Scientometrics (2022) 127:7751–7768
https://doi.org/10.1007/s11192-022-04312-x

Impact of model settings on the text-based Rao diversity index

Andrea Zielinski[1]

Received: 28 October 2021 / Accepted: 8 February 2022 / Published online: 12 March 2022

Abstract

Policymakers and funding agencies tend to support scientific work across disciplines, thereby relying on indicators for interdisciplinarity. Recently, text-based quantitative methods have been proposed for the computation of interdisciplinarity that hold promise to have several advantages over the bibliometric approach. In this paper, we provide a systematic analysis of the computation of the text-based Rao index, based on probabilistic topic models, comparing a classical LDA model versus a neural network topic model. We provide a systematic analysis of model parameters that affect the diversity scores and make the interaction between its different components explicit. We present an empirical study on a real data set, upon which we quantify the diversity of the research within several departments of Fraunhofer and Max Planck Society by means of scientific abstracts published in Scopus between 2008 and 2018. Our experiments show that parameter variations, i.e. the choice of the Number of topics, hyper-parameters, and size and balance of the underlying data used for training the model, have a strong effect on the topic model-based Rao metrics. In particular, we could observe that the quality of the topic models impacts on the downstream task of computing the Rao index. Topic models that yield semantically cohesive topics are less affected by fluctuations when varying over the number of topics, and result in more stable measurements of the Rao index.

Keywords LDA topic model · Rao Stirling · Interdisciplinarity

✉ Andrea Zielinski
andrea.zielinski@isi.fraunhofer.de

[1] Fraunhofer Institute for Systems and Innovation Research, Breslauer Strasse 48, 76131 Karlsruhe, Germany

Introduction

Interdisciplinary research (IDR) is a mode of research that integrates information, data, techniques, tools, perspectives, concepts, and theories from two or more scientific disciplines. According to innovation theory, research addressing social and economic needs is often beyond the scope of a single discipline and therefore policy-makers often promote IDR [see National Academies (2005)][1].

The most frequently used method to operationalize the concept of IDR is by means of the multi-dimensional Rao-Stirling indicator (Stirling, 2007) which contains three different dimensions: (1) variety: number of distinctive categories; (2) balance: evenness of distribution; and finally (3) disparity: degree to which the categories are different. In bibliometrics, the diversity score considers the number of publications in a scientific category and/or the percentage of references to documents into other scientific disciplines and relies on the metadata of scientific publications (Leydesdorff & Rafols, 2009).

According to Cassi et al. (2017), Rao is a relevant indicator at the scale of a research institution and can be adopted for comparing institutions' interdisciplinary practices but requires a proper delineation into research fields. Even though major publishers such as Elsevier provide a categorization scheme designed to define a scientific discipline, e.g. the ASJC codes in Scopus, the classification of articles is often too imprecise and course-grained for measuring interdisciplinarity, since articles are assigned to subject categories associated with the journal rather than the article (Zhang et al., 2016).

In contrast, clustering approaches based on machine learning (ML) allow to produce more fine-grained, faceted topics of the research literature. In addition, they are able to classify scientific knowledge into novel categories without the need to resort to human-defined subject categories that might be outdated (Suominen et al., 2016). In particular probabilistic topic models such as the classical latent Dirichlet allocation (LDA) (Blei et al., 2003, 2010) have been applied to the task of mapping research into fields of science (Yau et al., 2014).

Topic models have also been used to capture the notion of interdisciplinarity of research institutions, either based on scientific publications (Nanni et al., 2016; Paul & Girju, 2009) or research awards (Nichols, 2014; Talley et al., 2011).

When dealing with large datasets, employing ML algorithms that are able to calculate indicators in an unsupervised fashion are particularly attractive. An appealing work in this direction is provided by Bache et al. (2013) and Wang et al. (2014) who have re-interpreted the Rao Stirling indicator on the basis of topic modeling, relying exclusively on textual features. The authors conduct experiments on synthetic as well as real data sets (using abstracts, full papers, or grants) that suggest that also the text-based implementation of Rao's index correlates with human judgments.

Topic models are popular because of their data-driven nature that seeks to find emerging clusters of scientific disciplines automatically. Furthermore, they are multi-mixture models in which a document may contain several topics. Yet, it is well known that purely unsupervised models such as LDA often result in topics that do not fit the needs of a specific application, i.e. they do not necessarily align with an established subject domain classification schema. Moreover, hyper-parameter setting is important to produce high-quality topics (Chang et al., 2009; Syed & Spruit, 2018). According to Tang et al. (2014), LDA's

[1] https://www.nsf.gov/od/oia/additional_resources/interdisciplinary_research/definition.jsp.

performance depends mainly on the factors (a) number of topics, (b) the Dirichlet (hyper) parameters, (c) number of documents, and d) the length of individual documents.

One of the most crucial factors is the number of topics: Standard LDA requires that a good estimate of the number is known to avoid over-/underfitting of the data. By design, LDA topic models often make use of the sparse Dirichlet priors such that each document contains only a small number of topics and each topic uses only a small set of words frequently. Yet, setting these hyper-parameters has an impact on the document-topic and topic-word distribution and leaves room for variation.

This paper seeks to investigate in a pilot study in how much the LDA-based Rao measure is sensitive to parameter settings and if it can be used as a reliable indicator to automatically calculate a diversity ranking for an institute's research output, i.e. based on abstract and title as listed in Scopus.

The rest of this article is organized as follows. First, we present related work. In the second section, we summarize the definition of the Rao-based disciplinarity indicator, and discuss the topic-specific calculation of the metrics on the basis of LDA. In addition, we utilize a deep learning based neural topic model. Then, we briefly introduce the data used for the empirical analyses. Subsequently, we present the experimental results on the publication output of two research institutes. Finally, we conclude the article and state future directions.

Related work

Establishing methods for defining and measuring interdisciplinarity is central and intensively studied within bibliometrics (Wagner et al., 2011). The main goal of the task is to automatically define reliable indicators that are efficient to calculate, predictive, and robust regarding data errors (Guo et al., 2009).

A well-established indicator has been set up by Rao (1982) and Stirling (2007), i.e. the Rao-Stirling diversity, which considers variety (number of distinct categories), balance (evenness of the distribution), and disparity (distances or similarities between categories). Accordingly, variety is defined as the number of subject categories assigned to the papers' references and takes values between one and the number of subject categories, balance is a function of assignments across categories, and disparity is the complement of similarity and computed pairwise between the referenced subject categories. Yet, the bibliometric operationalization of diversity is actively discussed in the research community (Leydesdorff, 2018; Leydesdorff et al., 2019; Rousseau, 2019). Based on a case study on Web of Science data, Wang and Schneider (2020) found that many measures are inconsistent. This also holds for the Rao-Stirling indicator which has recently been criticized for its low discriminatory power (Zhou et al., 2012).

Starting from the pioneering works by Hall et al. (2008), Paul and Girju (2009), Griffiths and Steyvers (2004), among others, models of diversity have also spread in the area of computational linguistics, especially in connection with topic modelling. These approaches all rely on accepted subject classifications from journals or conference proceedings. Paul and Girju (2009) assess the interdisciplinary nature of distinct research fields based on their topic overlap. Document collections featuring different research fields are compared via their mean topic vectors using cosine similarity. Nichols (2014) applies LDA topic modelling to analyze research awards issued by the National Science Foundation (NSF), where the institutional structure serves as a proxy for research disciplines and topics are assigned to the discipline in which they occur most frequently. In contrast, Bache et al. (2013) define the Rao measure entirely on the LDA output, without mapping topics to pre-defined classes that reflect specific scientific disciplines, and without verifying the nature of the topics. In their work, the Rao index is derived in a fully data-driven way and computed on the level of a document over the LDA document-topic and word-topic matrices. The authors conduct various experiments on PubMed Open Access, NSF Grant Awards, and the ACL Anthology. The authors state that the topic-based Rao diversity measure outperforms alternative approaches like entropy in a classification task on pseudo documents. The authors hypothesize that the method would be invariant to the number of topics in the model. Wang et al. (2014) use the same approach as Bache et al. (2013), however, their LDA model is induced from a corpus that considers a paper's references and citations. The authors propose a discounting weight on the balance attribute as part of the diversity score.

Furthermore, a variety of LDA models has been proposed to address certain limitations of LDA and give better performance, when it comes to detecting rare topics in an imbalanced collection (Jagarlamudi et al., 2012) or short text (Newman et al., 2011; Quan et al., 2015;). Incorporating meta-information directly into the generative process of topic models can improve modelling accuracy and topic quality. Various authors have used document labels as a priori information to infer the underlying topic distributions (Chuang et al., 2012; Ramage et al., 2010). It has been shown that document regularization yields improved model performance, however requires reliable labeled data (Zhao et al., 2017).

Some recent works use neural topic models that offer additional flexibility over the traditional probabilistic approaches, since they allow to easily integrate prior knowledge, e.g., pre-trained word and text embeddings. This is important, because embedding techniques help to alleviate the language variation problem, i.e. the same concepts might be expressed in different ways in different scientific communities. The embedded topic model (ETM) proposed by Dieng et al. (2020) is a generative model and relies on word embeddings (Mikolov, 2013) and has shown improved topic quality across various datasets.

In our study, we compare the classical LDA versus the neural network topic model ETM. Note that the performance of the topic models can vary depending on the specific task to be solved (Doan et al., 2021), which in our case includes *depicting research fields* (science mapping), *uncovering cohesive topics*, and the downstream task of *computing the Rao index.*

Rao stirling diversity measures based on LDA

The classic Rao Stirling diversity index has been widely used to measure diversity and interdisciplinarity (e.g. Porter & Rafols, 2009; Wang et al., 2014). In this section, we will discuss the three different dimensions of diversity i.e. variety, balance, and disparity.

Variety

Instead of subject categories, the thematic diversity can be related to the number of distinct topics K. A characteristic of latent topics generated by LDA, however, is that every topic is in principle present in every document, with a non-zero proportion. A rough estimate is that a large number of topics is needed to account for small scientific communities. Current approaches set the number of topics between K = 300 (Griffiths et al., 2004), and K = 1000 (Nichols, 2014) to cover the whole scientific landscape. In practice, a higher number of

Table 1 Topic similarity measures based on the topic-word matrix

Metrics	Measure	Author
Divergence-based metrics	JS Divergence	Hall et al. (2008)
Coefficient-based metrics	Jaccard	Ramage et al. (2011)
Distance-based metrics	Hellinger Distance	Aletras et al. (2014)
	Cosine	Wang et al. (2019)

topics will necessarily result in a larger variety. This issue is crucial because the optimal number of topics in a corpus is unknown and based on a heuristic choice.

Balance

Generally, a more balanced document-topic distribution results in a higher thematic diversity estimate. The balance component as part of the Stirling Index can be calculated as follows:

$$\sum_{i=1}^{K}\sum_{j=1(i\neq j)}^{K} P(i|d)P(j|d) \quad \forall\, \mathrm{d}: \min^{\mathrm{T}\,(\mathrm{T}-1)} \leq \mathrm{B} \leq \max^{\mathrm{K}\,(\mathrm{K}-1)}$$

where P(i|d) is the probability of topic *i* in a paper d and individual pair scores take small values in the range of [$min : \sim 10^{-6}$,$max : \sim 0.25$]. Regarding the distribution of papers into scientific categories, it is likely that any database that seeks to monitor scientific research will consist of long-tailed, imbalanced data that is prevalent in any real-world setting. In order to deal with the issue of imbalanced data, it is necessary to have a good estimate of the scalar concentration parameter α that governs the shape of the document-topic distribution. Setting α to a value close to zero will result in a distribution where the probability mass is concentrated on a smaller set of topics. Moreover, an asymmetric α learns a non-uniform prior, assuming that certain topics might be more prominent in the collection. Thus, some topics may be the majority topic in a larger share of documents in the corpus overall and make up more of the total corpus. As an alternative, proper sampling methods that re-balance the data can help to mitigate the problem.

Disparity

Topic similarity metrics can be used to measure the (dis) similarity between two topics and are generally computed from the topics' word probability distributions. In this work, we use the distance function δ to estimate the similarity $\delta(i,j)$ between topics *i* and *j*. A systematic evaluation of different topic similarity measures for pairs of topics generated by LDA has been conducted by Aletras et al. (2014) and Wang et al. (2019), comparing which measure aligns best with human judgments. Their experiments show that intrinsic coherence scores like Jensen-Shannon, Hellinger, Jaccard Distance and cosine similarity applied on the original dataset are generally inferior to extrinsic metrics that make use of external data. However, it is crucial that the external datasets fit well to the domain of the data used to build the topic model. In the setting of Aletras et al. (2014), co-occurrences of words are drawn from Wikipedia, while Wang et al. (2019) use word embeddings, which have been specifically trained on Twitter data. Since external data that covers the immense variety of scholarly topics is not readily available, we use intrinsic measures to compute topic similarity. An alternative approach proposed in Bache et al. (2013) is to make use of the document-topic matrix in order to calculate the probability of distinct topics that co-occur in documents. The motivation for this approach is that topic distributions tend to be distinct by definition. We refrain from this approach, because standard LDA is unable to model relations among topics due to its use of a single Dirichlet distribution, and thus it is not possible to detect correlations amongst topics directly. In order to transform the similarity matrix between topics *i* and *j* into a dissimilarity matrix, a frequently applied solution is 1 − δ(i, j) and 1/δ(i, j). Based on prior studies, we choose the metrics listed in Table 1 for our evaluation study. The topic distance also indicates how well the topics are separated which is a sign for a high quality LDA model. In order to produce topics that are distinct from each other, a symmetric prior of the topic-word distribution is generally preferred, and the β hyper-parameter needs to be set to values ranging between 0.1 and 0.01, so that the topic vectors concentrate on fewer words (Wallach et al., 2009).

Summary of diversity measures

We apply the Rao-Stirling index (RS) to measure the degree of interdisciplinarity for each institute (aggregate over all publications of the institute) and experiment with different dissimilarity measures. The Rao Stirling diversity is defined as

$$RS(d) = \sum_{i=1}^{K} \sum_{j=1(i \neq j)}^{K} P(i|d)P(j|d)\delta(i,j)$$

In addition, the broadness of an institute can be determined by means of the Shannon Entropy (H) based on the distribution over latent topics for each institute. The measure combines the variety and balance dimension, while it ignores disparity. A high topic entropy signals an even distribution and broader spectrum of topics. Shannon Entropy is defined as

$$H(d) = -\sum_{i=1}^{K} P(i|d) \ln P(i|d)$$

The diversity measure can thus be obtained from the topic-document and word-topic distributions of the model. More concretely, we use Θ (topic-document probability matrix) for calculating the balance between topics and Φ (word-topic probability matrix) for computing the distance δ between topics. A limitation in our use case is obviously, that the underlying distributions are unknown and varying over the parameter setting for the number of topics K and hyper-parameters α and β might yield different Rao scores. Also, the size and length of the training data is crucial, since the priors are estimated from the observed counts in the data.

Datasets

In the present work, we use title and abstract from Scopus, a bibliographic database introduced in 2004 by Elsevier. Scopus provides a comprehensive collection of the scientific landscape, covering the world's leading journals, and is a real-time monitor

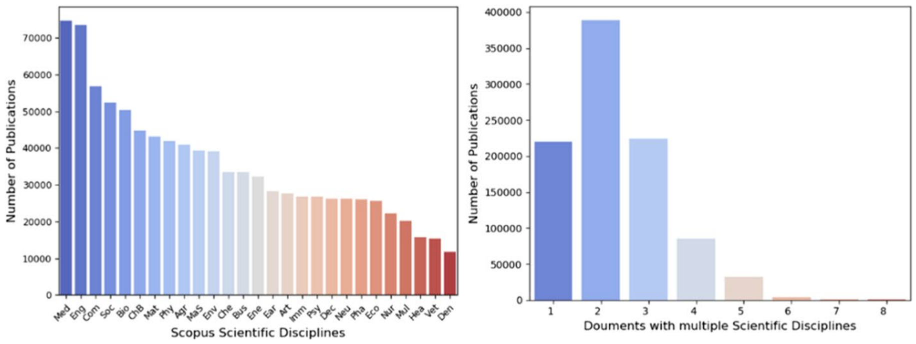

Fig. 1 Statistics for Scopus publications—Scopus world

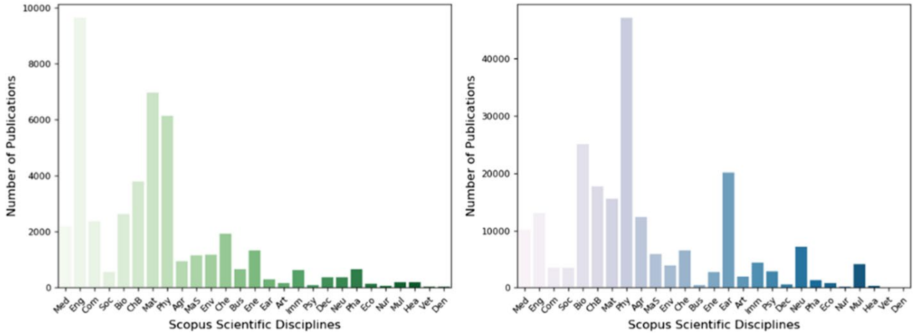

Fig. 2 Statistics for Scopus FH (left) Scopus MPG (right)

corpus that is both big in size and rich in metadata. It offers, e.g., research institutions of the authors as metadata records.

Scopus world 2018 (Scopus world)

To explore the interdisciplinarity of an institution, we aim to compute the diversity indicator on a balanced corpus that covers all scientific fields. Therefore, we sampled a corpus from Scopus where we seek to give equal weight to all scientific domains to mitigate the minority class problem, since the distribution of papers and journals over disciplines is heavily skewed (e.g., the humanities are underrepresented in the corpus). The result is a corpus of randomly selected publication abstracts and titles from all major fields of Scopus of the year 2018 (see Fig. 1).

In bibliometrics, the average number of subject categories of a publication, accumulated over an institute, can already serve as an indicator for interdisciplinarity (Levitt & Thelwall, 2008). The higher the value, the more interdisciplinary the institute. On the publication level, we see that the majority of documents is assigned to more than one discipline, i.e. on average there are 2.3 subject fields per publication (see Fig. 1, right).

Table 2 Dataset statistics

Data sets	Number of institutions	Number of abstracts
Scopus FH 2010–2018 (Scopus FH)	74	19,661
Scopus MPG 2010–2018 (Scopus MPG)	95	111,986
Scopus World 2018 (Scopus World)		517.516

Institute-specific publications: Scopus FH and Scopus MPG

As can be seen in Fig. 2, the research profiles of Fraunhofer (FH) and Max Planck Society (MPG) are rather imbalanced, e.g., Scopus FH contains a huge share of publication abstracts from Engineering, while Scopus MPG publishes mostly on Physics and Astronomy. Only a small fraction of articles is dedicated to, e.g., Dentistry. In the FH corpus, 82.41% are assigned to more than 1 field and on average there are 2.47 subject fields per publication, while for the MPG corpus, 70.59% are assigned to more than 1 field and on average there are 2,19 subject fields per publication. Table 2 provides a detailed breakdown of the datasets used in our study.

Empirical study

In this section, we describe settings used in our experiments. We start by describing the probabilistic topic models chosen. We then introduce the train and test corpora and the metrics used to examine the models.

Our goal is to test the effects of varying the topic model settings on the diversity measure, composed of disparity, balance and variety. Our research hypothesis is that to provide a good Rao Index of the data, it is desirable that the selected topics are both coherent and interpretable, and have a high coverage of the data.

Model selection and parameter settings

We address the computation of Rao, comparing two probabilistic topic models, namely the classical LDA model (Blei et al., 2003) which works purely on bag-of-words representation of documents *vs.* the neural topic model ETM (Dieng et al., 2020) that relies on word embedding representations. ETM jointly trains words and topics in a shared embedding space and is able to integrate pre-trained word embeddings. Note that topics reflect global semantic and syntactic features, while word embeddings encode more local aspects of a word. Their representations capture different aspects of word contexts and are therefore complementary.

Classical LDA

Variational inference (Hoffman et al., 2013) as implemented in *gensim* is used for model inference and standard Laplace smoothing factors with $\gamma = 0.1$ and 2000 iterations. We set the number of topics K = 100, 150, 200, 250, 300. As standard parameters of the Dirichlet

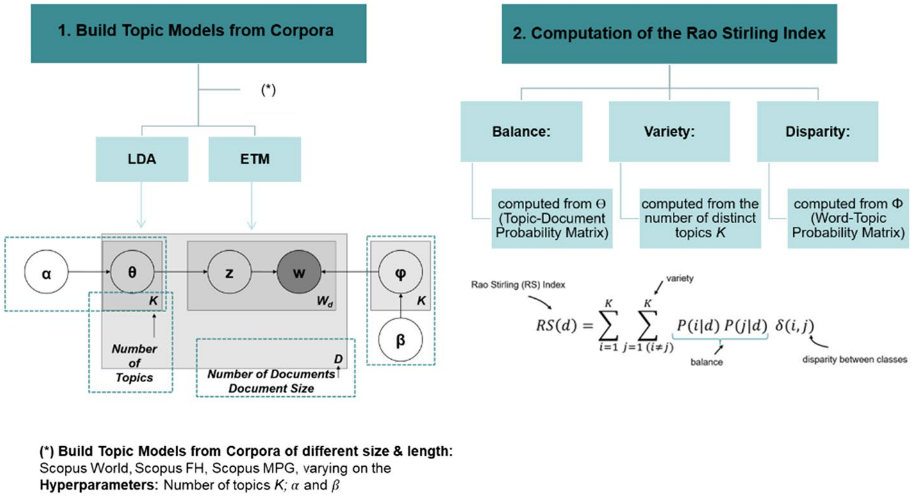
1. Build Topic Models from Corpora
(*)
LDA
ETM
α
θ
K
z
w
W_d
φ
K
β
Number of Topics
D
Number of Documents
Document Size
2. Computation of the Rao Stirling Index
Balance:
Variety:
Disparity:
computed from Θ (Topic-Document Probability Matrix)
computed from the number of distinct topics K
computed from Φ (Word-Topic Probability Matrix)
Rao Stirling (RS) Index
variety
RS(d) = Σ_{i=1}^{K} Σ_{j=1 (i≠j)}^{K} P(i|d) P(j|d) δ(i,j)
balance
disparity between classes
(*) Build Topic Models from Corpora of different size & length:
Scopus World, Scopus FH, Scopus MPG, varying on the
Hyperparameters: Number of topics K; α and β

Fig. 3 Rao-index computation process

prior we use (a) α = 0.1, (b) a non-uniform α estimated automatically from the data (Li et al., 2006) and (c) a fixed normalized asymmetric prior of 1/K (Wallach et al., 2009). Regarding the topics-word distributions, we set (a) $\beta = 0.1$ and (b) $\beta = 0.01$, unless otherwise specified.

ETM

Amortized inference (Gershman & Goodman, 2014) and the variational auto-encoder (Kingma & Welling, 2014) are used and an encoder-decoder architecture and recurrent neural network. The number of topics is set as above. The concentration parameters $\alpha = 1$ and $\beta = 1$ of the Dirichlet distributions are fixed model hyperparameters (Dieng, 2020). We train with $\gamma = 100$ epochs.

Pre-fitted word embeddings

For ETM, we incorporate Word2Vec embeddings (Mikolov et al., 2013) using the skip-gram neuronal net architecture trained on the *Scopus World* 2018 corpus. The size of the hidden-layer and subsequent word embeddings is 300 with parameters *min count* = 10 and a window size of 6.

Choice of the training and test corpora

As training corpus, we use *Scopus World*, and alternatively, *Scopus FH* and *Scopus MPG*. The last two corpora are composed of abstracts from FH and MPG published between 2008 and 2018 where we concatenate all abstracts by the same institute. We use the institute-specific corpora *Scopus FH* and *Scopus MPG* for testing.

The flow chart (see Fig. 3) shows the computation process for applying LDA and ETM to the computation of the text-based Rao Stirling Index.

Table 3 Coherence and Coverage for varying model size of LDA

Model	Dataset	100	150	200	250	300
LDA	Scopus World	− 7.53	− 8.77	− 9.75	− 10.90	− 11.74
LDA	Scopus FH	− 0.83	− 0.91	− 0.95	− 0.95	− 0.98
LDA	Scopus MPG	− 3.69	− 3.41	− 3.29	− 3.26	− 3.01

Table 4 Coherence (NPMI) for varying model size of ETM

Model	Dataset	100	150	200	250	300
ETM	Scopus World	0.128	0.128	0.126	0.115	0.114
ETM	Scopus FH	0.332	0.324	0.325	0.328	0.326
ETM	Scopus MPG	0.398	0.407	0.397	0.401	0.401

Preprocessing

We used sentence splitting, tokenization, lemmatization, and PoS tagging to filter all content words using the Stanford tools,[2] keeping only nouns, adjectives, verbs, and foreign words that consist of alphanumeric characters. We filtered out Named Entities (e.g. names of institutes, etc.) and tokens occurring in more than 70% of all documents. This resulted in 131,954, 12,598 and 36,381 unique words for *Scopus World*, *Scopus FH* and *Scopus MPG*, respectively.

Evaluation metrics

We assess modeling accuracy in terms of topic coherence under various settings of hyperparameters and number of topics. Even though determining the parameters is an established research area and various heuristics exist for real-life applications (Lau et al., 2014; Wallach et al., 2009), Chuang et al. (2012) have shown that a small change in term smoothing and prior selection can significantly alter the ratio of resolved and fused topics. Increasing the number of latent topics often leads to more junk and fused topics with a corresponding reduction in resolved topics.

Human assessment of the topic models show that while the classical LDA models are better in depicting the Scopus research fields, ETM outperforms it in uncovering cohesive topics. However, ETM has lower coverage of scientific fields and dismisses more topics than the LDA model with equal K (i.e., 300 topics).

Topic coherence versus coverage

The semantic coherence of the topics of the LDA model is measured using word co-occurrences within the original corpus by the UMass coherence score on the top 15 words from each topic (Mimno et al., 2011; Röder et al., 2015).

We compare the scores for varying model size of LDA, where models trained on *Scopus World* reach an average UMass score between − 7.53 (K = 100) to − 11.74 (K = 300) that

[2] https://stanfordnlp.github.io/CoreNLP.

Table 5 Diversity for varying model size of ETM

Model	Dataset	100	150	200	250	300
ETM	Scopus World	0.614	0.534	0.452	0.385	0.334
ETM	Scopus FH	0.561	0.466	0.401	0.353	0.317
ETM	Scopus MPG	0.525	0.439	0.407	0.341	0.324

decreases as we learn more topics. Even though models trained on *Scopus FH* and *Scopus MPG*, and thus less data, achieve higher UMass scores, they are inferior to the *Scopus World* model in terms of coverage (Table 3). Topics learned by the ETM model look overall more interpretable as is reflected in the topic coherence scores based on normalized pointwise mutual information (NPMI) (Aletras et al., 2013; Lau et al., 2014) that measures how related the top-10 words of a topic are to each other. As is shown in Table 4, we get a relatively high coherence score even for larger models.

Dieng et al. (2020) proposes a topic diversity metric that considers the percentage of unique words in the top 25 words of all topics (a value close to 1 indicates more varied topics). A high-quality topic model not only exhibits a high similarity within clusters (topic coherence), but also a low similarity between clusters (high topic diversity). The performance of a topic model can be assessed as the product of both measures, i.e. diversity and coherence. Comparing the topic quality this way, we notice that smaller topic model sizes perform best, since less topics lead to more diversity, while topics become similar to one another when increasing the number of topics (see Table 5), without having a negative impact on the coherence of topics.

Diversity index

Variety and balance scenario

We computed the evenness of the document-topic distribution for all FH institutes under various settings using Shannon Entropy, i.e. a high entropy signals interdisciplinarity.

Our experiments show that the choice of α impacts on the entropy values: Setting α to a value closer to zero results in a non-uniform document-topic distribution and lower entropy. Likewise, setting α = asym instead of α = auto has the effect that in the first setting the probability mass of the distribution will concentrate on fewer topics per document: Accordingly, entropy values are constantly lower for all topics. Additional experiments demonstrate the impact of proper sampling: Institutes show much higher equality and tendency to focus on more topics when the LDA model is computed on a data set, where samples were drawn such as to accommodate for balance beforehand, i.e. Scopus World, and results in high entropy values (Zielinski, 2021).

Disparity scenario

For LDA we observed that topical distance decreases, when β approaches 0. The inferred topics are a mixture of multiple topics and less separable when β = 0.1 instead of β = 0.01. Pairwise dissimilarity of topics is equally high for all other investigated distance metrics. For a model setting with 100 topics we receive Jensen-Shannon scores of 0.98 on average, ranging between 0.898 and 1 for β = 0.01 versus 0.79 and 1 for β = 0.1, respectively.

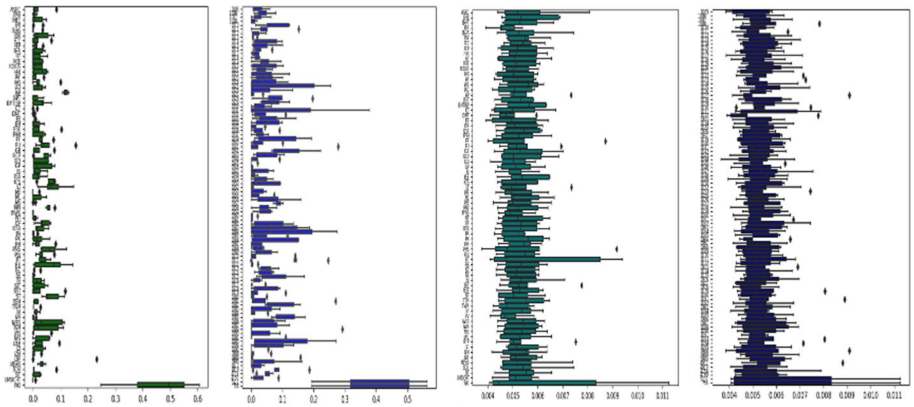

Fig. 4 Rao-Index for all Fraunhofer (green) and MPG (blue) institutes; computed for 100, 200, 300 topics on different LDA outputs, i.e. models are trained on *Scopus FH* versus *Scopus MPG* (left) versus Rao Index computed on Scopus *World* (right). (Color figure online)

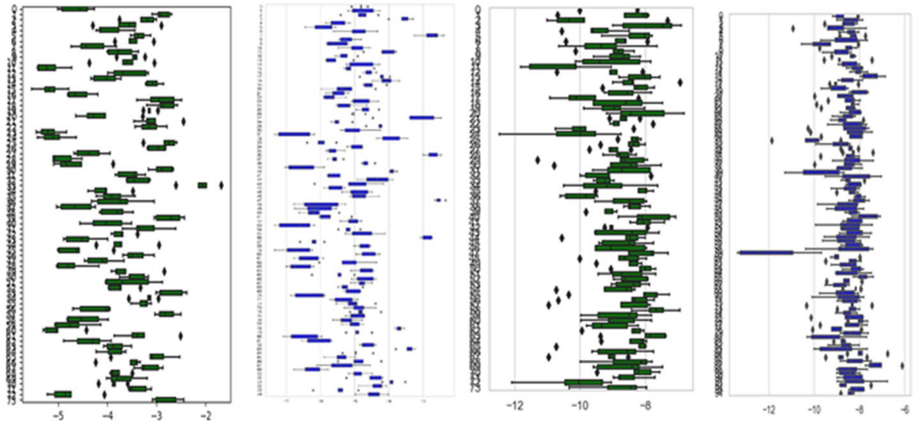

Fig. 5 Rao-Index for all Fraunhofer (green) and MPG (blue) institutes. (Color figure online)

Furthermore, the data is more separated when the number of topics becomes larger for both β = 0.1/0.01.

Experiments to assess the different dimensions of Rao

Rao scenario

We investigated the impact of different topic models on the Rao index. First, we calculated the index on the output of the LDA models trained on the institute-specific corpora *Scopus FH* and *Scopus MPG*. In the experiments, we could observe sharp fluctuations of the Rao index when varying over the number of topics (see Fig. 3, left). Rao index values range between 0.001 to 0.605 and 0 to 0.562, with a standard deviation of 0.062

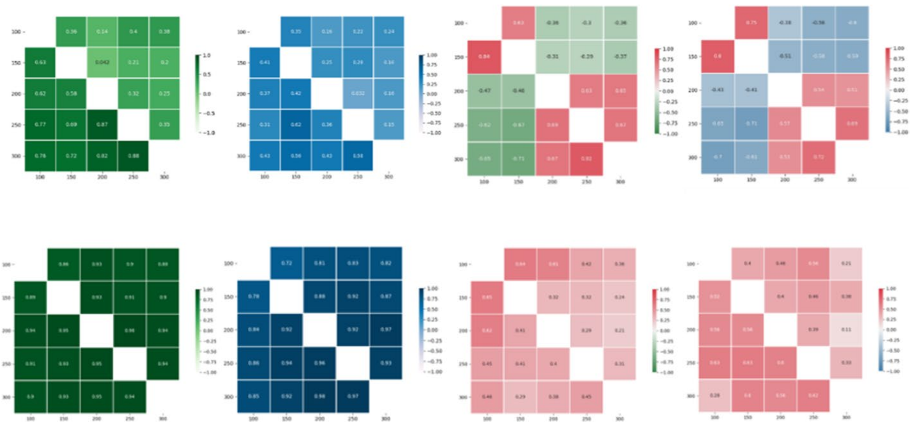

Fig. 6 Spearman (upper) and Pearson (lower) Correlation of Rao Index

and 0.077 for FH and MPG, respectively. Note that in this case, it was not possible to map the LDA topics fully to all ASJC fields, since the models have a relatively low coverage.

We also calculated the index for various LDA models trained on *Scopus World* and applied it to the *FH* and *MPG corpora*. The setting also makes comparisons between institutes possible and the LDA classifier is less prone to overfitting. However, as shown before, topic quality in terms of qualitative (human judgments) and quantitative (coherence) evaluation showed that many topics were not interpretable or meaningful.

For this setting, the standard deviations are much smaller. In this case, the Rao index takes small values, ranging between 0.004 and 0.011 for both institutes, and thus there is little difference between the values (see Fig. 4, right). The text-based Rao index thus suffers from the same limitations of low discriminating power as the bibliometric-based approach.

When comparing LDA to ETM, we can observe that the box plots for Rao trained on *Scopus FH* and *Scopus MPG* are more consistent with one from a normal distribution (median close to center of the box; whiskers of approximate equal length) and Rao also takes different values per institute, as opposed to the box plots for *Scopus World* which are more skewed, and values are more fixed (see Fig. 5).

Last but not least, we calculated the Spearman and Pearson Rank Correlation of the Rao Index varying on the number of topics and model size. Figure 6 shows the visualization of the coefficients based on the various outputs of Rao, depicting the pairwise correlations as a heatmap.

Computed from various LDA outputs, varying on the number of topics (on x-axis, y-axis) and model size of LDA (small models: left, large models: right), i.e. models are trained on *Scopus FH* (*green)* versus *Scopus MPG* (blue) vs. Rao Index computed on *Scopus World* (tested on *Scopus FH* (green–red), *Scopus MPG* (green–blue); Vanilla LDA model (upper) versus ETM model (lower).

As can be seen, the choice of K has a great influence on the Rao results: Pairwise comparisons of Rao results vary a lot, in particular for the LDA model, showing that there seems to be no association between the variables. In particular, Spearman correlation is weak, showing that the general rankings amongst institutes are not preserved when varying on the number of topics. Interestingly, a relatively strong Spearman and Pearson correlation (i.e., 0.80–1.0) can be achieved for the ETM model trained on the institute-specific corpora *Scopus FH* and *Scopus MPG.*

Conclusion

In this paper we investigated the Rao indicator for interdisciplinarity based on LDA for two German research institutes. Both institutions are specialized in certain scientific fields and have a more or less high propensity towards interdisciplinary research. It would be a benefit for politicians and decision makers to have an indicator that is able to truly reflect this trend and which can be computed automatically from any data set.

Our experiments show that the LDA-based Rao metrics has certain limitations, since the indicator crucially depends on the quality of the underlying topic model. However, automated measures such as the coherence or diversity measure have difficulties to select a proper topic model in an applied setting, particularly with no human-in-the-loop. We claim that the LDA-based Rao index can only serve as a useful indicator of interdisciplinarity, provided that the resulting topics carry coherent semantic meaning and have a high coverage of the data. When applied fully automatically, it might result in sharp fluctuations that make it an unreliable indicator. Our experiments on Scopus and two major German research associations show that Rao results that have been generated from different settings vary a lot. In fact, all parameter variations seem to have a strong effect on the output, i.e. choice of the number of topics, hyper-parameters, and size and balance of the underlying data used for training the model.

There seems to be a consensus in the research community that in order to select the best value of K, a qualitative evaluation of the performance of alternative LDA models with varying K is required (Suominen, 2016), ensuring that the topic model is able to represent and cover all major scientific fields. Moreover, it is crucial that hyper-parameters are set in such a way that they produce a topic model with sparse topic and word distributions. A qualitative analysis of the topics of various models reveals that the models fail to differentiate scientific topics from scientific discourse and junk topics. However, topics related to scholarly discourse not necessarily indicate interdisciplinary studies.

Appendix

We also assessed the identified topics generated by ETM along with their embeddings. As shown in Table 6, the embeddings provide additional semantically related terms for various scientific domains.

In our experiments, for each of the aforementioned models, we make use of the implementations provided by the authors.

Table 6 Topic embeddings trained on Scopus World (ETM)

Engineering	Material	Physics	Chemistry	Politics
Engineer bioengineering aerospace engineered construction application biotechnology nanotechnology geoscience electronics aeronautical industrial mechatronic biomedical nanoscience technology innovative science biomedicine	Nanomaterial elastomer composite printable polymer coating aerogel composites thermoset ceramic biomaterial filler plywood thermoplastic nanocomposite biopolymer elastomeric nanofiller nanostructure	Astrophysics astrophysical mechanic electrodynamics astronomy formalism cosmology relativistic mechanics physicist thermodynamics science quantum seismology mathematics geophysics hadron ultiphase analogy	Chemist organometallic catalysis chemical nanoscience biochemistry geochemistry physics solvents bioorthogonal nanomaterial frameworks thermodynamics photochemistry thermochemistry synthetic bioconjugation synthon synthesis	Political nationalism nationalist contestation hegemony discourse postcolonial neoliberalism rhetoric colonialism activist activism diplomacy ideology transnational politicize hegemonic governmentality slavery

Acknowledgements This paper is a substantially extended version of the ISSI2021 conference paper (Zielinski, 2021).

Funding Open Access funding enabled and organized by Projekt DEAL.

References

Aletras, N., & Stevenson, M. (2014). Measuring the similarity between automatically generated topics. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers (pp. 22–27)

Bache, K., Newman, D., & Smyth, P. (2013). Text-based measures of document diversity. In: Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 23–31)

Blei, D., Carin, L., & Dunson, D. (2010). Probabilistic topic models. *IEEE Signal Processing Magazine, 27*(6), 55–65.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993–1022.

Cassi, L., Champeimont, R., Mescheba, W., & De Turckheim, E. (2017). Analysing institutions interdisciplinarity by extensive use of Rao-Stirling diversity index. *PLoS ONE, 12*(1), e0170296.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J., & Blei, D. (2009). Reading tea leaves: How humans interpret topic models. *Advances in Neural Information Processing Systems, 22*, 288–296.

Chuang, J., Ramage, D., Manning, C., & 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 (pp. 443–452)

Dieng, A. B., Ruiz, F. J., & Blei, D. M. (2020). Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics, 8*, 439–453.

Doan, T., & Hoang, T. (2021). Benchmarking neural topic models: An empirical study. FINDINGS

Gershman, S., & Goodman, N. D. (2014). Amortized inference in probabilistic reasoning. *Cognitive Science, 36*, 1.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences, 101*(suppl 1), 5228–5235.

Guo, Z., Zhu, S., Chi, Y., Zhang, Z., & Gong, Y. (2009). A latent topic model for linked documents. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval (pp. 720–721)

Hall, D., Jurafsky, D., & Manning, C. D. (2008). Studying the history of ideas using topic models. In: Proceedings of the 2008 conference on empirical methods in natural language processing (pp. 363–371)

Hoffman, M. D., Blei, D. M., Wang, C., & Paisley, J. (2013). Stochastic variational inference. *The Journal of Machine Learning Research, 14*(1), 1303–1347.

Jagarlamudi, J., Daumé III, H., & Udupa, R. (2012). Incorporating lexical priors into topic models. In: Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (pp. 204–213)

Kingma, D.P., & Welling, M. (2014). Auto-encoding variational Bayes. CoRR. Retrieved from https://arxiv.org/abs/1312.6114

Lau, J. H., Newman, D., & Baldwin, T. (2014, April). Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality. In: Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (pp. 530–539)

Levitt, J. M., & Thelwall, M. (2008). Is multidisciplinary research more highly cited? A macrolevel study. *Journal of the American Society for Information Science and Technology, 59*(12), 1973–1984.

Leydesdorff, L. (2018). Diversity and interdisciplinarity: How can one distinguish and recombine disparity, variety, and balance? *Scientometrics, 116*(3), 2113–2121.

Leydesdorff, L., & Rafols, I. (2009). A global map of science based on the ISI subject categories. *Journal of the American Society for Information Science and Technology, 60*(2), 348–362.

Leydesdorff, L., Wagner, C. S., & Bornmann, L. (2019). Diversity measurement: Steps towards the measurement of interdisciplinarity? *Journal of Informetrics, 13*(3), 904–905. https://doi.org/10.1016/j.joi.2019.03.016

Li, W., & McCallum, A. (2006). Pachinko allocation: DAG-structured mixture models of topic correlations. In: Proceedings of the 23rd international conference on Machine learning (pp. 577–584)

Mikolov, T., Sutskever, I., Chen, K., Corrado, G., & Dean, J. (2013). Distributed representations of words and phrases and their compositionality. NIPS

Mimno, D., Wallach, H., Talley, E., Leenders, M., & McCallum, A. (2011). Optimizing semantic coherence in topic models. In: Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing (pp. 262–272)

Nanni, F., Dietz, L., Faralli, S., Glavaš, G., & Ponzetto, S. P. (2016). Capturing interdisciplinarity in academic abstracts. *D-Lib Magazine*. https://doi.org/10.1045/september2016-nanni

National Academies. (2005). National Science Foundation Committeeon Facilitating Interdisciplinary Research, Committee on Science, Engineering, and Public Policy (2004). Facilitating interdisciplinary research. Washington: NationalAcademy Press, p. 2.

Newman, D., Bonilla, E. V., & Buntine, W. (2011). Improving topic coherence with regularized topic models. *Advances in Neural Information Processing Systems, 24*, 496–504.

Nichols, L. G. (2014). A topic model approach to measuring interdisciplinarity at the National Science Foundation. *Scientometrics, 100*(3), 741–754.

Paul, M., & Girju, R. (2009). Topic modeling of research fields: An interdisciplinary perspective. In: Proceedings of the International Conference RANLP-2009 (pp. 337–342)

Porter, A., & Rafols, I. (2009). Is science becoming more interdisciplinary? Measuring and mapping six research fields over time. *Scientometrics, 81*(3), 719–745.

Quan, X., Kit, C., Ge, Y., & Pan, S. J. (2015). Short and sparse text topic modeling via self-aggregation. In: Twenty-Fourth International Joint Conference on Artificial Intelligence

Ramage, D., Manning, C. D., & McFarland, D. A. (2010). Which universities lead and lag? Toward university rankings based on scholarly output. In: Proceedings of NIPS Workshop on Computational Social Science and the Wisdom of the Crowds

Ramage, D., Manning, C. D., & Dumais, S. (2011). Partially labeled topic models for interpretable text mining. In: Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining (pp. 457–465).

Rao, C. R. (1982). Diversity and dissimilarity coefficients: A unified approach. *Theoretical Population Biology, 21*(1), 24–43.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In: Proceedings of the eighth ACM international conference on Web search and data mining (pp. 399–408)

Rousseau, R. (2019). On the Leydesdorff-Wagner-Bornmann proposal for diversity measurement. *Journal of Informetrics, 13*(3), 906–907. https://doi.org/10.1016/j.joi.2019.03.015

Stirling, A. (2007). A general framework for analyzing diversity in science, technology and society. *Journal of the Royal Society Interface, 4*(15), 707–719.

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

Syed, S., & Spruit, M. (2018). Selecting priors for latent Dirichlet allocation. In: 2018 IEEE 12th International Conference on Semantic Computing (ICSC) (pp. 194–202). IEEE.

Talley, E. M., Newman, D., Mimno, D., Herr, B. W., II., Wallach, H. M., Burns, G. A., Leenders, A. G., & McCallum, A. (2011). Database of NIH grants using machine-learned categories and graphical clustering. *Nature Methods, 8*(6), 443.

Tang, J., Meng, Z., Nguyen, X., Mei, Q., & Zhang, M. (2014). Understanding the limiting factors of topic modeling via posterior contraction analysis. In: International Conference on Machine Learning (pp. 190–198)

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

Wallach, H., Mimno, D., & McCallum, A. (2009). Rethinking LDA: Why priors matter. *Advances in Neural Information Processing Systems, 22*, 1973–1981.

Wang, Q., & Schneider, J. W. (2020). Consistency and validity of interdisciplinarity measures. *Quantitative Science Studies, 1*(1), 239–263.

Wang, X., Fang, A., Ounis, I., & Macdonald, C. (2019). Evaluating similarity metrics for latent Twitter topics. *European conference on information retrieval* (pp. 787–794). Springer.

Wang, K., Sha, C., Wang, X., & Zhou, A. (2014). Based on citation diversity to explore influential papers for interdisciplinarity. In *Asia-Pacific web conference* (pp. 343–354). Springer, Cham.

Yau, C. K., Porter, A., Newman, N., & Suominen, A. (2014). Clustering scientific documents with topic modeling. *Scientometrics, 100*(3), 767–786.

Zhang, L., Rousseau, R., & Glänzel, W. (2016). Diversity of references as an indicator of the interdisciplinarity of journals: Taking similarity between subject fields into account. *Journal of the Association for Information Science and Technology, 67*(5), 1257–1265.

Zhao, H., Du, L., Buntine, W., & Liu, G. (2017). MetaLDA: A topic model that efficiently incorporates meta information. In: 2017 IEEE International Conference on Data Mining (ICDM) (pp. 635–644). IEEE

Zhou, Q., Rousseau, R., Yang, L., Yue, T., & Yang, G. (2012). A general framework for describing diversity within systems and similarity between systems with applications in informetrics. *Scientometrics, 93*(3), 787–812.

Zielinski, A. (2021). Impact of model settings on the text-based Rao diversity index. In: 18th International Conference on Scientometrics and Informetrics Conference ISSI 2021 (pp. 1405–1416).