Toward a computational history of universities: Evaluating text mining methods for interdisciplinarity detection from PhD dissertation abstracts

Federico Nanni
International Centre for the History of Universities and Science, University of Bologna and Data and Web Science Group, University of Mannheim

Laura Dietz
Department of Computer Science, University of New Hampshire, USA

Simone Paolo Ponzetto
Data and Web Science Group, University of Mannheim

Abstract

For the first time, historians of higher education have large data sets of primary sources that reflect the complete output of academic institutions at their disposal. To analyze this unprecedented abundance of digital materials, scholars have access to a large suite of computational methods developed in the field of Natural Language Processing. However, when the intention is to move beyond exploratory studies and use the results of such analyses as quantitative evidences, historians need to take into account the reliability of these techniques. The main goal of this article is to investigate the performance of different text mining methods for a specific task: the automatic identification of interdisciplinary works from a corpus of PhD dissertation abstracts. Based on the output of our study, we provide the research community of a new data set for analyzing recent changes in interdisciplinary practices in a large sample of European universities. We show the potential of this collection by tracking the growth in adoption of computational approaches across different research fields, during the past 30 years.

Correspondence:
Federico Nanni, University of Mannheim, B6, 26, D-68159, Mannheim, Germany.
E-mail:
federico@informatik.uni-mannheim.de

1 Introduction

During the past decades, policymakers, government agencies, and private companies have been trying to encourage academia to conduct more and more interdisciplinary research, with the goal of addressing complex challenges and accelerating innovations across industries and branches of knowledge (Holm *et al.*, 2013; Allmendinger, 2015). Interdisciplinary practices, and in particular collaborative works employing computational methods, are sustained and

Digital Scholarship in the Humanities, Vol. 33, No. 3, 2018.
doi:10.1093/llc/fqx062 Advance Access published on 11 December 2017

fostered by dispensing grants, scholarships, and establishing direct collaborations between companies and research groups.

Historians of higher education, who aim to understand whether rhetoric and funding have played a significant role in recent years in orienting the research focus and practices of universities, now have at their disposal digital databases of PhD dissertations, which diachronically reflect academic outputs in their entirety (Ramage *et al.*, 2011). From the field of scientometrics (Van Raan, 1997), scholars can borrow many approaches for automatically mapping interdisciplinary works in large scientific corpora (Rafols and Meyer, 2010); unfortunately, these graph-based techniques (Lu and Wolfram, 2012) strictly depend on the direct accessibility of bibliographic data, which are in most cases not easily obtainable. Instead, data that are promptly available in large abundance across all countries and academic institutions are dissertation abstracts.[1]

In this article, we investigate whether text mining methods, such as Latent Dirichlet Allocation (LDA) topic models (Blei *et al.*, 2003), could represent a valid alternative for historians interested in identifying interdisciplinary practices directly from the textual content of dissertation abstracts. Next, we build upon the obtained results, to create a large-scale data set for the study of interdisciplinary research in academia; we show the usefulness of this new collection for tracking the so-called 'computational turn' (Berry, 2011), namely, the recent growth in adoption of computational methods across different research areas, from life sciences to the social sciences and the humanities. By addressing these two goals, our work aims to be both a contribution to the recent debate on the importance of tool criticism in the digital humanities (Traub and van Ossenbruggen, 2015) and a step toward a 'computational history' (Turkel, 2008) of interdisciplinary research.

1.1 Defining interdisciplinary research and recognizing it from text

Even if interdisciplinarity is a recurrent topic in research, defining it as a quantifiable property of an academic work remains extremely challenging, even today. In fact, as it has been already remarked (Wagner *et al.*, 2011), this concept relies on the existence of a clear distinction between academic disciplines, which is still a disputed issue in the literature on higher education (Repko, 2008; Sugimoto and Weingart, 2015).

In applied scientometric research, disciplines are often identified with metadata information associated with the publication, such as the ISI Subject Categories (Rafols and Meyer, 2010), and their existence is therefore accepted as a starting point of the work. In our study, while we also do not question the existence of disciplines, we will nevertheless discuss how the metadata information we employed have been initially assigned to the publications, highlighting the social context and implications of such decisions.

In the literature on higher education, interdisciplinary research is defined as a 'process of answering a question, solving a problem, or addressing a topic that is too broad or complex to be dealt with adequately by a single discipline, and draws on the disciplines with the goal of integrating their insights to construct a more comprehensive understanding' (Repko, 2008). In scientometrics research interdisciplinarity is generally quantified by examining the network of citations and measuring for instance the percentage of citations outside the main discipline of the citing paper; instead, in this work we intend to detect interdisciplinary practices by the way research is described in the abstract of an academic work.

The adoption of text mining approaches for examining scientific publications is not new: Dietz *et al.* (2007) used LDA topic models to quantify the impact that research papers have on each other. A few years later, Gerrish and Blei (2010) showed that LDA is able to identify a qualitatively different set of relevant articles, when compared to traditional citation-count metrics; with the same method, Hall *et al.* (2008) identified different methodological trends in the field of computational linguistics across almost 30 years of publications.

Even the automatic detection of interdisciplinary practices from text has been already attempted with text mining approaches (Ramage *et al.*, 2011; Chuang *et al.*, 2012; Nichols, 2014), and in

particular with the use of LDA topic models. However these studies, which mainly employed LDA as a corpus exploration method, did not establish its reliability for spotting interdisciplinary works. As opposed to them, as a first step of this work, we intend to verify the usefulness of topic models for identifying interdisciplinary practices. To do so, we employ a corpus of PhD thesis abstracts collected from the Digital Library of the University of Bologna,[2] and we compare the performance of LDA to the results obtained by using other text mining methods.

2 Corpus

Recently, Italian universities started to offer online institutional repositories of the doctoral theses defended at their institutions. Each publication is stored under legal deposit at the National Libraries of Florence and Rome and is uniquely identified by the National Bibliography Number (NBN) and the Digital Object Identifier (DOI).

One of the largest data sets available is offered by the Digital Library of the University of Bologna. When this research was conducted, it consisted of 4,556 theses, defended between 2007 and 2015. Each of these theses is described with a series of metadata: title, author, short abstract (the majority of which are in English), names of the supervisors, etc. From this data set, we selected all the theses with an abstract in English (2,954) as our starting data set.

2.1 Discipline annotations

Our data set contains an explicit mention of the main discipline of each dissertation in the field 'Settore disciplinare' (subject area). This label, selected by the PhD candidate in agreement with the supervisor of the thesis, is extremely relevant in the Italian academic environment, as it identifies (and conditions) the future field of study of the researcher. The way the subject areas are defined by the Italian Ministry of Education, Universities and Research is still a widely debated issue, as the process is necessarily conditioned by long-term academic, historical, and sociopolitical reasons (Pascuzzi, 2014). Opposite examples of this phenomenon could be identified by the existence of two subject areas dedicated to Logic (the first under the main research-area of Mathematics and the second under Philosophy)—reconfirming the importance of the Italian school of Philosophy of Logic (Ballo and Franchella, 2006)—and by the absence of a subject area dedicated to the Digital Humanities (Orlandi and Mordenti, 2003), despite the large contribution of Italian scholars to the international research community.

Knowing that these specific limitations could slightly influence the output of our analysis (e.g. a thesis discussing modal logic will inevitably appear as interdisciplinary between Mathematics and Philosophy), we employ here the twenty-eight main disciplines as currently defined by the Ministry (statistics provided in Fig. 1).

Discipline	All	Int-Disc	Mono-Disc
Agriculture	233	14	22
Anthropology	13	1	0
Arts	55	4	5
Biology	303	4	16
Chemistry	262	8	15
Civil Engineering and Architecture	117	7	11
Classical Languages	29	0	3
Computer Engineering	203	4	6
Computer Science	58	2	7
Earth Science	110	5	6
Economics	122	1	7
Geography	11	0	0
History	69	3	3
Industrial Engineering	216	8	12
Law	179	6	6
Linguistics	70	6	3
Mathematics	36	3	1
Medicine	322	3	17
Oriental Studies	7	0	0
Pedagogy	19	0	1
Philology and Literary Studies	54	1	0
Philosophy	25	2	6
Physics	172	4	18
Political and Social Sciences	68	0	3
Psychology	73	1	6
Sport Science	9	0	0
Statistics	30	2	0
Veterinary	89	4	5

Fig. 1 The total number of abstracts for each discipline in our data set (All) and the number of inter-disciplinary (Int-Disc) and monodisciplinary (Mono-Disc) theses in our gold standard

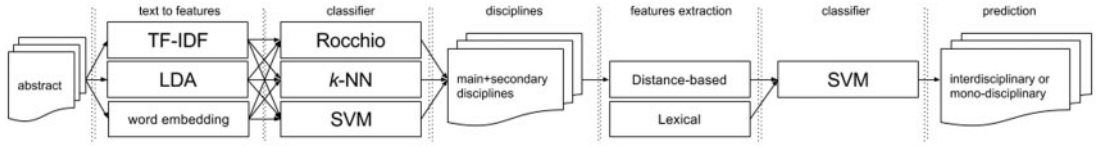

Fig. 2 Schema of the methods for discipline classification and interdisciplinary detection. Boxes with different options are depicted, for example distance-based features obtained from the results of Rocchio TF-IDF, SVM TF-IDF, etc

2.2 Interdisciplinary annotations

For assessing the quality of text mining methods for spotting interdisciplinary works, it is necessary to manually create a so-called 'gold standard' with annotations of (1) other secondarily related fields (e.g. 'this thesis is focused on Biology, but it involves the use of Computer Science methods'), and (2) whether the thesis should be considered interdisciplinary. We obtained the labels conducting a survey among all the supervisors of theses in our corpus. For each of their supervised dissertations, supervisors were asked:

- By considering the Subject Areas as main disciplines, was the thesis interdisciplinary?
- Which are the secondary disciplines?

The survey leads to a collection of expert human assessment on a subset of 272 theses (93 interdisciplinary and 179 mono-disciplinary theses). The frequency statistics is presented in Fig. 1.

3 From Text to Feature-Vectors

To automatically analyze the abstracts, it is necessary to represent each of them with a vector of numeric values (as depicted in Fig. 2). In this work, we consider three different ways of generating such vectors from textual contents. The first is a widely adopted approach in Natural Language Processing (NLP): the term frequency–inverse document frequency (TF-IDF). This method maps each word in a thesis-abstract to a numeric value: it gives higher scores to terms with high frequency within a few documents and decreases the importance of words occurring in many documents (e.g. stopwords).

The second method is based on the use of LDA topic models. As in Chuang *et al.* (2012),[3] we first ran LDA on the entire corpus, and next we represented each thesis-abstract as a vector of LDA topic-values. Therefore, the vector-size is determined by the number of topics.

Both LDA and TF-IDF are vector representation methods based on the 'bag-of-words' assumption, meaning that the order of words does not play any role in the model. The third method we consider for representing the abstract as a vector of values is based on the use of so-called 'word-embeddings' (Mikolov *et al.*, 2013). These are a class of language modeling approaches that, by studying the local context of each word (meaning the other surrounding words), define the position of that word in a multidimensional vector space. Representing each word as word vector permits to capture 'semantic' information (such as similarity and relatedness). As in Lauscher *et al.* (2017), we created domain-specific word-embeddings on Core,[4] a large collection of scientific dissertations, and then we represented each abstract with a single vector, computed by averaging the embeddings of each of its words.

4 Discipline Identification and Interdisciplinary Detection

Previous attempts to detect interdisciplinary practices from text have been based on two main assumptions: (1) the difference between a set of predefined academic disciplines can be automatically identified from text (by using, for example, LDA topic models) and (2) this knowledge will lead us to know which theses are interdisciplinary (intuitively, the ones that are the most difficult to classify as belonging to a single discipline).

In this article we decided to evaluate the correctness of both assumptions. We start by examining whether it is true that text mining methods can

detect the main discipline of an academic abstract; to accomplish this, Chuang *et al.* (2012) used the Rocchio classifier (Manning *et al.*, 2008) with LDA topic values as features. We therefore decided to adopt the same approach, and to compare its performance to the results obtained with other classifiers.

Given a set of vector-representation of abstracts and the related discipline-labels, the Rocchio classifier first creates a centroid for each discipline, which represents the center of mass of all the members. Next, when considering a new unlabeled abstract, it computes the distances between its vector representation and all the centroids (using—in our case—the cosine similarity): the closest one is returned as the most suitable label.

The Rocchio classifier could already provide good results; however, due to the fact that it generalizes each class to a single centroid, this method may have issues with classes that are broad and general, such as academic disciplines (for example, it represents all History dissertations with a single centroid). For this reason, we compare its performance with the *k*-nearest neighbors classifier (*k*-NN). This is an alternative classification method that, instead of computing a centroid, labels each new observation with the majority class of the *k* most similar labeled documents.

The third method we considered in this work is a support vector machine (SVM). SVMs are one the most adopted approaches for text-classification tasks (Joachims, 1998). In this model, examples are represented as points in a multidimensional feature-space. Learning the classification consists in finding hyper-planes that separate the points while maximizing the margin between the hyper-planes and the closest points. In this work, we train a series of binary classifiers that distinguishes between one of the labels and the rest (one-versus-all) and finally assign each thesis to the classifier with the highest confidence.

Each of the presented classifiers produces a ranking of disciplines for each abstract, from the most similar to the farthest away. Previous works (Chuang *et al.*, 2012) have used this knowledge, which we call here 'distance-based' information, to generate a graphical representation of the corpus that helped them distinguishing interdisciplinary and mono-disciplinary theses: the interdisciplinary ones should be in fact situated 'between' different disciplines, while monodisciplinarty ones should be closer to their main discipline representation. While these visualizations are useful to explore the corpus, to quantitatively assess whether it is true that the obtained distances distinguish interdisciplinary and monodisciplinary works, a second classification step is necessary. Therefore, we feed a second classifier with the produced distance-based information, and we train it to decide if an abstract is interdisciplinary or not.

To understand whether this notion of distance is essential for recognizing interdisciplinary research, we compare the performance of such classifier with the results produced by a second classifier that simply employs features directly extracted from text (e.g. TF-IDF). The assumption in this case was that to distinguish interdisciplinary and mono-disciplinary theses is sufficient to examine the differences in their language, without any notion of 'distance' between disciplines. For these final experiments, we use in both cases a SVM classifier.

5 Experiments

As a first step, we compare the performance of the above introduced feature-vector representations and classification algorithms to assess their reliability on recognizing the main discipline of a dissertation. We evaluate them with 10-fold cross-validation, a common practice in NLP, where the original data set is randomly divided in ten equal-sized subsamples. For the Rocchio classifier we use nine parts as a training corpus and one part as a test set. For *k*-NN and SVM, we use eight parts as a training corpus, one part for parameters tuning (in *k*-NN the number of k, in SVM the parameter C[5]), and one part for testing. In all experiments we use a SVM with linear kernel. When using LDA for generating topic-values, we tested values for the parameter *k* (the number of topics) in range 50–1,000. We report the performance of the methods that consistently performed best in the experiments.

5.1 Predicting the main discipline

The assumption behind previous works on the topic is that identifying the main discipline of

interdisciplinary theses will be more difficult compared to mono-disciplinary theses. This would, in fact, confirm the fact that interdisciplinary theses borrow words from different disciplines, while mono-disciplinary dissertations use a more 'defined' language. For this reason, in Fig. 3 we reported the results (on the main-discipline classification tank - MDC), by considering both the performance of the methods over the entire corpus and only on the subset of 93 interdisciplinary and 179 mono-disciplinary theses. As can be seen in Fig. 3, the classification quality on mono-disciplinary theses is much higher than on interdisciplinary theses, which therefore confirms this starting conjecture. However, an important finding of this experiment is that using lexical features (i.e. TF-IDF weighted term-vectors) within a SVM or a Rocchio classifier consistently outperformed other models, and in particular the results of LDA topic models. Additional experiments on this task could be found in Nanni *et al.* (2016).

5.2 Predicting secondary disciplines

Each of the ninety-five interdisciplinary dissertations presented in our gold standard is associated with a set of secondary disciplines suggested by the thesis supervisor. We employed this information to examine the correctness of the list of secondary disciplines that each classifier produces as an output. In Fig. 3 we report the mean average precision (MAP) of the rankings produced by each classifier (secondary discipline ranking (SDR)). Once again, the rankings produced by the SVM and the Rocchio classifier using TF-IDF representation of the abstracts were better than the other classifiers. This indicates that often the second discipline associated with a dissertation represented using LDA topic values as features is not the correct one.

	MDC			SDR
	All	Mono-Disc	Int-Disc	Int-Disc
Rocchio TF-IDF	0.71^{-}	0.84	0.62	**0.55**
Rocchio LDA	0.62^{-}	0.71^{-}	0.53^{-}	0.43^{-}
k-NN TF-IDF	0.67^{-}	0.69^{-}	0.57^{-}	0.27^{-}
SVM TF-IDF	**0.75**	**0.87**	**0.68**	**0.55**
SVM w. Emb.	0.68^{-}	0.8^{-}	0.62	0.52

Fig. 3 Results on discipline classification (main: F1-Score, secondary: MAP). Methods over which SVM TF-IDF achieves significant improvements are marked with –

5.3 Detecting interdisciplinary dissertations

While these findings are relevant from a text classification perspective, the real takeaway of this work is understanding whether the induced distances between disciplines are signals for distinguishing interdisciplinary and monodisciplinary dissertations. In Fig. 4 we report the performance (F1 Score) of the second classifier. As can be seen, the results are in contrast with what previously suggested in the literature and show that distant-based features are generally not informative enough for correctly distinguishing interdisciplinary from monodisciplinary theses. On the opposite, the performance of the SVM using only textual information is evidently more robust. For better understanding how this 'lexicalized' classifier performs the task, we examined the features that appeared to be the most relevant for distinguishing between the two classes. Here we noticed that what really characterize interdisciplinary dissertations is the language used to present their research, where words focused on methodology ('research', 'approach', 'technology', 'method') or adopted for describing research project with a wide scope ('pain', 'population', 'environment') are prominent.

In addition to these 'technical' outcomes, a second important finding of this experiment is that word-features are consistently outperforming topic-features from LDA. The impact of this finding goes beyond this specific work and should be relevant to any other type of research that employ text mining methods for generating quantitative evidence. The reliability of a computational approach should never be assumed in advance; on the contrary, it should be tested and proven, and its error

	Interdisciplinary detection
Rocchio TF-IDF	0.51
Rocchio LDA	0.56
k-NN TF-IDF	0.46
SVM TF-IDF	0.44
SVM w. Emb.	0.35
SVM (Textual Feat.)	**0.74**

Fig. 4 Performance on the interdisciplinary detection task (F1-Score). Underlined method/features are significantly better than all others

modes have to be clearly understood before adopting it and interpreting its results for a research project (as remarked in Traub and van Ossenbruggen, 2015; Nanni *et al.*, 2016).

6 Conclusion: Enhancing the Study of the Computational Turn

The results presented in the quantitative evaluation highlight that detecting interdisciplinary research from textual content is more complex that what previously remarked in the literature. As a matter of fact, we have shown that the use of a mixture of discipline-specific words is not always a direct signal of interdisciplinary collaborations and that other components play an essential role, such the overall research question or the topic of the work (e.g. studying a cross-discipline topic such as 'global warming'). We have also remarked on the fact that is type of signal is better captured by directly extracting features from text, instead of generating distances between discipline-centroids. Based on this knowledge, we present in the final part of this article a new data set we have enriched with discipline information, to allow historians of higher education to study the topic of interdisciplinary research directly from text.

The online portal DART-Europe (Digital Access to Research Theses-Europe), a partnership of research libraries and library consortia who are working together to improve global access to European research theses, offers over 700,000 theses from 28 European countries and 596 universities. While this corpus provides an unprecedented amount of primary sources for historians interested in the changes in research practices in academia, the available collection does not consistently offer metadata regarding the discipline of each thesis. As a consequence, this reduces the navigation of the corpus and does not allow diachronic and discipline-based comparative study (such as examining the changes in biological research across the past 30 years).

To allow this kind of research, we have collected from DART-Europe around 200,000 doctoral theses published between 1980 and 2015, which provide an abstract in English. Next, we have classified each thesis using the previously described SVM (with textual features); the classifier was trained on a subset of the data set which offers information regarding the main discipline (all theses from Italian universities with an abstract in English, for 11,726).[6]

To show the usefulness of this new resource, we have used it to track a specific interdisciplinary-related phenomenon, namely, the computational turn (Berry, 2011). To do so, as a first step we have examined the main disciplines of each thesis: if 'Computer Science' (or 'Computer Engineering') appeared to be one of the top two disciplines detected, we have considered the thesis as having a 'computational' aspect (which, as we have already remarked, does not necessarily imply that the thesis is 'interdisciplinary'). To better understand the type of cross-discipline collaboration, we have additionally examined the language used in its abstract highlighting words that emphasize interdisciplinary concepts, such as novelty, collaboration, and method-oriented research.

In Fig. 5, results of our study are presented. We have grouped disciplines together considering the European Research Council (ERC) domains: Physical Sciences and Engineering, Life Sciences, Social Sciences, and Humanities, to offer a macro-overview of the analysis. As can be noticed, the 'computational turn' is detected, but it strongly differs across different macro-areas; for instance Physical Sciences and Engineering has an average of 17% computational theses, while Life Sciences 8%, Social Sciences 9%, and Humanities 6%. Moreover, the time-trend reveals that, for instance, Physical Sciences and Engineering started the earliest, they experienced a quick and steep growth between the 80s and the 90s, and has been relatively stable in the past 10 years, while the Humanities, started in later years and still present a very unstable profile, with less clear growth trends. More analyses on this dataset could be found in Nanni and Paci (2017).

Starting from these analyses, we envision historians of higher education moving beyond our work and employ the presented large-scale collection for digging deeper into the driving forces of such computational turn across different European countries, academic environments, and academic disciplines, ranging from Biology to Neuroscience to, of course, History.

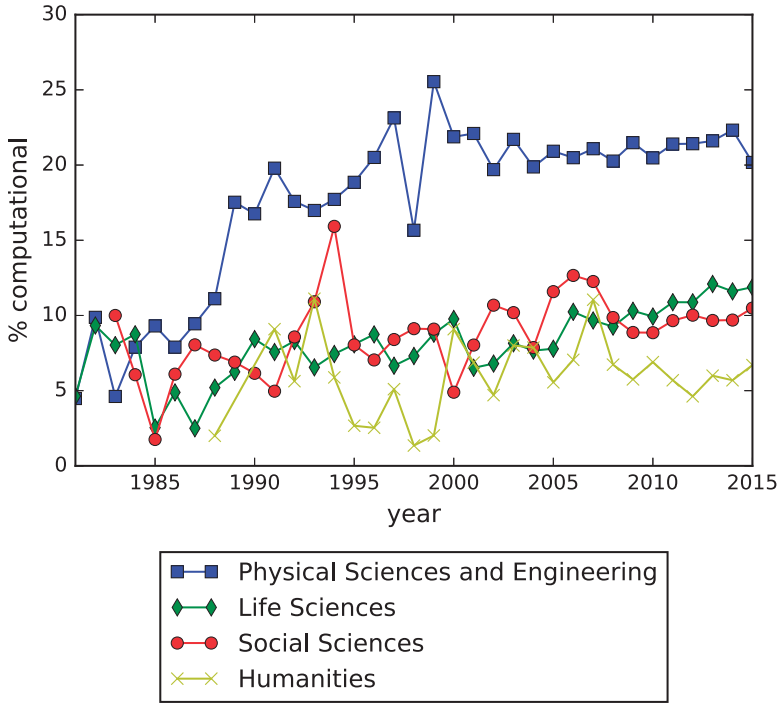

Fig. 5 Per-discipline growth of theses with a computational aspect, between years 1985 and 2015

Funding

This work was conducted at the University of Mannheim—Data and Web Science Group, B6, 26, D-68159, Mannheim, Germany.

References

Allmendinger, J. (2015). Quests for interdisciplinarity: a challenge for the ERA and HORIZON 2020. In *Policy Brief by the Research, Innovation, and Science Policy Experts (RISE)*. Brussels: European Commission. https://ec.europa.eu/research/openvision/pdf/rise/allmendinger-interdisciplinarity.pdf.

Berry, D. (2011). The computational turn: thinking about the digital humanities. *Culture Machine*, **12**: 1–22.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**: 993–1022.

Ballo, E. and Franchella, M. (2006). *Logic and Philosophy in Italy: Some Trends and Perspectives: Essays in Honor of Corrado Mangione on his 75th Birthday*. Monza, Milan, Italy: Polimetrica.

Chuang, J., Ramage, D., Manning, S. D., and Heer, J. (2012). Interpretation and trust: designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, May 5–10, 2012, Austin, Texas, USA, pp. 443–52.

Dietz, L., Bickel, S., and Scheffer, T. (2007). Unsupervised prediction of citation influences. In *Proceedings of the ICML*, June 20–24, 2007, Corvallis, OR, USA, pp. 233–40.

Gerrish, S. and Blei, D. M. (2010). A language-based approach to measuring scholarly impact. In *Proceedings of the ICML*, 21–24 June 2010, Haifa, Israel, pp. 375–82.

Hall, D., Jurafsky, D., and Manning, C. D. (2008). Studying the history of ideas using topic models. In *Proceedings of ACL*, June 25–20 2008, Columbus Ohio.

Holm, P., Goodsite, M. E., Cloetingh, S., Agnoletti, M., Moldan, B., Lang, D. J., Leemans, R., Moeller, J. O., and Zondervan, R. (2013). Collaboration between the natural, social and human sciences in global change research. *Environmental Science and Policy*, **28**: 25–35.

Joachims, T. (1998). *Text Categorization with Support Vector Machines: Learning with Many Relevant Features*. Heidelberg, Germany: Springer.

Lauscher, A., Glavaš, G., Ponzetto, S. P., and Eckert, K. (2017). Investigating convolutional networks and domain-specific embeddings for semantic classification of citations. In *Proceedings of WOSP 2017*. New York: ACM.

Lu, K. and Wolfram, D. (2012). Measuring author research relatedness: a comparison of word-based, topic-based, and author cocitation approaches. *Journal of the American Society for Information Science and Technology*, **63**(10): 1973–86.

Manning, C., Schütze, H., and Raghavan, P. (2008). *Introduction to Information Retrieval*. Cambridge: Cambridge University Press.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. New York: Curran Associates, Inc. pp. 3111–19.

Nanni, F., Dietz, L., Faralli, S., Glavaš, G., and Ponzetto, S. P. (2016). Capturing interdisciplinarity in academic abstracts. *D-lib magazine*, **22**(9/10).

Nanni, F. and Paci, G. (2017). A discipline-enriched dataset for tracking the computational turn of European universities. In *Proceedings of WOSP 2017*, ACM, New York.

Nanni, F., Kümper, H., and Ponzetto, S. P. (2016). Semi-supervised textual analysis and historical research helping each other: some thoughts and observations. In *International Journal of Humanities and Arts Computing*. United Kingdom: Edinburgh University Press, pp. 63–77.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the national science foundation. *Scientometrics*, **100**(3): 741–54.

Orlandi, T. and Mordenti, R. (2003). Lo status accademico dell'Informatica umanistica, con Appendice di M. Catacchio. *Archeologia e Calcolatori*, **14**, 7–32.

Pascuzzi, G. (2014). Soldatini e danni collaterali: i settori scientifico-disciplinari. ROARS.

Rafols, I. and Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, **82**(2): 263–87.

Ramage, D., Manning, C. D., and Dumais, S.T. (2011). Partially labeled topic models for interpretable text mining. In *Proceedings of SIGKDD*, August 21–24, 2011, San Diego, California, USA, pp. 457–65.

Repko, A. F. (2008). *Interdisciplinary Research: Process and Theory*. Thousand Oaks, California: Sage.

Sugimoto, C. and Weingart, S. (2015). The kaleidoscope of disciplinarity. *Journal of Documentation*, **71**(4): 775–94.

Traub, M. C. and van Ossenbruggen, J. (2015). Workshop on tool criticism in the DH. In *Workshop on Tool Criticism in the DH*, p. 7. https://pdfs.semanticscholar.org/d337/ce558c2fd1d8be793786c9cfc3fab6512dea.pdf.

Turkel, W. J. (2008). *Towards a Computational History*. http://digitalhistoryhacks.blogspot.de/2008/07/towards-computational-history.html (accessed 25 January 2016).

Van Raan, A. (1997). Scientometrics: State-of-the-art. *Scientometrics*, **38**(1): 205–18.

Wagner, C. S., Roessner, J. D., Bobb, K., Klein, J. T., Boyack, K. W., Keyton, J., Rafols, I., and Börner, K. (2011). Approaches to understanding and measuring interdisciplinary scientific research (IDR): a review of the literature. *Journal of Informetrics*, **5**(1): 14–26.

Notes

1 As, for example, on DART-Europe (http://www.dart-europe.eu/).
2 http://almadl.unibo.it/
3 Chuang *et al.* (2012) also employ Labeled LDA. This method needs a multi-labeled corpus, where each dissertation is identified with more than one discipline-label. This does not apply to our data set.
4 https://core.ac.uk/
5 $C = \{0.0001, 0.001, 0.1, 1, 10, 100, 1{,}000, 10{,}000\}$
6 We obtain a micro F1-Score of 0.72, which is consistent with the results presented for the task of main discipline classification on the University of Bologna small data set. The data set is available here: https://madata.bib.uni-mannheim.de/259/.