Into the bibliography jungle: using random forests to predict dissertations' reference section

Silvia E. Gutiérrez De la Torre[1,*], Julián Equihua[2], Andreas Niekler[1] and Manuel Burghardt[1]

[1]*Computational Humanities Group (Leipzig University)*
[2]*Helmholtz-Centre for Environmental Research (Leipzig)*

Abstract

Cited-works-lists in Humanities dissertations are typically the result of five years of work. However, despite the long-standing tradition of reference mining, no research has systematically untapped the bibliographic data of existing electronic thesis collections. One of the main reasons for this is the difficulty of creating a tagged gold standard for the around 300 pages long theses. In this short paper, we propose a page-based random forest (RF) prediction approach which uses a new corpus of Literary Studies Dissertations from Germany. Moreover, we will explain the handcrafted but computationally informed feature-selection process. The evaluation demonstrates that this method achieves an F1 score of 0.88 on this new dataset. In addition, it has the advantage of being derived from an interpretable model, where feature relevance for prediction is clear, and incorporates a simplified annotation process.

Keywords

electronic theses and dissertations, bibliographic reference parsing, information retrieval, machine learning

CEUR-WS.org/Vol-3220/invited-talk1.pdf

1. Introduction

Citation analysis (CA) of dissertations, that is, the examination of cited works in theses, has a long story within bibliometric studies but still no computational operationalization. Most approaches address collection development needs in libraries and thus seek to ascertain what types of documents are the most frequently used in the doctoral research stage [1, 2, 3, 4, 5]. To a lesser degree, some studies have investigated other research behaviors such as interdisciplinarity [6, 7], language use [8, 9, 10], and domain-specific trends in specific domains such as chemistry [4, 2], library science [8, 6, 11], sociology/anthropology [10], atmospheric science [12], agriculture/biology [13] and mathematics education [14]. Because of tedious manual extraction, so far, only small-scale studies are to be found; the big picture of citation strategies in Ph.D. theses is, however, yet to be painted. In this paper, we propose a computational approach to automatically mine references from a large corpus of – mostly – German-language dissertations in the field of literary studies. The ultimate goal of this endeavor is to investigate the epistemological

Understanding LIterature references in academic full TExt at JCDL 2022, June 24, 2022, Köln, Germany
*Corresponding author.
silviaegt@uni-leipzig.de (S. E. G. D. l. Torre); julian.equihua@ufz.de (J. Equihua); aniekler@informatik.uni-leipzig.de (A. Niekler); burghardt@informatik.uni-leipzig.de (M. Burghardt)
0000-0002-0877-7063 (S. E. G. D. l. Torre); 0000-0002-3036-3318 (A. Niekler); 0000-0003-1354-9089 (M. Burghardt)
© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

foundations of this field by means of systematic, large-scale, citation analysis. While there are a number of examples of citation analyses in the humanities [12] and even more specifically in literary studies[15, 16, 17, 18, 19], each of those has only used monographs and journals as their sources, but not dissertations. Only one study from the 1990s has tapped into the richness of citations in literary studies dissertations, manually analyzing 26 theses defended in one university in India [20]. The reasons for a gap in large-scale dissertations' CA may be threefold: First, the extension of the reference section of dissertations (which in our corpus averages 25 pages in length) implies a great challenge for in-depth tagging, which is a necessary step for machine learning approaches. Second, the assumed inconsistency of references in the humanities[21] is particularly true for dissertations, which tend to lack conventionalized editorial guidelines. Third, and related to the first two: while reference mining tools for the automatic extraction of citations have become more sophisticated in recent years, they are all trained on gold standards of English papers from the natural and applied sciences[22, 23, 24, 25]. The aftermath of this is that they are hardly applicable to other domains, languages, and document types. To close this gap and address these challenges, we present ongoing work on a page-based random forest prediction approach, which allows us to extract bibliography sections (in various styles and formats) from a corpus of 1,330 literary studies dissertations. This task is an important prerequisite for the later exploration of specific bibliographic information, which will be used to reveal citation strategies and trends in this field.

2. Related work

As for related studies to our approach, the EXCITE project aims at extracting citations outside the natural and applied sciences Anglosphere. Its purpose is to develop "a set of algorithms for the extraction of citation and reference information" for the social sciences[26]. Yet, the transparency with which they share their gold standard shows a heavy inclination towards journal articles and the inclusion of fewer than five dissertations. The only large-scale precedent on the dissertation parsing front is the "Opening Books and the National Corpus of Graduate Research" project[27]. Their wide scope of tasks includes reference mining. However, so far, the team has only released its code on DOI metadata retrieval and has expressed its intention to identify "particular pieces of information within the citations such as the author names"[28]. There are no available pipelines yet for the first task in reference mining: detecting the reference section[26]. Another interesting approach is a line-based conditional random field, which showed to reduce model complexity of reference string extraction by detecting relevant strings without first identifying the reference section[26]. Yet, this supervised learning algorithm requires line-wise annotations that are hardly scalable for the ca. 12,000 lines long dissertations. Especially considering that all dissertations have a bibliography section, unlike certain journals.

3. Methodology

Corpus: To test our approach, we gathered 1,330 electronic theses with their corresponding metadata from the German National Library. The selection criteria were subject and temporality: Literary Studies dissertations, which were defended in German Universities after the

dates_sw	pubplaces	page_n	bibtest
0.06211	0.01527	152	0
0.06338	0.00352	153	0
0.06609	0.00575	154	1
0.08397	0	155	1

Table 1
Example of tagged dataset to train RF model, each line is one page of the same PDF file

feature	MDA
pubplaces_sw	32.3184181
dates_sw	24.87243224
dates	22.39062468
position	22.08909718
pubplaces	20.99889236

Table 2
Top 5 features by mean decrease accuracy (MDA)

reunification (1990) until 2020. After cleaning duplicates and misclassified documents, we were left with 1,116 electronic dissertations and their corresponding PDF files. *Model, features, and sample selection*: The machine learning method we chose for its interpretable model is random forests (RF). Breiman (2001) developed RF as a machine learning technique for classification and regression[29]. An RF is a collection of "decision trees". Each node in a tree (i.e., each decision) is based on a random subset of the available features, automatically predicting the likelihood that an item belongs to a specific class (in our case: if the page was part of the bibliography section or not). We used the randomForest package implementation in R[30].

The selected features were the relative frequencies of bibliographic elements that appear in other reference mining projects: 1) dates, both in a four numbers date and with special but common formatting in bibliographies, namely between parenthesis; 2) publication places with a consistent frequency (we selected fifteen); 3) different bibliography-section-headers in English, Spanish, French, and German. Common abbreviations for references in these languages such as: "Ed(s)", "Hrsg.", "Vol", etc. were added as suggested by previous work on reference mining German academic texts[26]. We also added typical abbreviations in the footnotes (such as "vgl.", "ebd.", "ders."). We added this last feature to push the label to a negative class (i.e. "not bibliography section"). Finally, we considered page number ("n_pag"), the total number of pages ("n_pags"), and position (i.e. n_pag/n_pags).

Furthermore, we used a sliding windows approach to these features. Since bibliographies are a section, we wanted to get a feature measure that overcomes single pages as barriers. Sliding windows are useful to compute a running average over adjacent pairs, this allows to catch the presence of features across a range of pages rather than on singleton ones, which also contain some selected features but are not part of the section. For instance, the sliding-window mean for dates in the first three pages is 0.042 (that is: the average of 1/37 + 1/32 + 2/29) and it is compared against those of the second window (0.054) and the third one 0.044 (see Figure 1). We used this sliding windows approach for dates, publication places, abbreviations, page numbers, positions, and section headings. In order to create a representative sample of different types of dissertations' bibliographies, we selected documents with the most varied distribution of these features. We selected 11 different k-means clusters of differentiated features distribution and 5 different documents from each one of these clusters. We then proceeded to create a training data frame. Each row represents while columns contain the relative frequency for each feature except for the last one which is for the class to be predicted ("bibtest"). This last column contains 0's for non-bibliography sections, and 1's for positive cases (see Table 1).

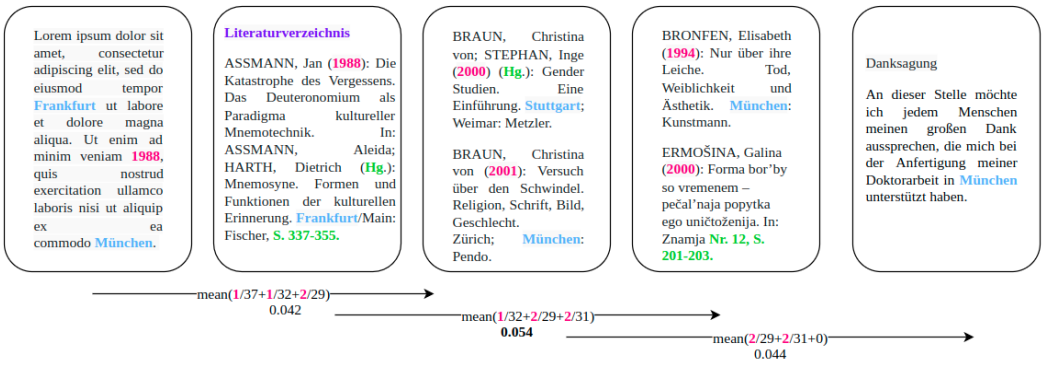

Figure 1: Example of sliding windows calculation.

4. Results and future work

We trained the RF on 1,000 trees and followed a bootstrap strategy, which roughly contains two-thirds of the observations[31]. The confusion matrix shows more errors in classifying content text as part of the reference section (304 false positives (FP), class error 0.016) while only 9 out of 1,145 reference section pages got erroneous predictions (false negatives, FN). In other numbers, we got a precision of 0.79, a recall of 0.99, and an overall F1 score of 0.87. Moreover, we were able to identify the most relevant features by their mean decrease accuracy and observe the relevance of both our proposed sliding windows method ('_sw' features in Table 1 and 2) and of the selected features.

Besides these extremely promising results, the problems of our approach are worth mentioning. As previously stated, we get a relevant amount of FPs. Looking at concrete examples of this, we can identify pages which are quite similar to true bibliographies that get misclassified. For instance, we found containing pages containing a single footnote with references or where the author lists works that are not cited in the scholarly sense. Likewise, lists of figures and of abbreviations are sources of errors, as they superficially appear very similar to the bibliographies. Furthermore, derived from our sliding windows approach, bibliography-like pages that are nearby the reference section also get incorrect predictions. On the FN classification side, the formatting and layout of the bibliographies is often very unusual, or they do not begin on a new page. We can experimentally trace both observations back to the feature level and show that the feature structure differs in these examples. The biggest difference is in position, as FP are on a lower page number than true-positives. Also, the distribution of dates and publication places is very different, as FP contain a lower number of both. We thus need to complement heuristic plausibility checks for the application of the approach, which eliminate possible misclassifications based on additional conditions, e.g., checking the position in the text. In addition, we consider complementary Regular Expressions that can automate additional plausibility checking of the result set. However, we realize that these additional corrections would have to be adapted for dissertations from other fields.

Acknowledgments

Authors Silvia Gutiérrez and Julián Equihua have completed these experiments while receiving a doctoral research grant from the German Academic Exchange Service (DAAD)

References

[1] P. M. Beile, D. N. Boote, E. K. Killingsworth, A Microscope or a Mirror? A Question of Study Validity Regarding the Use of Dissertation Citation Analysis for Evaluating Research Collections, The Journal of Academic Librarianship 30 (2004) 347–353. URL: http://www.sciencedirect.com/science/article/pii/S0099133304001041. doi:10.1016/j.acalib.2004.06.001.

[2] N. Vallmitjana, L. G. Sabaté, Citation Analysis of Ph.D. Dissertation References as a Tool for Collection Management in an Academic Chemistry Library, College & Research Libraries 69 (2008) 72–82. URL: http://crl.acrl.org/index.php/crl/article/view/15913. doi:10.5860/crl.69.1.72.

[3] T. P. Franks, D. S. Dotson, Book Publishers Cited in Science Dissertations: Are Commercial Publishers Worth the Hype?, Science & Technology Libraries 36 (2017) 63–76. URL: https://doi.org/10.1080/0194262X.2016.1263172. doi:10.1080/0194262X.2016.1263172.

[4] L. Zhang, A Comparison of the Citation Patterns of Doctoral Students in Chemistry versus Chemical Engineering at Mississippi State University, 2002–2011, Science & Technology Libraries 32 (2013) 299–313. URL: http://www.tandfonline.com/doi/abs/10.1080/0194262X.2013.791169. doi:10.1080/0194262X.2013.791169.

[5] P. C. Johnson, Dissertations and discussions: engineering graduate student research resource use at New Mexico State University, Collection Building 33 (2013) 25–30. URL: http://www.emeraldinsight.com/doi/10.1108/CB-09-2013-0037. doi:10.1108/CB-09-2013-0037.

[6] C. R. Sugimoto, Mentoring, collaboration, and interdisciplinarity: An evaluation of the scholarly development of Information and Library Science doctoral students, Ph.D. Thesis, University of North Carolina at Chapel Hill, 2010.

[7] W. R. Fernandes, B. V. Cendón, C. A. A. Araújo, Ciência da informação e áreas correlatas: um estudo de caso na Universidade Federal de Minas Gerais, Brazilian Journal of Information Science 5 (2011) 3–35. Publisher: Universidade Estadual Paulista.

[8] T. LaBorie, M. Halperin, Citation Patterns in Library Science Dissertations, Journal of Education for Librarianship 16 (1976) 271–283. URL: https://www.jstor.org/stable/40322465. doi:10.2307/40322465, publisher: Association for Library and Information Science Education (ALISE).

[9] S.-J. Gao, W.-Z. Yu, F.-P. Luo, Citation analysis of PhD thesis at Wuhan University, China, Library Collections, Acquisitions, and Technical Services 33 (2009) 8–16. URL: http://linkinghub.elsevier.com/retrieve/pii/S1464905509000281. doi:10.1016/j.lcats.2009.03.001.

[10] Z. Rosenberg, Citation Analysis of M.A. Theses and Ph.D. Dissertations in Sociology and Anthropology: An Assessment of Library Resource Usage, The Journal of Academic

Librarianship 41 (2015) 680–688. URL: http://www.sciencedirect.com/science/article/pii/S0099133315001007. doi:10.1016/j.acalib.2015.05.010.

[11] R. Echezona, Principal, V. Okafor, S. C. Ukwoma, Information Sources Used by Postgraduate Students in Library and Information Science: A Citation Analysis of Dissertations 2011 (2011).

[12] S. Kaczor, A Citation Analysis of Doctoral Dissertations in Atmospheric Science at the University at Albany, Science & Technology Libraries 33 (2014) 89–98. URL: http://www.tandfonline.com/doi/abs/10.1080/0194262X.2013.866067. doi:10.1080/0194262X.2013.866067.

[13] P. U. Kuruppu, D. C. Moore, Information Use by PhD Students in Agriculture and Biology: A Dissertation Citation Analysis, portal: Libraries and the Academy 8 (2008) 387–405. URL: http://muse.jhu.edu/content/crossref/journals/portal_libraries_and_the_academy/v008/8.4.kuruppu.html. doi:10.1353/pla.0.0024.

[14] A. Fernández Cano, M. Torralbo, L. Rico, P. Gutiérrez, A. Maz, Análisis cientimétrico, conceptual y metodológico de las tesis doctorales españolas en Educación Matemática (1976-1998), Revista Española de Documentación Científica 26 (2003) 162–176. URL: https://redc.revistas.csic.es/index.php/redc/article/view/135/189, publisher: Universidad de Granada.

[15] C. O. Frost, The Use of Citations in Literary Research: A Preliminary Classification of Citation Functions, The Library Quarterly 49 (1979) 399–414. URL: https://www.journals.uchicago.edu/doi/abs/10.1086/600930. doi:10.1086/600930.

[16] J. Ardanuy, C. Urbano, L. Quintana, The Evolution of Recent Research on Catalan Literature through the Production of PhD Theses: A Bibliometric and Social Network Analysis, Information Research: An International Electronic Journal 14 (2009). URL: https://eric.ed.gov/?id=EJ851921.

[17] J. W. Thompson, The death of the scholarly monograph in the humanities? Citation patterns in literary scholarship, Libri 52 (2002) 121–136.

[18] R. Heinzkill, Characteristics of References in Selected Scholarly English Literary Journals, The Library Quarterly: Information, Community, Policy 50 (1980) 352–365. URL: https://www.jstor.org/stable/4307248.

[19] D. S. Nolen, H. A. Richardson, The search for landmark works in English literary studies: a citation analysis, The Journal of Academic Librarianship 42 (2016) 453–458.

[20] V. N. Deo, S. M. Mohal, Bibliometric study of doctoral dissertations on English language and literature, Annals of Library and Information Studies 42 (1995) 81–95.

[21] D. Rodrigues Alves, G. Colavizza, F. Kaplan, Deep Reference Mining From Scholarly Literature in the Arts and Humanities, Frontiers in Research Metrics and Analytics 3 (2018). URL: https://www.frontiersin.org/articles/10.3389/frma.2018.00021/full. doi:10.3389/frma.2018.00021.

[22] P. Lopez, Grobid, 2019. URL: https://github.com/kermitt2/grobid.

[23] D. Tkaczyk, P. Szostek, M. Fedoryszak, P. J. Dendek, L. Bolikowski, CERMINE: automatic extraction of structured metadata from scientific literature, International Journal on Document Analysis and Recognition (IJDAR) 18 (2015) 317–335. URL: https://doi.org/10.1007/s10032-015-0249-8. doi:10.1007/s10032-015-0249-8.

[24] I. G. Councill, C. L. Giles, M.-Y. Kan, ParsCit: an Open-source CRF Reference String Parsing

Package., in: LREC, volume 8, 2008, pp. 661–667.
[25] A. Prasad, M. Kaur, M.-Y. Kan, Neural ParsCit: a deep learning-based reference string parser, International Journal on Digital Libraries 19 (2018) 323–337. URL: https://doi.org/10.1007/s00799-018-0242-1. doi:10.1007/s00799-018-0242-1.
[26] M. Körner, B. Ghavimi, P. Mayr, H. Hartmann, S. Staab, Evaluating Reference String Extraction Using Line-Based Conditional Random Fields: A Case Study with German Language Publications, in: M. Kirikova, K. Nørvåg, G. A. Papadopoulos, J. Gamper, R. Wrembel, J. Darmont, S. Rizzi (Eds.), New Trends in Databases and Information Systems, Communications in Computer and Information Science, Springer International Publishing, Cham, 2017, pp. 137–145. doi:10.1007/978-3-319-67162-8_15.
[27] W. A. Ingram, E. A. Fox, J. Wu, Opening Books and the National Corpus of Graduate Research, LG-37-19-0078-19, 2019. URL: https://www.imls.gov/grants/awarded/lg-37-19-0078-19.
[28] B. Ingram, B. Banerjee, S. Kahu, Classification and extraction of information from ETD documents, 2019. URL: https://github.com/Opening-ETDs/CS6604-ETD.
[29] L. Breiman, Random Forests, Machine Learning 45 (2001) 5–32. URL: https://doi.org/10.1023/A:1010933404324. doi:10.1023/A:1010933404324.
[30] A. Liaw, M. Wiener, Classification and Regression by randomForest, R News 2 (2002) 18–22. URL: https://cran.r-project.org/doc/Rnews/Rnews_2002-3.pdf.
[31] B. Efron, R. Tibshirani, Improvements on Cross-Validation: The .632+ Bootstrap Method, Journal of the American Statistical Association 92 (1997) 548. URL: https://www.jstor.org/stable/2965703?origin=crossref. doi:10.2307/2965703.