Poster

STI 2022 Conference Proceedings

Proceedings of the 26th International Conference on Science and Technology Indicators

All papers published in this conference proceedings have been peer reviewed through a peer review process administered by the proceedings Editors. Reviews were conducted by expert referees to the professional and scientific standards expected of a conference proceedings.

Proceeding Editors

Nicolas Robinson-Garcia
Daniel Torres-Salinas
Wenceslao Arroyo-Machado

OPEN ACCESS

Citation: Arhiliuc, C. (2022). Discipline classification of research publications using keyword extraction algorithms. In N. Robinson-Garcia, D. Torres-Salinas, & W. Arroyo-Machado (Eds.), *26th International Conference on Science and Technology Indicators*, STI 2022 (sti22128). https://doi.org/10.5281/zenodo.6975435

Collection: https://zenodo.org/communities/sti2022grx/

Discipline classification of research publications using keyword extraction algorithms

Cristina Arhiliuc*

*cristina.arhiliuc@uantwerpen.be
ECOOM, University of Antwerp, Prinsstraat 13, Antwerp 2000 (Belgium)

Introduction

This research aims to use keyword extraction algorithms to perform an efficient and transparent classification of scientific publications into disciplines. Furthermore, having the publications represented by keywords offer a better insight into the apparition and the evolution of different topics.

Our research focuses mainly on data from VABB-SHW database, which contains the publications authored by researchers from the Social Sciences and Humanities (SSH) departments of the Flemish for the period 2000-2019. Currently, the database contains two types of discipline classification: organisational classification (based on the discipline(s) of the unit(s) of research of the authors) and channel-based cognitive classification (based on the discipline(s) assigned to the channel of publication – journal, conference, etc.). We aim to extend these two types of classification with a content-based cognitive classification based on text analysis, as we don't have any citation data available.

Previous attempts to classify and cluster research papers in disciplines have been done in (Glänzel et al., 1999; Kandimalla et al., 2021; Matwin & Sazonova, 2012; Suominen & Toivanen, 2016; Weber et al., 2019), including on this data (Eykens et al., 2021), however this approaches relied solely on document embedding techniques and don't offer any intermediate information about the publication. A keyword-based classification technique provides more information about the publication and offers more interpretability to the results. Moreover, the intermediate results (keywords and key phrases) can be used for further research that doesn't concern discipline classification or that require a more fine-grained classification based on the subject(s) of the publication.

Data

The database contains initially only the abstracts for a part of the publications. However, as a result of an on-going project that collects affiliation data and other relevant data, full-texts in .pdf format have become available for some publications.

Currently the database contains 119 652 publications, out of which 49 510 have available abstracts and 3023 have full text. Table 1 shows the distribution of the abstracts, full texts and the two main languages (Dutch and English) across disciplines. Although the database

contains other languages too, the number of publications is largely inferior. The disciplines with generic names as "Other disciplines" have been excluded.

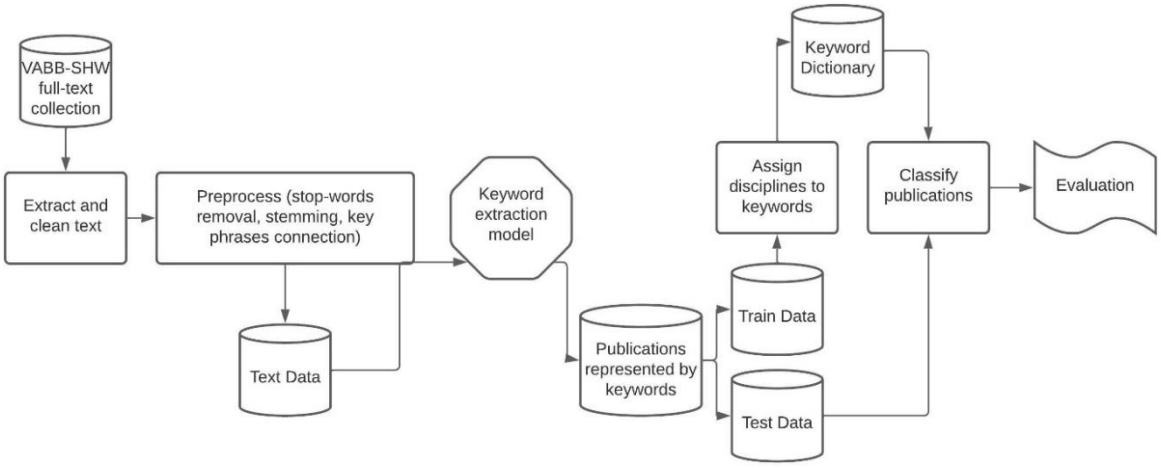
Table 1. Distribution of publications, abstracts and full texts for different disciplines based on organisational classification

Discipline	Total pub.		Abstracts		Full-text	
	English	Dutch	English	Dutch	English	Dutch
Archaeology	2114	449	765	130	67	42
Art History	4630	971	2382	303	118	42
Communication Studies	4015	668	2331	242	191	47
Criminology	1669	1903	765	446	68	114
Economics & Business	15746	1541	8638	366	443	132
Educational Sciences	5384	936	3093	236	144	88
History	3928	1671	1498	408	83	61
Law	7437	9311	2320	1114	187	213
Linguistics	7545	1448	3069	450	180	63
Literature	2690	1382	809	276	68	33
Philosophy	6170	1116	2778	249	97	44
Political Sciences	4516	1165	1772	232	125	106
Psychology	12030	1046	6625	276	161	99
Social Health Sciences	16924	1264	11091	379	227	121
Sociology	5561	1310	3080	298	127	166
Theology	2422	711	596	118	20	14

In terms of data, it is challenging to collect and process the full text .pdf files, especially given the structural differences between them. Currently, GROBID[1] library is used to get the text from .pdf, followed by additional cleaning steps to ensure the correct management of tables and formulae, that are usually explained in the core text anyway. However, no perfect process has been found. Consequently, for a first iteration to test the model, 100 texts extracted from publications will be manually checked and cleaned.

Methodology

Figure 1. Graphic representation of the basic methodology



[1] https://grobid.readthedocs.io/

Keywords extraction process depends on the data provided, so we will firstly clean and prepare the data. As explained in the Data section, the first step is making sure that the data is well aligned and additional data like footers, headers, captions, reference and table data is removed or transferred to another file for further use. Additional pre-processing steps will depend on the model used. If the model can't recognize key phrases (frequent relevant word n-grams), then they will be identified in the pre-processing phase and connected so the algorithm would see them as one word. For this a combination of simple identification of words frequently used together and noun groups will be used. Stop words and words that are very frequent in all or the big majority of publications will also be removed.

For the keyword extraction, TF-IDF and KeyBert will be applied in the first iteration. The KeyBert algorithm offers the possibility to extract key phrases too in a given n-gram range. The two algorithms will be applied for both English and Dutch, although KeyBert might require small adaptations to use BERTje (de Vries et al., 2019) model for Dutch instead of BERT (Devlin et al., 2019). As a result of the keyword extraction, top n keywords/key phrases are going to be kept according to their relevance. The n parameter is to be determined after further experiments. For example, for the article (Luukkonen et al., 1992), the keyword extraction model based on KeyBert without additional pre-processing has identified keywords as: "countries collaboration", "research collaboration", "international scientific collaboration", "international collaboration tendency", "authorship measures economic", "international science" and others.

Further, we are going to use the current organizational and channel-based cognitive classification of the publications to statistically assign disciplines to the extracted keywords. To avoid the noise, only top m disciplines per keyword/key phrase are going to be kept according to their statistical frequency and/or the sum of scores accumulated in the previous step.

Next, the keywords and their associated disciplines are going to be used for the classification of new unseen test data. This classification can be rigid (one discipline per publication) or fuzzy (more disciplines per publication). Possible extensions of the algorithm are possible at this step. The dictionary obtained from the training data can be extended with new unseen keywords and key phrases from the new data. For this, the new keyword will be ignored during the classification and the disciplines assigned to the publications will serve as a base for the keyword. A keyword should not be used for classification unless it reached a pre-fixed number of occurrences in publications, otherwise its discipline assignment is unreliable. In a similar way, the dictionary can be updated after each new publication classified, however, this might present further risks if used incorrectly (for example, one text introduced multiple times).

A manual evaluation of the model will happen when extracting keywords (20 random publications), assigning disciplines to keywords (100 random keywords) and final classification (100 random publications). Additionally, we will do an evaluation based on the organizational and channel-based cognitive classification for comparison purposes. However, we do not aim for very high scores for this type of evaluation.

Discussions

The methodology proposed in this research in progress paper aims to provide a good content-based classification for the articles from VABB-SHW. However, the keywords extracted could be used for further analysis like the evolution of topics and their dynamics in different

disciplines, countries, universities, departments etc. It could also show the emergence of new subjects and disciplines. We expect, by the time of the conference, to have more results.

For further research, it would be interesting extend our data with publications from other national and international databases and with non-SSH publications.

References

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019). BERTje: A Dutch BERT Model. *ArXiv:1912.09582 [Cs]*. http://arxiv.org/abs/1912.09582

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *ArXiv:1810.04805 [Cs]*. http://arxiv.org/abs/1810.04805

Eykens, J., Guns, R., & Engels, T. C. E. (2021). Fine-grained classification of social science journal articles using textual data: A comparison of supervised machine learning approaches. *Quantitative Science Studies*, *2*(1), 89–110. https://doi.org/10.1162/qss_a_00106

Glänzel, W., Schubert, A., & Czerwon, H.-J. (1999). An item-by-item subject classification of papers published in multidisciplinary and general journals using reference analysis. *Scientometrics*, *44*(3), 427–439. https://doi.org/10.1007/BF02458488

Kandimalla, B., Rohatgi, S., Wu, J., & Giles, C. L. (2021). Large Scale Subject Category Classification of Scholarly Papers With Deep Attentive Neural Networks. *Frontiers in Research Metrics and Analytics*, *5*. https://www.frontiersin.org/article/10.3389/frma.2020.600382

Luukkonen, T., Persson, O., & Sivertsen, G. (1992). *Understanding Patterns of International Scientific Collaboration*. *Science, Technology, & Human Values*, *17*(1), 101–126. https://doi.org/10.1177/016224399201700106

Matwin, S., & Sazonova, V. (2012). Direct comparison between support vector machine and multinomial naive Bayes algorithms for medical abstract classification. *Journal of the American Medical Informatics Association*, *19*(5), 917. https://doi.org/10.1136/amiajnl-2012-001072

Suominen, A., & Toivanen, H. (2016). Map of science with topic modeling: Comparison of unsupervised learning and human-assigned subject classification. *Journal of the Association for Information Science and Technology*, *67*(10), 2464–2476. https://doi.org/10.1002/asi.23596

Weber, T., Kranzlmüller, D., Fromm, M., & de Sousa, N. T. (2019). Using Supervised Learning to Classify Metadata of Research Data by Discipline of Research. *ArXiv:1910.09313 [Cs, Stat]*. http://arxiv.org/abs/1910.09313