes assigning instances to concepts which are organized and structured in dedicated ontologies. Entity and relation recognition thus represent a valuable pre-processing step for subsequent search procedures. Medical entity recognition (cf. Abacha & Zweigenbaum, 2011) seeks to extract instances from classes such as “Disease”, “Symptom” or “Drug” to enrich the retrieval process. In bioinformatics, Zweigenbaum et al., 2007 identify biological entities, for example, instances of “Protein”, “DNA” or “Cell Line”, and extract the relations between these entities as facts or events. Research assistants such as BioRAT or FACTA then can offer an added value employing this type of semantic information. BioRAT (cf. (Corney et al., 2004)) is given a query, and autonomously, finds a set of papers, applies natural language processing to identify biomedical entities, and highlights the most relevant facts. FACTA (cf. (Tsuruoka et al., 2008)) searches Medline abstracts with an emphasis on biomedical concepts.

Liakata et al., (2012) departed from a mere content-level enrichment and focused on the discourse structure to characterize the knowledge conveyed within the text. For this purpose, they identified 11 core scientific concepts including “Motivation”, “Result” or “Conclusion”. In the Partridge system, Ravenscroft et al., (2013) build upon the automated recognition to automatically categorize articles according to their types such as Review or Case Study. The TeamBeam (cf. (Kern et al., 2012)) algorithm aims to extract an article’s meta-data, such as the title, journal name and abstract, as well as explicit information about the article's authors. Implicit information about an author includes her writing style which reflects among others the writer’s personality as well as directly relates to characteristics such as readability, clarity, aso. Stylometry represents the line of research which focuses on defining features to quantify an author's writing style (Holmes, 1998). Bergsma, Post & Yarowsky (2012) used stylometric

[1] https://doi.org/10.5281/zenodo.437461

features to detect the gender of an author and to distinguish between native vs. non-native speakers and conference vs. workshop papers. Stylometry is, for example, employed to attribute authorship, i.e. from a set of candidate authors the author of a questioned article is to be selected (cf. Stamatatos (2009), Juola (2008)).

With respect to bibliometrics, citation information have been recently explored to enrich the retrieval process. Dabrovska & Larson, (2015) extracted citation contexts from citing articles and used them in the scientific search process. Preliminary results indicated that including citation contexts had a small but positive impact. Eck, N.J., & Waltman, L. (2014) introduced CitNetExplorer, a software tool for analysing and visualizing citation networks, and that can thus be used for citation-based scientific literature retrieval.

Experimental Setup

In order to understand whether humans are able to identify the authorship once the topological information has been removed, we conduct a pilot study. In this study, we provide human annotators with one source and four target textual snippets in different experiments. In the first, one of the targets is written by the same author as the source and the other three are written by different authors as the source. Then, we let the annotators rank the snippets from the most to the least similar with respect to the writing style, asking them to rank as "most similar" the snippet written by the same author (see Figure 2).

For the study, we selected data from Pubmed[2], a free database created by the US National Library of Medicine. This database holds full-text articles from the biomedical domain together with a standard XML markup that rigorously annotates the complete content of the published document. It also contains valuable metadata like the authors and the journal in which the article is published. At first, we retrieve documents written by only a single author to obtain "pure" writing styles. Note that it can happen that some articles can be written by ghostwriters or by colleagues of authors helping them with English writing. Yet, we believe that this is a very rare case.

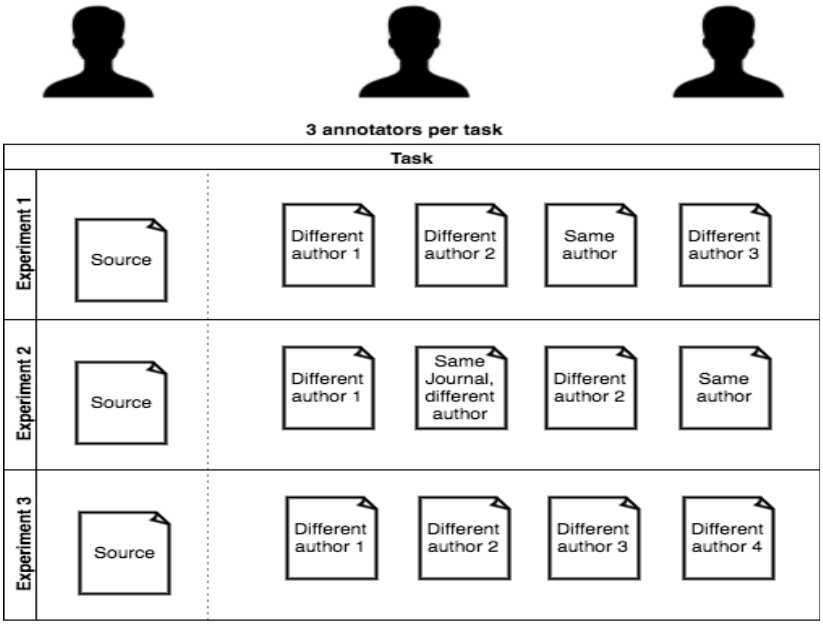


Figure 2: Three annotators are assigned to each task (each consisting of three experiments) and rank the given snippets of the respective experiment from "most similar" to "least similar".

[2] https://www.ncbi.nlm.nih.gov/pubmed (last accessed Feb. 12th, 2017)

From the previously selected documents, we choose a subset and decide to make the annotators rank text snippets which are drawn from the beginning of the introduction section (we select the first sentences until the one ending after the 400-th character). The rationale behind this choice is twofold: i) it gets more and more difficult for the user to remain focused on the task while reading a long text; ii) we hypothesize that the introduction contains less topological information than other parts of the scientific papers.
Having selected the text snippets, we designed three experiments (which we call a "task") for each annotator. For each of the experiments we present the annotators a source and four target snippets, subject to different problem settings (see Figure 2):

- In Experiment 1, we present a target snippet written by the same author as the source as well as three others written by different authors as the source.
- In Experiment 2, we provide the annotators with one target article written by the same author as the source, one target article from a different author but published in the same journal as the source, and two targets written by different authors and published in different journals as the source. This experiment is designed to capture any correlation from the writing style within the same journal, presumably within the same scientific topic.
- In Experiment 3, we want to gain as much information as possible from the user's thinking while ranking based on the similarity. Thus, we show four target snippets written by different authors as the one of the source snippet, while still suggesting to the annotator that one of the targets is written by the same author as the source.

To conduct an exploratory data analysis, we presented the same set of experiments ("task") to three different annotators (see Figure 2). In the last design step of our pilot study, we selected 90 random snippets from the PubMed database as candidate source snippets. We indexed the database of the snippets from single authors by stemming the words and removing the stop-words. We assigned 30 snippets to each of the categories of the experiments, and we performed for each of them a search according to the request:

- In Experiment 1, we searched for 10 similar articles from the same author and 100 from different authors.
- In Experiment 2, we searched for 10 similar articles from the same author, 10 from the same journal but different author, and 100 from different authors and journals
- In Experiment 3, we searched for 100 similar articles from different authors.

Based on these results, we perform a cosine similarity between the vector of the words with the source snippet and select the most similar ones accordingly to the experiment description. This way, the topological information should be removed as a source of information for authorship identification. For example, for Experiment 1 we selected the most similar article from the same author and 3 from different authors. Additionally we also apply a manual check and remove experiments that we assume to contain topological hints for texts written by the same author (mainly based on keywords or phrases). At the end of this phase, we chose 66 experiments (22 per each category of experiments previously described).

The pilot study was performed using the crowed-sourcing platform CrowdFlower[3]. The platform provides workforce from different countries helping to label and to enrich data. In the next section, we present the outcome of this study as well as analysis of the result.

[3] https://www.crowdflower.com/ (last accessed Feb. 12th, 2017)

Results

The job in CrowdFlower was performed by 56 different annotators from 29 different countries. Being our goal to rank based on the writing style, the level of understanding English isn't of a big concern for the task. To avoid random selection, we have configured the system to disallow annotation in less than 20 seconds.

At first glance, the annotators have a small agreement in the ranking of the similarity between source and target snippets. Without considering the rank itself, it was achieved a full agreement in 26 targets, 160 have an agreement of two annotators, and 78 of the targets have no agreement at all. For a more detailed analysis, we use Krippendorff's alpha measure to determine the inter-rater agreement for the ranking of each target. This was computed using the library "*DK-Pro statistics*"[4]. The results show:

- An Inter-rater Agreement of 0.299
- An Observed Disagreement of 0.699
- An Expected Disagreement of 0.999

We continue to explore the annotator's rank by considering the snippets written by the same author and those written within the same journal (but by different authors). Table 1 shows the amount of times users selected the articles in each category of similarity.

Snippet/Ranking	Most Similar	Similar	Less Similar	Least Similar
Same Author	25	60	31	16
Same Journal	14	27	17	8

Table 1: The annotators' ranking for snippets from the same author and snippets from the same journal but different author.

As we can notice the agreement is low and the annotators mainly fail to recognize the same author and the same journal. To further investigate the annotators' ranking behaviour, we make a visual analysis between the ranking of the users, the content similarity, and the stylometric similarity.

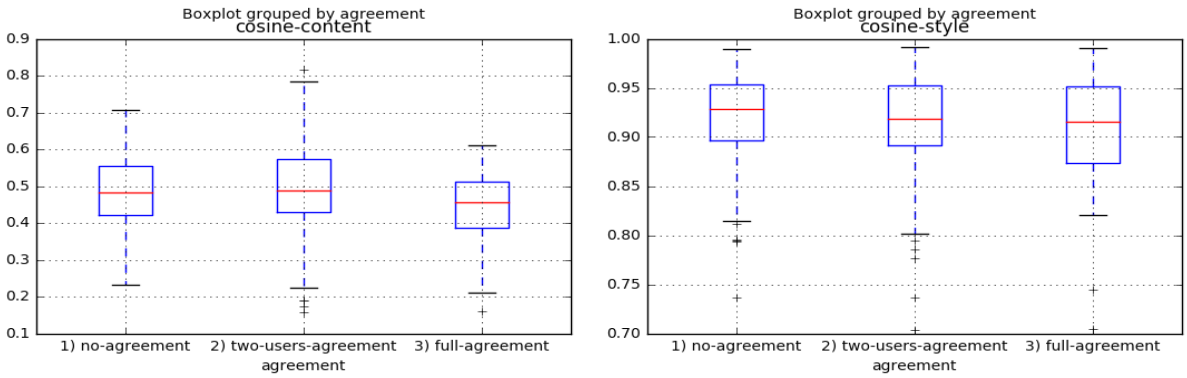
Feature	Description
alpha-chars-ratio	the fraction of total characters in the paragraph which are letters
digit-chars-ratio	the fraction of total characters in the paragraph which are digits
upper-chars-ratio	the fraction of total characters in the paragraph which are upper-case
white-chars-ratio	the 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

Table 2: List of stylometric features used (Tweedie & Baayen, 2002).

[4] https://dkpro.github.io/dkpro-statistics/

First, we select a list of stylometric features to extract from the source and the target texts. The literature suggests a broad amount of stylometric features (Mosteller & Wallace, 1964; Tweedie & Baayen, 2002; Stamatatos, 2009). Table 2 presents the list of features we extract for each snippet. In addition, we calculate the minimum, maximum, average and variance for each of those features across every snippet.

We consider the similarity between the source and the targets as a cosine similarity between the stylometric feature vectors. As depicted in Figure 3, we created box-plots to study whether there is a correlation between the user agreement and the content similarity (a) and one between the user agreement and the writing style similarity (b).



(a) (b)

Figure 3: Box-plots representing the distribution of the annotators' agreement over: a) the similarity of the content; b) the similarity of the writing style.

There is no clear evidence that explains the agreement/disagreement among annotators from the considered features. To dig more deeply, in Figure 4 we created a scatter plot in order to comprehend whether there is a correlation between the three similarities, i.e. content similarity, the writing style similarity and the inter-rater agreement.

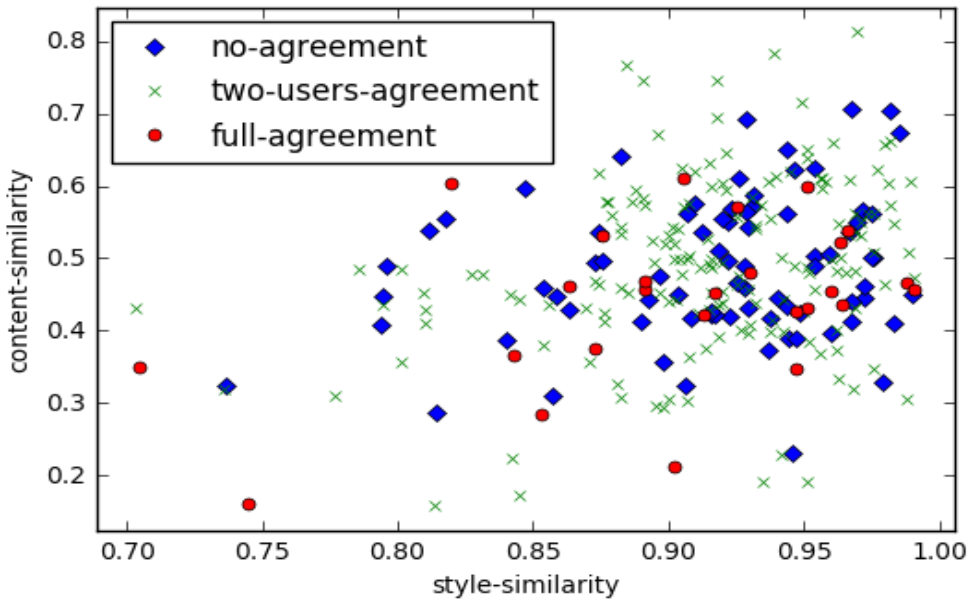


Figure 4: Scatter plot relating three dimensions: style similarity, content similarity and agreement between annotators.

The scatter plot does not provide any visual hint about the annotators' agreement/disagreement. In addition, we plotted every combination considering, instead of the whole vector of the aforementioned features (see Table 2), each of them singularly. Yet, we didn't notice any clear pattern. As there is no additional information added to the previous plot we omitted them in this paper.

Finally, we empirically measured whether the annotators did their ranking in a random manner. We run 500.000 rounds of random studies and, for each of them, calculate the inter-rater agreement using the Krippensdorff's alpha. The results show an average of 0.250 with a variance of 0.020. 28% of the cases have a larger agreement than our study, thus we can conclude with a confidence of 72% that the annotators in our experiment didn't rank in a random manner.

Conclusion and Future Work

In this paper, we proposed to include author's writing style as a new dimension of retrieving scientific literature. As preparation to automate this process, we have conducted a pilot study to learn whether the humans can distinguish text written by the same author when the topological information is removed. Our analyses show that this is challenging task, and there is no clear indicator for the choices of the humans. We provide the dataset of this study for the research community to make further investigations. In future work, we also plan to extend the study by increasing and diversifying the set of experiments aiming to capture, from human annotators, properties of the thinking process while performing this task.

Acknowledgements

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

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 Human Language Technology Conference / North American Chapter of the Association for Computational Linguistics.

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

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

Dabrowska, A. & Larsen, B. (2015). Exploiting citation contexts for physics retrieval. 3rd International Workshop on Bibliometric-enhanced Information Retrieval (BIR).

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.

Eck, N.J., & Waltman, L. (2014). Systematic Retrieval of Scientific Literature based on Citation Relations: Introducing the CitNetExplorer Tool. 2nd International Workshop on Bibliometric-enhanced Information Retrieval (BIR).

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), pp. 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 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.

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. *Advances in Information Retrieval* (pp. 798-801). Springer International Publishing.

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

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, pp. 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 33rd SGAI International Conference.

Rexha, A., Klampfl, S., Kröll, M. & Kern. R. (2015). Towards authorship attribution for bibliometrics using stylometric features. Proceedings 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.

Rexha, A., Klampfl, S., Kröll, M. & Kern. R. (2016). Towards a more fine grained analysis of scientific authorship: Predicting the number of authors using stylometric features. 4th International Workshop on Bibliometric-enhanced Information Retrieval (BIR).

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.

Zweigenbaum, P., Demner-Fushman, D., Yu, H., and Cohen, K. 2007. Frontiers of biomedical text mining: Current progress. *Briefings in Bioinformatics*, *8(5)*.