ISSI 2021

18th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2021

Virtual event

12–15 July 2021
KU Leuven, Belgium

PROCEEDINGS

2021

18th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2021

Virtual event

12–15 July 2021

KU Leuven, Belgium

PROCEEDINGS

18th INTERNATIONAL CONFERENCE ON SCIENTOMETRICS & INFORMETRICS

ISSI2021

12–15 July 2021
KU Leuven, Belgium

PROCEEDINGS

Editors
Wolfgang Glänzel, Sarah Heeffer, Pei-Shan Chi, Ronald Rousseau

ISBN 9789080328228 – ISSN 2175-1935
July 2021 Printed in Belgium

Sponsors and Funding

Diamond sponsor – Elsevier
Gold sponsor – Digital Science
Gold sponsor – RISIS
Silver sponsor – JDIS
Silver sponsor – MIT Press
Silver sponsor – Clarivate ISI
Bronze sponsor – Springer Nature
Bronze sponsor – FRONTIERS

Research Foundation – Flanders (FWO)

Organisers

Koenraad Debackere, KU Leuven (Belgium), *Chair*
Wolfgang Glänzel, KU Leuven (Belgium)
Bart Thijs, KU Leuven (Belgium)
Tim Engels, UAntwerp (Belgium)

Programme Committee: Chairs

Ronald Rousseau, KU Leuven (Belgium), *Chair*
Lin Zhang, Wuhan University (China)/KU Leuven (Belgium)
Henk F. Moed, Sapienza University of Rome (Italy)

Program Committee: Members

Kevin Boyack, SciTech Strategies Inc. (USA), *Chair Workshops and tutorials*
Claudia Gonzalez Brambila, ITAM (Mexico)
Julie Callaert, KU Leuven (Belgium)
Pei-Shan Chi, KU Leuven (Belgium), *Poster Session*
Cinzia Daraio, Sapienza University of Rome (Italy)
Jacqueline Leta, UFRJ (Brazil), *Poster Session*
Raf Guns, UAntwerp (Belgium)
Stefanie Haustein & Juan Pablo Alperin, Ottowa (Canada)
Sarah Heeffer, KU Leuven (Belgium)
Omwoyo Bosire Onyancha, University of South Africa (South Africa)
Walter Ysebaert, VUB (Belgium)
Lin Zhang, Wuhan University (China)/KU Leuven (Belgium)

Doctoral Forum Committee

Jens Peter Andersen, Aarhus University (Denmark), *Chair*
Guillaume Cabanac, University of Toulouse (France)
Vincent Larivière, Université de Montréal (Canada)
Andrea Scharnhorst, KNAW (Netherlands)
Cassidy Sugimoto, Indiana University (USA)
Alesia Zuccala, University of Copenhagen (Denmark)

Eugene Garfield Doctoral Dissertation Award 2021 Committee

Birger Larsen, University of Aalborg (Denmark), *Chair*
Elías Sanz Casado, Carlos III University of Madrid (Spain)
Gemma Derrick, Lancaster University (UK)
Nees Jan van Eck, Leiden University (Netherlands)

Committee for the ISSI Student Travel Award

Kevin Boyack, SciTech Strategies Inc. (USA), *Chair*
Aparna Basu, South Asian University (India)
Cinzia Daraio, Sapienza University of Rome (Italy)

Scientific Committee

Abramo Giovanni
Adams Jonathan
Adie Euan
Aguillo F. Isidro
Ahlgren Per
Ajiferuke Isola
Akbaritabar Aliakbar
Aman Valeria
Amez Lucy
Andersen Jens Peter
Archambault Eric
Arvanitis Rigas
Astrom Fredrik
Ausloos Marcel
Azagra-Caro Joaquín M.
Badin Luiza
Baiget Tomas
Ball Rafael
Ballester-Gonzalez Omar
Bartol Tomaz
Bauer Guntram
Benner Mats
Bhattacharya Sujit
Bianchi Federico
Bollen Johan
Bonaccorsi Andrea
Borner Katy
Bornmann Lutz
Boshoff Nelius
Bouabid Hamid
Bowman Timothy
Buehrer Susanne
Cabezas-Clavijo Álvaro
Calero-Medina Clara
Campbell David
Cañibano Carolina
Cao Cong
Carayol Nicolas
Casnici Niccolò
Catalano Giusepppe
Caviggioli Federico
Chang Ching-Chun
Chaurasia Neeraj
Chavarro Diego
Checchi Daniele
Chen Chaomei
Chen Dar-Zen
Chen Yue
Chincilla-Rodriguez Zaida
Ciarli Tommaso
Colavizza Giovanni
Confraria Hugo
Costas Rodrigo
Côté Grégoire
Criado Regino
D'Angelo Ciriaco Andrea
Dastidar Prabir G.
De Bellis Nicola
De Bordes Carole
De Moya Felix
Dehdari-rad Tahereh
Deinzer Gernot
Díaz-Faes Adrián A.
Ding Ying
Donner Paul
Fatchur Rochim Adian
Ferreira-Goncalves Márcia
Franssen Thomas
Frietsch Rainer
Furner Jonathan
Gallart Jordi Molas
Garcia Romero Antonio
Gauch Stephan
Getz Daphne
Ghiasi Gita
Glänzel Wolfgang
Gläser Jochen
Gok Abdullah
Gorraiz Juan
Gorry Philippe
Gregori Martina
Guevara Miguel
Gumpenberger Christian
Gzoyan Edita
Halevi Gali
Hammarfelt Björn
Haunschild Robin
Hinze Sybille
Holmberg Kim
Hörlesberger Marianne
Hornbostel Stefan
Hu Xiaojun
Huang Mu-Hsuan
Ingwersen Peter
Ivancheva Ludmila
Jana Siladitya
Jeng Wei
Jimenez-Contreras Evaristo
Jonkers Koen
Jovanovic Milos
Kajikawa Yuya
Katz Sylvan
Kelchtermans Stijn
Klavans Dick
Konkiel Stacy
Kousha Kayvan
Krauskopf Manuel
Kulczycki Emanuel
Kumaravel J P S
Lamers Wout
Laudel Grit
Laurens Patricia
Lawson Cornelia
Lemke Steffen
Leng Rhodri
Lepori Benedetto
Li Jiang
Licea Judith
Liu Yuxian
Lo Szu-Chia

López-Illescas Carmen
Luwel Marc
Maisano Domenico
Markusova Valentina
Martín-Martín Alberto
Maynard Diana
Mayr Philipp
Mêgnigbêto Eustache
Meho Lokman I.
Melkers Julia
Meoli Michele
Milojevic Stasa
Mongeon Philippe
Morillo Fernanda
Moskaleva Olga
Mouton Maria
Mugabushaka Alexis Michel
Mugnaini Rogerio
Must Ülle
Mutz Ruediger
Nane Tina
Nederhof Anton
Nieminen Pentti
Noyons Ed
Ochsner Michael
Olmeda-Gómez Carlos
Orozco Luis Antonio
Ortega José Luis
Pandiella Andres
Pellegrino Gabriele
Perianes-Rodríguez Antonio
Peritz Cheila, Bluma
Peset Fernanda
Peter Viola
Peters Isabella
Pinheiro Diogo
Pislyakov Vladimir
Pontille David
Porter Alan L.
Pouris Anastassios
Prathap Gangan
Prozesky Heidi
Qiu Junping
Raffo Julio D.
Rafols Ismael
Rao Ravichandra
Reale Emanuela
Rigby John
Rigolin Camila Carneiro Dias
Robinson Garcia Nicolas
Roche Ivana
Romanello Matteo
Ronda-Pupo Guillermo
Rons Nadine
Rotolo Daniele
Rousseau Ronald
Rousseau Sandra
Ruggiero John
Ruocco Giancarlo
Sachwald Frederique
Sahoo Bibhuti
Sainte-Marie Maxime
Salinas Daniel
Sangam Shivappa Lingappa
Sargsyan Shushanik
Scherngell Thomas
Schiebel Edgar
Schmidt Marion
Schmoch Ulrich
Schneider Jesper
Schubert András
Schubert Torben
Seeber Marci
Shapira Philip
Shen Zhesi
Shu Fei
Sivertsen Gunnar
Small Henry
Smit Nynke-Jo
Soós Sándor
Sorz Johannes
Srivastava Divya
Sterzi Valerio
Sun Yuan
Tang Li
Tesch Jakob
Thelwall Mike
Traag Vincent
Tunger Dirk
Van den Besselaar Peter
Van der Meulen Barend
Van Leeuwen Thed
Van Looy Bart
Van Raan Anthony
Vancauwenbergh Sadia
Vargas-Quesada Benjamin
Velden Theresa
Vinkler Peter
Visser Martijn
Wagner Caroline
Waltman Ludo
Wang Lili
Wang Qi
Wang Xianwen
Welch Eric
Winnink Jos
Wolfram Dietmar
Woolley Richard
Wouters Paul
Wroblewski Angela
Wu Jinshan
Wu Yishan
Yan Erjia
Yang Liying
Ye Fred Y.
Yegros Alfredo
Youtie Jan
Zahedi Zohreh
Zhao Dangzhi
Zhao Yajuan
Zhou Ping
Zirulia Lorenzo

PREFACE

The 18th International Society of Scientometrics and Informetrics Conference is held on 12–15 July 2021 and organized by KU Leuven in close collaboration with the university of Antwerp under the auspices of ISSI the International Society for Scientometrics and Informetrics.

Due to the constraints imposed by the COVID-19 pandemic situation and the resulting risks for planning and organising a full-fledged conference with physical presence of attendees, the organisers decided to go virtual this time. This decision warranted a smooth organisation without unpredictable events and unnecessary modifications and adjustments during the preparation process. However, organising a virtual event of this size has posed a true challenge for the organisers. We were very pleased that despite the various different time-zones in the world, we could organise part of the conference, notably all keynotes, invited talks and plenary sessions as well as the workshops and tutorials and, last not least, all ceremonies as live events. The special tracks and the usual "parallel sessions" had, however, to be organised using pre-recorded on-line video streams. Making a virtue of necessity, we can provide all talks and the discussion for three months after the conference and thus make the conference a hopefully sustainable event.

The goal of ISSI2021 is to provide an international open forum for scholars and practitioners in the domain of informetrics, bibliometrics, scientometrics, webmetrics and altmetrics to discuss new research directions, advanced methods and theories, and to highlight the best research in this area. In order to achieve this goal, we asked researchers worldwide to submit original research manuscripts, particularly full papers, research-in-progress papers or posters, to propose and organise tutorials and workshops, with a special emphasis on the future of this area and on its interdisciplinary links with other fields. We succeeded in attracting a sufficient number of contributions to organise two special tracks, one on bibliometric approaches to measure and evaluate interdisciplinary research (IDR) organised by KU Leuven and one on the bibliometrics of social sciences and humanities (SSH) organised by our partner at University Antwerp. University Antwerp also took the opportunity to introduce their project of publishing on a new handbook on "Research Assessment in the Social Sciences".

We would like to thank the three keynote speakers, who have accepted our invitation, namely Katy Börner from Indiana University in Bloomington (USA), Albert-László Barabási from Northeastern University in Boston (USA), and David Sweeney, Executive Chair of Research England (UK).

Despite the pandemic-induced constraints, we could accept about 230 contributions, roughly two thirds of which are full and research-in-progress papers and nearly one third is present in pre-recorded poster sessions.

The assignment of contributions to session topics was facilitated by the organisation as a virtual event since time constraints did not influence the schedule. We have chosen the following major themes.

- Research evaluation & bibliometrics in support of science policy
- Effects of research funding
- Patent analysis
- Individual-level bibliometrics
- Collaboration and mobility
- Domain studies and regional issues
- Open Science, Open Access and editorial impact
- Gender and equality studies
- Webmetrics, altmetrics and media impact of research
- Advanced methods in citation analysis
- Data sources and data processing
- Document and journal analysis
- Network analysis

Furthermore, six workshops deal with the tracing of epistemic change, with creation and application of models, cited reference analysis, national and institutional research assessment, and collaborative archive and data research environment. This is completed by two tutorials. The organisers look forward to the future publication of the outcomes of these workshops.

All accepted presentations and posters are incorporated in the conference proceedings. The first part comprises invited and full papers as well as research-in-progress papers, while the second part is devoted to posters.

At this place, we would like to express our thanks to all participants and contributors for their understanding for the constrains due to the special situation and their active support. Our thanks are due to the ISSI board for its trust and support. We also thank the Leuven Congress Office – PC for their enormous effort and dedicated work and MEEPLE for providing the online platform. In particular, we thank Liesbeth Michiels, Marie-Laure Bettens and Ann Moerenhout.

On behalf of the organisers and the programme committee of ISSI2021
Wolfgang Glänzel

INDEX OF PAPERS
(FULL PAPERS AND RESEARCH IN PROGRESS)

INDEX OF POSTERS

An appraisal of publication embedding techniques in the context of conventional bibliometric relatedness measures

Wout S. Lamers[1,*], Nees Jan van Eck[1] and Giovanni Colavizza[2]

[1] {w.s.lamers, ecknjpvan}@cwts.leidenuniv.nl
Leiden University, Centre for Science and Technology Studies (CWTS), Leiden (The Netherlands)

[2] g.colavizza@uva.nl
University of Amsterdam, Faculty of Humanities, Department Mediastudies, Amsterdam (The Netherlands)

Abstract
Modern natural language processing techniques have given rise to embedding techniques that can represent documents based on their content or context, and several papers have operationalized these to perform bibliometric tasks. The relationship between these embeddings and conventional citation based or title and abstract based mappings remains unclear. Contrary to citation-based or term-based relatedness, embedding-based relatedness is not immediately interpretable. We consider four embedding-derived publication relatedness measures, based on: 1) word2vec embeddings of citation labels, sentence embeddings using 2) BERT and 3) SciBERT, and 4) title and abstract embeddings using SPECTER, and compare them with conventional bibliometric publication relatedness measures derived from citation relations and title and abstract noun phrases. We show that there is stronger overlap between these embedding-derived relatedness measures and citation-based relatedness than with title and abstract noun phrase-based relatedness, and that embedding-derived relatedness measures outperform conventional techniques when used to cluster publications cited with the same citation intent.

Introduction
The notion of relatedness of scientific publications is at the heart of many open problems in bibliometrics. Publication clustering and mapping techniques rely on publication-to-publication relatedness measures, conventionally based on citation relations, textual similarities of titles and abstracts, or occasionally combinations of these. It should come as no surprise that different relatedness measures produce different results (Gläser et al., 2017), which makes the informed choice of appropriate relatedness measures an imperative aspect of many bibliometric research tasks.

Meanwhile, advances in natural language processing (NLP) and information retrieval have introduced a host of new text-based methods for establishing relatedness of not only documents to one another, but also documents to terms in the vocabulary of these documents. The introduction of word2vec (Mikolov, Chen, et al., 2013; Mikolov, Sutskever, et al., 2013) marked a substantial advance, allowing vector representations of words to be learned that not only allow for easy computation of semantic similarities, but also preserve—to some extent—the semantic and syntactic relationships between terms. Doc2vec (Le & Mikolov, 2014) adapted this method, allowing for the generation of similar vector representations for larger bodies of text. Further advances in the field since then such as the emergence of large pre-trained language models like BERT (Devlin et al., 2018) have only expanded the suite of options for establishing relatedness between documents.

The case for using NLP techniques in science mapping and clustering is compelling. At their core lies the distributional hypothesis, the assertion that terms with similar distributions have similar meaning, i.e., similar terms are used in similar context. What is citation, if not the deliberate act of putting previous work in the context of a new author's design? Hence, papers that are cited in similar context, or that introduce similar language, can be assumed to be related to each other, and embedding techniques can capture this relatedness. Another important benefit of these embedding methods is that publications can be embedded along with terms, allowing

researchers to learn relationships between embedded literature and concepts in the vocabulary, directly from the data.

Several recent publications have leveraged these new NLP techniques to explore new ways of representing publications, visualize research landscapes, and facilitate document search and retrieval. Approaches range from representing documents using the text in their abstracts (e.g. Hain et al., 2020) to embeddings derived from citation context (e.g. Berger et al., 2017; He & Chen, 2018), and a number of studies have explored combining text-based approaches with citation-based approaches (e.g. Cohan et al., 2020; Ganguly & Pudi, 2017). However, while many of these papers find that their representations of scientific literature capture useful and relevant aspects of publication context, topic, purpose, or overall relatedness, and often outperform traditional techniques in specific tests selected by the authors, embedding-derived relatedness measures remain difficult to interpret, and their resulting vector representations of papers are essentially black boxes. The nature of the resulting relatedness measures is difficult to compare to traditional bibliometric techniques, and this ambiguity makes the use of these new embedding techniques in science mapping and publication clustering less attractive.

Our research aims to compare a set of common NLP-based publication representations and their resulting relatedness measures with traditional citation-based and abstract-based publication relatedness measures. We follow the methodology introduced by Waltman et al. (2020), allowing us to compare the accuracy of clustering results obtained using various publication relatedness measures based on a selected baseline relatedness measure. We compare four methods for generating publication embeddings with two established bibliometric relatedness measures, one based on publication titles and abstracts, and one based on citations, in two data sets. Our goal is to establish whether embedding-derived publication relatedness measures more strongly resemble established text-based relatedness measures or citation-based relatedness measures.

Methods

Our first data set is selected from the Elsevier ScienceDirect corpus previously analyzed in detail by Boyack et al. (2018). We select scientometrics-related publications using the meso-level of the CWTS hierarchical publication classification system (Waltman & van Eck, 2012). This produced a total of 7825 publications, that are cited in 51003 full-text sentences (not limited to the scientometrics field). 3169 of these publications feature in at least five citing sentences, for a total of 44464 citing sentences. If a sentence features multiple citations, only the first citation is considered, to avoid associating the exact same context with multiple works.

The second dataset is the SciCite citation intent dataset introduced by Cohan et al. (2019). This consists of 6427 publications and 10817 sentences citing these publications, as well as annotated citation intent labels noting whether a citing sentence cites a paper as *background*, for its *methods*, or to compare *results*. Titles and abstracts of these publications, as well as publications cited by or citing them, were retrieved using the Semantic Scholar API.

Word2vec

Our first embedding-derived relatedness measure is a simple word2vec embedding of citation labels. In each citing sentence, the label referencing the cited publication is replaced with a token unique to the cited work, and a word2vec embedding is trained on the resulting corpus using the *gensim* python library (Rehurek & Sojka, 2010). For a schematic representation, see figure 1. Relatedness of publications is calculated as the cosine similarities of the unique publication tokens. The word2vec architecture has a few adjustable hyperparameters. We varied

the embedding size, window size, epochs, and negative sample size, and tried both the CBOW and skip-gram methods. Skip-gram with a window size encompassing the entire sentence, keeping other parameters set to their default values, produced the most stable publication-to-publication similarities over multiple model runs.

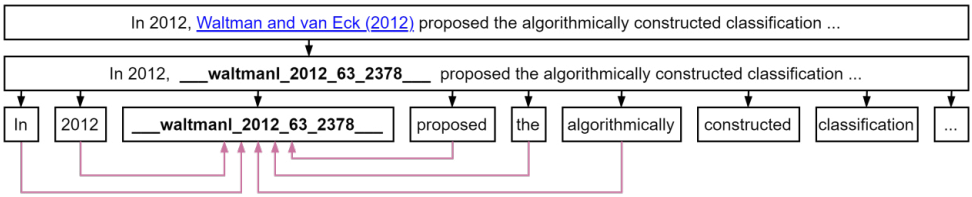

Figure 1: example of word2vec embedding of citation labels, window size three.

Sentence-BERT

Entire sentences can be embedded using Sentence-BERT (Reimers & Gurevych, 2019). Averaging the embeddings of sentences that cite the same paper then offers us another way to generate representations for cited publications based on their citation context. In this procedure, we again associated sentences only with the first cited publication, and omitted citation labels entirely from the sentence. Since the Sentence-BERT architecture can use a variety of language models, we repeated this process twice, once using the original BERT base model (Devlin et al., 2018), and once using the SciBERT model (Beltagy et al., 2019) which was trained specifically on scientific text.

SPECTER

Our final embedding-based relatedness measure uses SPECTER (Cohan et al., 2020) to generate publication representations based on titles and abstracts alone. SPECTER's core feature is that it generates document-level representations using only titles and abstracts, in a manner informed by pre-training on a citation graph. This allows it to outperform other architectures that rely solely on title and abstract information in a variety of performance benchmarks such as citation prediction and topic classification (Cohan et al., 2020).

Conventional relatedness measures and comparison

Waltman et al. (2020) introduce a methodology for comparing publication relatedness measures based on the accuracy of clustering solutions, compared to a baseline relatedness measure. Following their work, we use the BM25 text similarity measure based on title and abstract noun phrases, and the combined direct citation, bibliographic coupling and co-citation (DC-BC-CC) relatedness measure, to represent baseline text-based and citation-based relatedness measures, respectively. All relatedness measures are also simplified and normalized, fist by keeping only the 20 most strongly related connections for each document, and subsequently normalizing their weights by dividing the relatedness of each outgoing edge by the total weight of outgoing edges for each node.

Waltman et al. (2020) then define the accuracy of a clustering solution based on relatedness measure X, evaluated against measure C, as

$$A^{X|C} = \frac{1}{N}\sum_{ij}\left(c_i^X = c_j^X\right)r_{ij}^C$$

where $\left(c_i^X = c_j^X\right)$ equals 1 if the clustering c of document i is the same as that of document j, and 0 otherwise, with r_{ij}^C the relatedness between documents i and j per measure C. The normalization of the relatedness measures constrains this accuracy score between 0 and 1. Accuracy is computed for a variety of clustering solutions obtained using the Leiden algorithm (Traag et al., 2019), using a range of clustering resolution parameters, and the obtained accuracy is then plotted against the granularity of the resulting clustering solution in a granularity-accuracy plot, allowing us to easily compare the clustering accuracy of relatedness measures over a range of resolutions.

Results

Figure 2a displays the accuracy of the clustering solutions resulting from the various relatedness measures, using the BM25 title and abstract based relatedness as a baseline, over different clustering granularities. This analysis is limited to only the 2979 publications that are connected by relatedness in each of the six networks. Figure 2b uses the DC-BC-CC citation relatedness baseline, while 2c and 2d use word2vec and SPECTER-derived relatedness instead.

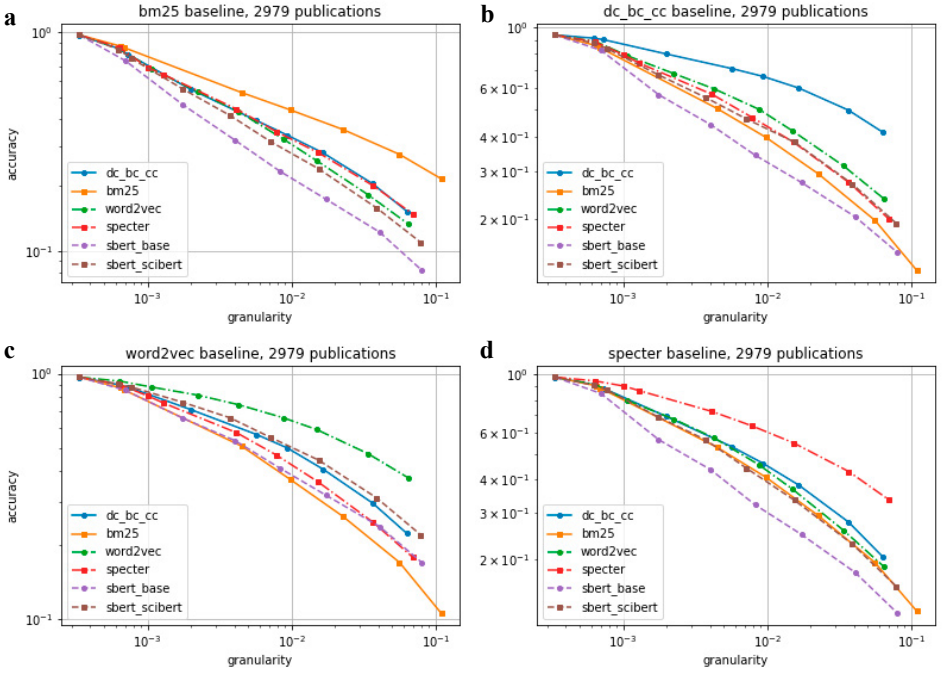

Figure 2: granularity-accuracy plots, scientometrics data set.

Figure 2c shows us that, in our scientometrics data set, the word2vec embedding derived from citing context resembles the citation-based relatedness measures more strongly than it does the title and abstract text-based relatedness measure. This is perhaps surprising, as it is fundamentally text-based, though this text originates from sentences that cite, which would explain the similarities to citation relatedness patterns. More surprising, figure 2d shows that the SPECTER-embeddings, despite being based on the same title and abstract text as the BM25 relatedness, behave more like the DC-BC-CC citation relatedness. Indeed, figure 2a shows that SPECTER and DC-BC-CC are approximately equally accurate when compared directly against the BM25 relatedness. Finally, figure 2b shows that word2vec, SPECTER and the SciBERT-based sentence embedding more closely resemble citation relatedness than BM25, which was

previously found to be the best-performing title and abstract relatedness measure, at least in our sample in which publications with 5 or more in-text citations are considered. When also including publications with fewer than 5 in-text citations (not pictured), SPECTER and BM25 are the most similar to DC-BC-CC, followed by word2vec and SciBERT sentence embedding.

The SciCite dataset, moreover, contains not only citation context for each publication, but is hand-labeled with citation intent. These labels (*background*, *methods*, and *results*) make for a poor relatedness measure but can be assessed at the cluster level. We can express the diversity of these labels over the clusters using Rao-Stirling diversity (see e.g. Leydesdorff et al., 2019)

$$\Delta = \sum_{i,j} d_{ij}(p_i p_j)$$

in which d_{ij} is a distance measure between clusters i and j, and p_i is the proportion of nodes in cluster i. To compute a distance measure, we take the numbers of citing sentences labeled *background*, *methods* and *results* in each cluster as a three-length vector and take the cosine distance of these vectors between each cluster pair. This means that diversity is maximized if publications cited with the same intent are clustered together. We vary the clustering resolution and plot the resulting diversity of SciCite intent labels against cluster granularity in Figure 3.

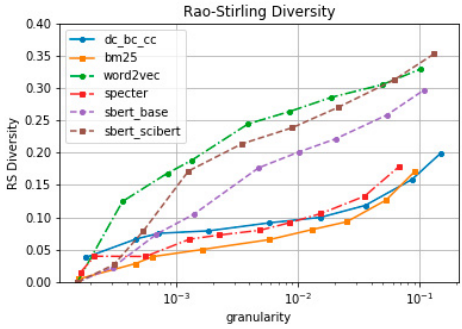

Figure 3: diversity-granularity plot of citation intent labels, SciCite dataset.

Figure 3 shows us that, with some margin, the citation context derived embeddings outperform the conventional relatedness measures, and that SPECTER improves only marginally over BM25 when it comes to clustering citation intent. Even BERT base, while lagging in figure 2, improves markedly on the conventional relatedness measures.

Discussion

Our research shows that embedding representations of publications have the potential to resemble conventional citation-based relatedness structures more accurately than the best-available conventional title and abstract-based relatedness measure. This holds true even for a simple word2vec embedding, provided enough citation context is available. Context-based embeddings also show a stronger correlation to citation intent labels in the SciCite data set than conventional citation-based or abstract-based relatedness measures.

It must be noted that the embedding-based relatedness measures used in this research might be further explored and optimized for the task of establishing publication relatedness. Especially BERT-based representations are typically fine-tuned for downstream tasks. Nevertheless, our research demonstrates the base feasibility of using embedding representations of publications

and provides important insight into how they relate to traditional publication relatedness measures.

References

Beltagy, I., Lo, K. & Cohan, A. (2019). SciBERT: A pretrained language model for scientific text. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing*, 3615–3620.

Berger, M., Mcdonough, K. & Seversky, L. M. (2017). cite2vec: Citation-Driven Document Exploration via Word Embeddings. *IEEE Transactions on Visualization and Computer Graphics,* 23(1), 691-700.

Boyack, K. W., van Eck, N. J., Colavizza, G. & Waltman, L. (2018). Characterizing in-text citations in scientific articles: A large-scale analysis. *Journal of Informetrics*, 12(1), 59–73.

Cohan, A., Ammar, W., Van, M. & Cady, Z. F. (2019). Structural Scaffolds for Citation Intent Classification in Scientific Publications. *Proceedings of NAACL-HLT*, 3586–3596.

Cohan, A., Feldman, S., Beltagy, I., Downey, D. & Weld, D. (2020). SPECTER: Document-level Representation Learning using Citation-informed Transformers. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2270–2282.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 4171–4186.

Ganguly, S. & Pudi, V. (2017). Paper2vec: Combining graph and text information for scientific paper representation. *Lecture Notes in Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, *10193 LNCS*, 383–395.

Gläser, J., Glänzel, W. & Scharnhorst, A. (2017). Same data—different results? Towards a comparative approach to the identification of thematic structures in science. *Scientometrics*, 111(2), 981–998.

Hain, D., Jurowetzki, R., Buchmann, T. & Wolf, P. (2020). *Text-based Technological Signatures and Similarities: How to create them and what to do with them*. http://arxiv.org/abs/2003.12303

He, J. & Chen, C. (2018). Temporal Representations of Citations for Understanding the Changing Roles of Scientific Publications. *Frontiers in Research Metrics and Analytics*, 3, 27.

Le, Q. V. & Mikolov, T. (2014). Distributed Representations of Sentences and Documents. *31st International Conference on Machine Learning, ICML 2014*, 4, 2931–2939.

Leydesdorff, L., Wagner, C. S. & Bornmann, L. (2019). Interdisciplinarity as diversity in citation patterns among journals: Rao-Stirling diversity, relative variety, and the Gini coefficient. *Journal of Informetrics*, 13(1), 255–269.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013, January 16). Efficient estimation of word representations in vector space. *1st International Conference on Learning Representations, ICLR 2013 - Workshop Track Proceedings*.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. & Dean, J. (2013). Distributed Representations of Words and Phrases and their Compositionality. *Advances in Neural Information Processing Systems*, 26.

Rehurek, R. & Sojka, P. (2010). Software Framework for Topic Modelling with Large Corpora. *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, 45–50.

Reimers, N. & Gurevych, I. (2019). Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *EMNLP-IJCNLP 2019 - 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing,* 3982–3992.

Traag, V. A., Waltman, L. & van Eck, N. J. (2019). From Louvain to Leiden: guaranteeing well-connected communities. *Scientific Reports*, 9(1), 1–12.

Waltman, L., Boyack, K. W., Colavizza, G. & Van Eck, N. J. (2020). A principled methodology for comparing relatedness measures for clustering publications. *Quantitative Science Studies*, 1(2), 691–702.

Waltman, L. & van Eck, N. J. (2012). A new methodology for constructing a publication-level classification system of science. *Journal of the American Society for Information Science and Technology*, 63(12), 2378–2392.