D-Lib Magazine

September/October 2016
Volume 22, Number 9/10

Capturing Interdisciplinarity in Academic Abstracts

Federico Nanni
Data and Web Science Research Group, University of Mannheim, Germany
International Centre for the History of Universities and Science, University of Bologna, Italy
federico@informatik.uni-mannheim.de

Laura Dietz, Stefano Faralli, Goran Glavaš and Simone Paolo Ponzetto
Data and Web Science Research Group University of Mannheim, Germany
{dietz,stefano,goran,simone}@informatik.uni-mannheim.de

DOI: 10.1045/september2016-nanni

Abstract

In this work we investigate the effectiveness of different text mining methods for the task of automated identification of interdisciplinary doctoral dissertations, considering solely the content of their abstracts. In contrast to previous attempts, we frame the interdisciplinarity detection as a two step classification process: we first predict the main discipline of the dissertation using a supervised multi-class classifier and then exploit the distribution of prediction confidences of the first classifier as input for the binary classification of interdisciplinarity. For both supervised classification models we experiment with several different sets of features ranging from standard lexical features such as TF-IDF weighted vectors over topic modelling distributions to latent semantic textual representations known as word embeddings. In contrast to previous findings, our experimental results suggest that interdisciplinarity is better detected when directly using textual features than when inferring from the results of main discipline classification.

Keywords: Interdisciplinarity; Text Classification; Scientometrics; Tool Criticism

1 Introduction

In recent years, both governments and funding agencies have been trying to encourage more interdisciplinary research than ever before [7]. As recently remarked in a special issue of Nature dedicated to the topic, researchers are attempting greater and greater challenges combining techniques from multiple disciplines, blurring the traditional borders between academic areas.

While there is a large body of work in scientometrics focusing on new solutions for identifying, quantifying, and visualising the diffusion of interdisciplinary research in large collections of scientific works [13], most of these methods exploit bibliometric measures such as co-author networks and citation graphs. The downside of such techniques is that they obviously depend on the availability of the bibliometric network data, which can often be difficult to obtain. In fact, it is generally much easier to obtain the full-text of the publications (as through DART-Europe) than the bibliometrics network data, especially for comparisons across different research areas.

In this article we focus on detecting interdisciplinarity in a setting where bibliometric network information is not available. The few previous attempts towards content-based interdisciplinarity detection mostly relied on unsupervised machine learning methods. In particular among them, topic modeling has been adopted to detect mixtures of discipline-specific terminologies [14, 11], often in combination with partially supervised methods [2, 6]. Having noticed a general lack of gold standard and quantitative evaluation for the task of interdisciplinarity detection, this study aims to fill this gap and provide answers to the following research questions:

- How well can the main discipline of scientific publications be predicted using supervised models with content-based features?
- In the case of interdisciplinary research, can similar techniques be used to robustly detect the secondary discipline(s)?
- Do prediction confidences of the main discipline classifier contribute to the detection of interdisciplinary research (i.e., do they improve the performance of the second-step binary interdisciplinarity classifier)?

Another important aspect of this work is the examination of different text-based features, both lexical (TF-IDF-weighted term-vectors) and semantic (topic model distributions, word embeddings), in both classification tasks.

We provide an overview of the related work in Section 2. Section 3 covers the classification pipeline used, which consists of two supervised steps: (1) discipline classification and (2) interdisciplinarity detection. Section 4 details the evaluation setting, gold standard creation, and provides quantitative evidence that the proposed supervised pipeline can be used to accurately predict the main discipline, the set of secondary disciplines, and also to identify interdisciplinary dissertations from their abstracts. Finally, we conclude in Section 5 with a summary of the main findings.

2 Related Work

In the field of scientometrics [16], the use of bibliometrics approaches and network analysis techniques is the most common methodology for evaluating and quantitatively describing the scientific output [10]. In particular, for measuring interdisciplinarity, researchers usually opt for numerous variants of citation and co-citation analyses [17]. As already remarked [13], the percentage of citations outside the main discipline of the citing paper is the most useful indicator of interdisciplinarity.

More recently, researchers started exploiting textual content of the publications in scientometric studies and several methods using lexical and topic-based information were proposed. Dietz *et al*. [4] extend Latent Dirichlet allocation (LDA) [1] to quantify the impact that research papers have on each other, whereas Gerrish *et al*. [5] extend LDA to model authoritative publications. Furthermore, Lu *et al*. [10] compare LDA to co-citation methods for measuring the relatedness between authors and their research and remark that the topic modelling approach produces a more complete map of these relations.

Concerning, in particular, the automatic detection of interdisciplinarity, several studies observed the usefulness of topic modelling for the task [14, 2, 11]. One of the few supervised approaches towards interdisciplinarity detection from text is proposed by Giannakopoulos *et al*. [6]. In that paper the researchers adopt a Naive Bayes classifier to establish the probability of each publication belonging to a specific discipline. The obtained results are used to visualise "distances" between academic fields, and are interpreted by the authors as a sign of interdisciplinary practices.

The approach most related to ours is the one described in Chuang *et al*. [2]. The authors employ the Rocchio Classifier [8] and, as in Giannakopoulos *et al*. [6], derived distances between dissertation abstracts and representations of disciplines with the goal of identifying and visualising interdisciplinary research in the Stanford corpus of doctoral dissertations. These distances are obtained by two different vector-based representations of dissertations abstracts: (1) term-vectors, typically weighted according to the term frequency-inverse document frequency (TF-IDF) scheme [15] and (2) vectors capturing the distributions of latent topics in the abstract, obtained using the topic models. It has been observed that different representations of abstracts yield outputs that "seem plausible, but each makes different predictions" [2].

Effects of different design decisions such as the choice of the vector representation of the abstracts or the choice of the learning algorithm have not yet been systematically explored in a quantitative evaluation setup. The main contribution of this work is a quantitative evaluation of how different features sets and learning algorithms perform on the interdisciplinarity detection task.

3 Methodology

Following Chuang *et al*. [2], we study a two-step pipeline where first the abstract is classified into one of many research disciplines, then the outputs are analysed to predict a binary interdisciplinarity attribute. The full pipeline of the proposed approach is depicted in Figure 1.

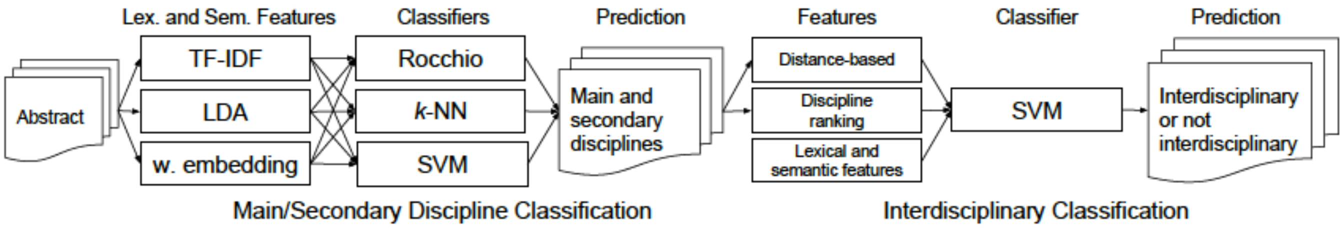
Lex. and Sem. Features
Classifiers
Prediction
Features
Classifier
Prediction
Abstract
TF-IDF
LDA
w. embedding
Rocchio
k-NN
SVM
Main and secondary disciplines
Distance-based
Discipline ranking
Lexical and semantic features
SVM
Interdisciplinary or not interdisciplinary
Main/Secondary Discipline Classification
Interdisciplinary Classification

Figure 1: Graphic representation of the adopted pipeline for interdisciplinarity detection

3.1 Main/Secondary Discipline Classification

Our first task is to classify dissertation abstracts according to their main (primary) discipline. To this end, we experiment with several different classifiers for which we use the following set of lexical and semantic features:

1. The TF-IDF-weighted term vector of the dissertation abstract, computed over the corpus of available dissertation abstracts (cf. Section 4.1);
2. The topical distribution (i.e., a topic-probability vector) of the abstract. The topic model was trained over the corpus of available dissertation abstracts;
3. The semantic embedding of the dissertation abstract, which is computed as the element-wise average of the embeddings of the words the abstract contains. Let *A* be the set of unique words in the dissertation abstract. The embedding of the abstract (V_a) is then computed as follows:

$$v_a = \frac{1}{N}\sum_{w\in A}\text{freq}(w)\cdot v_w$$

   where freq(*w*) is the frequency with which word *w* occurs in the abstract, *vw* is the embedding vector of the word *w*, and *N* is the total number of words in the dissertation abstract. We used the state-of-the-art GloVe word embeddings [12] for the computation of the abstract embeddings. We experimented both with the pre-trained GloVe embeddings (trained on the merge of Wikipedia and Gigaword corpus) and word embeddings trained by us on the CORE Corpus of scientific publications.

Using a *k*-fold cross-validation setup, we experiment with three different supervised classification models using the above described set of features: (1) Rocchio classifier [8], (2) *k*-Nearest Neighbors (*k*-NN) [3], and (3) Support Vector Machines (SVM) [9]. All these supervised classifiers provide confidence values based on the distance towards a class centroid or a separating hyperplane. In a multi-class classification setting, all three classifiers offer a mechanism for ranking the discipline classes for every given dissertation abstract instance: SVM and Rocchio directly provide confidence scores for each class, and for *k*-NN we rank the classes according to the number of neighbours (out of the total *k*) labeled with a particular class. We then use these class rankings, implicitly produced by the classifiers when predicting the main discipline of an abstract, as a method for capturing the secondary disciplines.

3.2 Interdisciplinarity Classification

We study feature sets for binary interdisciplinarity detection.

Original. The feature vectors used for multi-class discipline classification can also be used in this secondary classifier (trained with different truth labels). These are TF-IDF term vectors, topic distributions and semantic abstract embeddings. These features capture general language patterns of abstracts, without any further discipline information.

Distance-based features. These features directly encode the confidences of the main discipline classifier from the first step. Given a set of *n* disciplines, this will be a numerical vector of length *n* where an element at position *m* of the vector is the classifier's confidence that the *m*-th discipline is the main discipline of the abstract. The rationale behind this feature is that a more uniform confidence distribution across disciplines (when classifying the main discipline) is a stronger indicator of an interdisciplinary dissertation, whereas the very skewed confidence distributions should be more common for the monodisciplinary dissertations.

Discipline-ranking features. In some cases, the ranking of disciplines induced by confidence scores (i.e., primary discipline versus secondary discipline) might be more informative than the confidence values. Furthermore, some main disciplines might be more appropriate for interdisciplinary research (e.g., biology) than others (e.g., history). This is encoded in a numerical feature vector of length *n* X *n*, encoding the matrix of 28 disciplines across 28 possible ranks, where within the row of rank *r*, we place a 1 in the column corresponding to discipline *m* on that rank.

We train a binary SVM classifier with the above-described features using ground truth annotations in a *k*-fold cross-validation setting.

4 Evaluation

We first describe the dataset of dissertation abstracts used to evaluate our supervised models, after which the experimental setup is covered. Finally, we discuss the classification performance of discipline classification and binary interdisciplinarity detection of different model variants.

4.1 Dataset

We use a corpus of PhD dissertations for evaluation, as these collections reflect the greatest common denominator of academic output available across all disciplines, research schools, and countries. The digital library of the University of Bologna provides us with a collection of a total of 2,954 dissertations with an English abstract. The mean length of these abstracts is 320 tokens. Each dissertation was first assigned one main discipline from the list of 28 disciplines defined by the Italian Ministry of Education, University, and Research (see Table 1). As expected, the distribution of dissertations over disciplines is heavily skewed (ranging anywhere from 322 for medicine to 7 for oriental studies), reflecting the natural order in which some disciplines are more "popular" than others. Additional information on the dataset is available here.

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

Table 1: The total number of abstracts for each discipline in our dataset (All) and the number of interdisciplinary (Int-Disc) and monodisciplinary (Mono-Disc) theses in our gold standard.

The following example snippets of dissertation abstracts from three disciplines demonstrate the highly technical content of these abstracts.

> **Medicine:** *IL-33/ST2 axis is known to promote Th2 immune responses and has been linked to several autoimmune and inflammatory disorders, including inflammatory bowel disease (IBD), and evidences show that it can regulate eosinophils (EOS) infiltration and function.*
>
> **Computer Science:** *We have tried for the first time to explore the relation among mutations at the protein level and their relevance to diseases with a large-scale computational study of the data from different databases.*
>
> **History:** *The present study aims at assessing the innovation strategies adopted within a regional economic system, the Italian region Emilia-Romagna, as it faced the challenges of a changing international scenario.*

In the next step, we enrich the dataset with (1) the ground truth annotations of secondary disciplines and (2) a binary assessment of whether a thesis is interdisciplinary or not. The highly technical nature of the abstracts' content requires expert annotators in different academic fields. In a survey, we asked supervisors of the dissertations to provide assessments for each of their supervised dissertations whether they consider it to be interdisciplinary and if so, to list the secondary disciplines of the dissertation. This resulted in a collection of 272 dissertations which are manually annotated for interdisciplinarity (93 interdisciplinary and 179 monodisciplinary dissertations).

In order to estimate the difficulty of detecting interdisciplinarity for humans, we asked two other expert annotators to re-annotate the 272 dissertations according to their expertise; one for natural sciences and a second for technical and social sciences. With respect to interdisciplinarity annotations, the annotators agree with the dissertation supervisors in 82% of the cases, with the interannotator agreement in terms of Cohen's Kappa score being 0.63 (which is considered to be substantial agreement). In cases where an annotator disagreed with a supervisor, we kept the annotations assigned by the dissertation supervisor.

4.2 Experiments

We evaluate the three parts of our pipeline: Main disciplinary classification, Secondary discipline classification, and Interdisciplinarity detection. Unless indicated otherwise, we use a linear SVM (i.e., no non-linear kernel), with model hyperparameters optimised on a hold-out tuning set and evaluated the performance using 10-fold cross-validation. For SVM, best hyperparameter C=1 in all experiments. For *k*-NN, optimal *k* varied in the range 30-250. When using LDA, we tested values for the parameter *k* (the number of topics) in range 50-1,000. Best performance with 250 topics.

Main discipline classification

In Table 2 we report (1) the classification results for the main discipline classification task (MDC) in terms of micro-averaged F_1 score and (2) the ranking performance for the assignment of secondary disciplines (SDR) measured in terms of mean-average precision (MAP) against the manually assigned secondary disciplines.

	MDC			**SDR**
	All	**Mono-Disc**	**Int-Disc**	**Int-Disc**
Random	0.04^{-}	0.04^{-}	0.04^{-}	0.10^{-}
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
SVM All feats.	**0.75**	**0.87**	**0.68**	**0.55**

Table 2: Results on discipline classification (main: F1, secondary: MAP). Methods over which SVM TF-IDF achieves significant improvements (std error test) are marked with $^{-}$.

In order to validate the underlying assumption that predicting the main discipline for interdisciplinary dissertations is more difficult than for monodisciplinary dissertations, we separately measure the performance of the models on the subset of monodisciplinary dissertations ("Mono-Disc") from the performance on the interdisciplinary dissertations ("Int-Disc"). As expected, the performance on the monodisciplinary subset is significantly higher than the performance on the interdisciplinary subset, which only confirms our assumption that it is more difficult to detect the main discipline for interdisciplinary publications. Using lexical features (i.e., TF-IDF weighted term-vectors) within a SVM or Rocchio classifier consistently outperformed other models. An unusual finding is that incorporating semantic features like topic distributions and abstract embeddings does not significantly improve the performance of any of the classifiers.

Secondary discipline classification

The quality of the ranking of the secondary disciplines was evaluated only on the subset of the corpus consisting of 93 interdisciplinary dissertations ("Int-Disc"). The trends in the results are consistent with experiments on main discipline classification: SVM and Rocchio classifier based only on lexical features outperform other combinations of classifiers and features.

Interdisciplinarity detection

The performance of different models on the interdisciplinarity detection task, in terms of micro-averaged F_1 score, is shown in Table 3. In all cases, we used a binary SVM for the classification, but varied different features sets in accordance with methods discussed for discipline classification (denoted in the row headers of Table 3) as follows. The first column (*orig*) presents results of the SVM model for interdisciplinarity classification that uses the same set of features as the main discipline classifier given by the corresponding table row (therefore row 2,4,5 are identical w.r.t. the orig column). The remaining columns present performance achieved when using output of the discipline classifier as input features: (1) only distance-based features (*dt*), (2) only discipline-ranking features (*dr*), (3) distance-based and discipline-ranking features (*dt* + *dr*), and (4) distance-based and discipline-ranking features together with the original features used by the main discipline classifier (*all* = *orig* + *dt* + *dr*).

	Interdisciplinary detection				
	orig	**dt**	**dr**	**dt + dr**	**all**
Random	0.04	0.04	0.04	0.04	0.04
Rocchio TF-IDF	0.52	0.51	0.54	**0.56**	0.57
Rocchio LDA	0.52	**0.56**	0.45	0.53	0.59
k-NN TF-IDF	0.52	0.46	0.52	0.42	0.58
SVM TF-IDF	0.52	0.44	0.49	0.47	0.56
SVM w. Emb.	0.60	0.35	0.37	0.36	0.60
SVM All feats.	**0.74**	**0.56**	**0.55**	0.55	**0.70**

Table 3: Performance on the interdisciplinarity classification task (F1). Underlined method/features is significantly better than all other methods.

Our results clearly show that inferring interdisciplinarity from the results of the main discipline classifier (i.e., using distance-based and discipline-ranking features), as suggested in literature [2], cannot be done robustly. Using the prediction confidences of the main discipline classifier for interdisciplinarity classification (the scores in the *dt*, *dr*, and *dt* + *dr* columns of Table 3) yields worse performance than using directly the same lexical and semantic features (*orig*) that were used for the main discipline classifier. The best interdisciplinarity classification performance is achieved by the SVM model directly using lexical and semantic features (TF-IDF term vector, topic distribution, and abstract embedding), which is not exploiting in any way the confidence scores produced by the main discipline classifier.

These results disprove the assumption from previous work [14, 2] that a notion of "distance" between publications and representation of disciplines are the best way for detecting interdisciplinarity. Our experiments reveal that the distinction between interdisciplinary and monodisciplinary practices can be better explained by the differences in the language used than by the distances to disciplines' classification centroids or hyperplanes. This might indicate that, instead of the ranking of involved disciplines, what really distinguishes interdisciplinary dissertations is their focus and the language in which they present and describe their research.

5 Conclusion

In this paper we present a supervised pipeline for the main discipline classification and interdisciplinary detection in academic thesis abstracts. Striving for a tool applicable in practise, we focus on a setting where only the textual content of the abstract is considered (i.e., where no scientometric network data is available). Following the insights from related studies [2], we frame the pipeline as a sequence of two supervised classification tasks: (1) classification of the main discipline of the publications, and (2) binary interdisciplinarity detection, with features based on results of the main discipline classification as (additional) input. In both classification steps, the study included several sets of lexicalised and semantic features such as term-vectors, topic distributions, and semantic word embeddings.

The results of our experiments on the corpus of doctoral dissertation abstracts show that both robust identification of the main discipline and ranking of the secondary disciplines can be achieved using a simple (and computationally inexpensive) classification model (e.g., Rocchio classifier with only TF-IDF weighted term-vectors). Regarding the interdisciplinarity classification, results of our experiments do not indicate that discipline classification is needed for interdisciplinarity detection, which is in contrast to findings of previous work. On the contrary, we show that the best performance on the inderdisciplinarity detection task is achieved using the set of lexical and semantic features derived directly from the text. We speculate that the lexicalised classifier learns from cue words such as "interdisciplinary", "innovative" and "collaboration" as well as discipline-specific words together.

Acknowledgements

This work was funded in part by a scholarship of the Eliteprogramm for Postdocs of the Baden-Württemberg Stiftung (project "Knowledge Consolidation and Organization for Query-specific Wikipedia Construction"). Furthermore, this work was partially funded by the Junior-professor funding programme of the Ministry of Science, Research and the Arts of the state of Baden-Württemberg (project "Deep semantic models for high-end NLP application"). We thank the digital library of the University of Bologna for providing the dataset of thesis abstracts and all our annotators.

References

[1] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3:993-1022, 2003.

[2] J. Chuang, D. Ramage, C. Manning, and J. Heer. Interpretation and trust: Designing model-driven visualizations for text analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 443-452. ACM, 2012. http://doi.org/10.1145/2207676.2207738

[3] T. M. Cover and P. E. Hart. Nearest neighbor pattern classification. *Information Theory, IEEE Transactions on*, 13(1):21-27, 1967. http://doi.org/10.1109/TIT.1967.1053964

[4] L. Dietz, S. Bickel, and T. Scheffer. Unsupervised prediction of citation influences. In *Proceedings of the 24th International Conference on Machine Learning*, pages 233-240. ACM, 2007. http://doi.org/10.1145/1273496.1273526

[5] S. Gerrish and D. M. Blei. A language-based approach to measuring scholarly impact. In *Proceedings of the 27th International Conference on Machine Learning*, pages 375-382, 2010.

[6] T. Giannakopoulos, I. Foufoulas, E. Stamatogiannakis, H. Dimitropoulos, N. Manola, and Y. Ioannidis. Discovering and visualizing interdisciplinary content classes in scientific publications. *D-Lib Magazine*, 20(11):4, 2014. http://doi.org/10.1045/november14-giannakopoulos

[7] P. Holm, M. E. Goodsite, S. Cloetingh, M. Agnoletti, B. Moldan, D. J. Lang, R. Leemans, J. O. Moeller, M. P. Buendía, and W. Pohl. Collaboration between the natural, social and human sciences in global change research. *Environmental Science & Policy*, 28:25-35, 2013. http://doi.org/10.1016/j.envsci.2012.11.010

[8] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. Technical Report, DTIC Document, 1996.

[9] T. Joachims. *Text Categorization with Support Vector Machines: Learning with many Relevant Features*. Springer, 1998.

[10] K. Lu and D. Wolfram. 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):1973-1986, 2012. http://doi.org/10.1002/asi.22628

[11] L. G. Nichols. A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics*, 100(3):741-754, 2014. http://doi.org/10.1007/s11192-014-1319-2

[12] J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Proceedings of EMNLP*, volume 14, pages 1532-1543, 2014.

[13] I. Rafols and M. Meyer. Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, 82(2):263-287, 2010. http://doi.org/10.1007/s11192-009-0041-y

[14] D. Ramage, C. D. Manning, and S. Dumais. Partially labeled topic models for interpretable text mining. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 457-465. ACM, 2011. http://doi.org/10.1145/2020408.2020481

[15] G. Salton and M. J. McGill. *Introduction to Modern Information Retrieval*. 1983.

[16] A. Van Raan. Scientometrics: State-of-the-art. *Scientometrics*, 38(1):205-218, 1997. http://doi.org/10.1007/BF02461131

[17] C. S. Wagner, J. D. Roessner, K. Bobb, J. T. Klein, K. W. Boyack, J. Keyton, I. Rafols, and K. Böörner. Approaches to understanding and measuring interdisciplinary scientific research (idr): A review of the literature. *Journal of Informetrics*, 5(1):14-26, 2011. http://doi.org/10.1016/j.joi.2010.06.004

About the Authors

Federico Nanni is a final year Ph.D. Student at the International Centre for the History of Universities and Science, University of Bologna, and a researcher at the Data and Web Science group, University of Mannheim. His research interests include considering web materials as primary sources for historical studies and adopting natural language processing solutions for supporting research in the history of academic institutions.

Laura Dietz is a professor at the University of New Hampshire, where she teaches Information Retrieval and Machine Learning. Before that she worked in the Data and Web Science group at the University of Mannheim, at the University of Massachusetts, and she obtained her Ph.D. from the Max Planck Institute for Informatics. Her scientific contributions range from information retrieval with knowledge base information to the prediction of influences in citation graphs with topic models.

Stefano Faralli is a postdoctoral researcher at the Data and Web Science group, University of Mannheim. His research interests include Word Sense Disambiguation and Ontology learning.

Goran Glavaš is a postdoctoral researcher at the Data and Web Science group of University of Mannheim. His research interests are in core natural language processing and cover several application areas, including (but not limited to): information extraction, information retrieval and semantic search, lexical and computational semantics, and text summarization and simplification.

Simone Paolo Ponzetto is a professor at the University of Mannheim, where he leads the Natural Language Processing group. His research interests include knowledge acquisition and ontology learning from text, computational semantics, and the application of computational methods and digital technologies to problems in the humanities as well as in the social and political sciences.

Return to Article