Impact of Model Settings on the Text-based Rao Diversity Index

Andrea Zielinski[1]

[1] *andrea.zielinski@isi.fraunhofer.de*
Fraunhofer Institute for Systems and Innovation Research ISI, Breslauer Strasse 48, 76139 Karlsruhe (Germany)

Abstract
Topic models such as Latent Dirichlet Allocation (LDA) have been proved to be effective tools to discover latent topics in text collections in a data-driven way. These topics can be further utilized to investigate academic disciplines in terms of *interdisciplinarity* by means of indicators that reflect the diversity of the scientific output. This study provides a systematic analysis of model parameters that affect the diversity scores which are computed directly from the output of the LDA model.
We present an empirical study on a real data set, upon which we quantify the diversity of the research within several departments of Fraunhofer (FH) and Max Planck Society (MPG) 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 LDA-based Rao metrics. In particular, we could observe sharp fluctuations of the Rao index when varying over the number of topics. Due to its instability, it might not be a useful indicator of *interdisciplinary.*

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)[i].
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 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 and Toivanen, 2016). In particular probabilistic topic models such as 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 (Paul and Girju, 2009; Nanni et al., 2016) 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 judgements.

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 where 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 (Syed and Spruit, 2018; Chang et al., 2009). According to Tang et al. (2014), LDA's 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 according to 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 briefly discuss 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. 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[ii], balance is a function of assignments across categories and ranges between zero and one[iii], and disparity is the complement of similarity and computed pairwise between the referenced subject categories. Its value also ranges between zero and one[iv]. Yet, the bibliometric operationalization of diversity is actively discussed in the research community (Leydesdorff, 2018; Leydesdorff, 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) apply LDA topic modelling to analyze research awards issued at the National Science Foundation (NSF), inducing 1,000 latent topics from 170,000 project award descriptions. The institutional structure serves as a proxy for research disciplines and topics are assigned to the discipline in which they occur most frequently. The author observed a high variation in the topic frequency over years, while the aggregation of the topics into disciplines accounted for the temporal stability and resulted in a relatively constant score between 0.11 and 0.125.

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 and build a topic model for each corpus, varying over the number of topics (K = 10, 30, 100 and 300) and keeping the hyper-parameters fixed, in order to compute the Rao diversity scores. 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, for instance, 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, using training data with known labels in a semi-supervised setting (Ramage et al., 2010). Topic models have been often used in combination with partially supervised methods (Chuang et al., 2012). It has been shown that document regularization yields improved model performance, however requires reliable labeled data (Zhao et al., 2017).

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., 2015). 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=1,000 (Nichols, 2014) to cover the whole scientific landscape. Griffiths et al. (2004) determine the number of topics based on the log-likelihood of the data, while Nichols (2014) set the number of topics according to the number of research divisions at NSF. In practice, a higher number of 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:

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

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 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 applied to estimate the (dis)similarity $\delta(i, j)$ between topics i and j, and are generally computed from the topics' word probability distributions. 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 judgements. 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 were 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 or cosine distance 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).

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. (2009)
Distance-based metrics	*Hellinger Distance*	Aletras et al. (2014)
	Cosine	Wang et al. (2019)

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 LDA. 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 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).

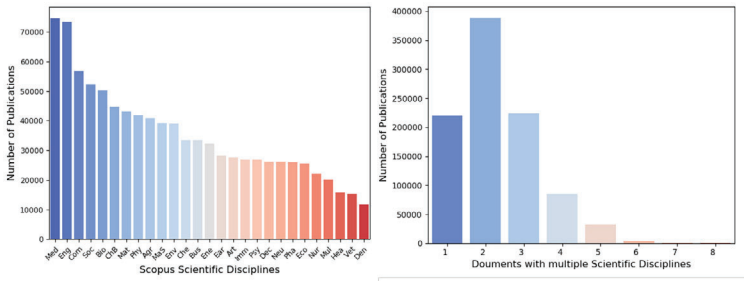


Figure 1: Statistics for Scopus Publications – *Scopus World*

In bibliometrics, the average number of subject categories of a publication, accumulated over an institute, can already serve as an indicator for interdisciplinarity (Levitt und 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 Figure 1, right).

Institute-specific Publications: Scopus FH and Scopus MPG

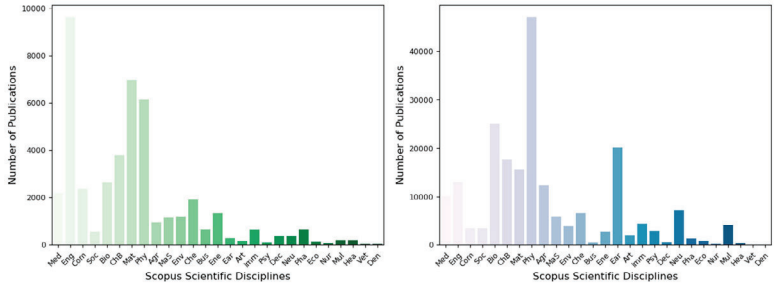


Figure 2. Statistics for *Scopus FH (left) Scopus MPG (right)*

As can be seen in Figure 2, the research profiles of FH and 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.

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

Empirical Study

Our goal is to test the effects of varying the LDA 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.

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- 2018 where we concatenate all abstracts by the same institute to obtain longer documents (with more co-occurrences) that yield better quality topics (Jónsson & Stolee, 2015). We use the institute-specific corpora *Scopus FH* and *Scopus MPG* for testing.

Model Selection and Parameter Settings

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 2,000 iterations. We set the number of topics K = 100, 150, 200, 250, 300 topics. As standard parameters of the Dirichlet

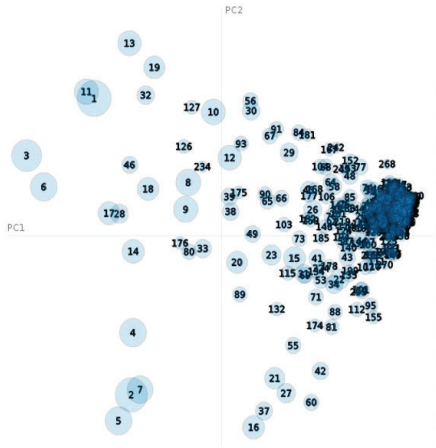
prior we use a) $\alpha = 0.1$, b) a non-uniform α estimated automatically from the data (Li et al., 2006) and c) a fixed normalized asymmetric prior of *1/K* (Wallach et al., 2009). Regarding the topics-word distributions, we set a) $\beta = 0.1$ and b) $\beta = 0.01$, unless otherwise specified.
For pre-processing, we used sentence splitting, tokenisation, lemmatization, and PoS tagging to filter all content words using the Stanford tools[v], keeping only nouns, adjectives, verbs, and foreign words that consist of alphanumeric characters. This resulted in 131,954, 12,598 and 36,381 unique words for *Scopus World*, *Scopus FH* and *Scopus MPG*, respectively.

Model Accuracy for Different Settings

We assess modelling accuracy in terms of topic coherence under various settings of hyper-parameters and number of topics. Even though determining the parameters is an established research area and various heuristics exist for real-life applications (Wallach et al., 2009; Lau et al., 2014), 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. Since LDA results are often difficult to interpret (Chang et al., 2009), we first investigate the outcome of the topic models by humans.

Topic Evaluation by Humans

A qualitative analysis of the topics by manual inspection reveals that there are topics that correspond to scientific domains and others to specific modes of discourse (e.g., description of experimental settings). Topics related to scholarly discourse - signaled by keywords such as *method*, *apply*, or *examine* - are most dominant in the corpus, relative to other topics. In these topics, authors make claims about the key contributions of their paper (Motta et al., 2000). They are more likely in the corpus, because they are relevant for all scientific disciplines.
We also wanted to see how well the topics correlate with an existing categorization schema. To this aim, we compared the topics induced by different LDA models to investigate in how much they correspond to the scientific fields in Scopus ASJC. In order to understand the distribution of topics in terms of size and their overall significance in the corpus, we use the visualization created by LDAvis (Sievert et al., 2014). It also allows assessing how well topics are separated from each other, where topics distance is based on KL divergence.



Alignment	*ASJC Category*	*Topic Words*
20,27	Engineering	temperature heat thermal pump
14,25	Materials Science	nanoparticle spectroscopy scanning coating
44	Physics & Astronomy	simulation magnet electron laser controller
12	Chemistry	ratio compound chemical derivative
47	Biochemistry	organic molecular biological fluorescence wavelength
50,193	Computer Science	application tool processing architecture automate
10,18	Medicine	patient cancer therapy tumor survival clinical
9	Chemical Engineering	reaction film synthesis catalyst conductivity
228	Energy	solar atomic newton force graft
20,148	Environmental Science	environmental wind pollution climate atmospheric
76,83	Mathematics	solution equation linear nonlinear
90	Agriculture & Biology	grow leaf fruit crop cultivar farming
145	Pharmacology	medication prescription risk pharmacy
261	Business	investment financing portfolia profitability
10	Immunology & Microbiology	tumor chemotherapy cancer patient
53,87	Social Sciences	survey questionnaire respondent
248	Decision Sciences	decision making rural household
24	Neuroscience	syndrome fatigue disease dysfunction
39	Earth & Planetary Sciences	natural seismic earthquake hazard typology landfill
26,100	Health Professions	model cell method patient
291	Arts & Humanities	artistic artist foam passive hardness
12,46	Economics	real firm sector economy bank commercial
56,187	Psychology	disorder mental disturbance psychiatric psychological
0	Nursing	patient treament analysis time test
165	Veterinary	veterinary life living mikcrobiota
25	Dentistery	implant ceramic arch crown screw dental

Figure 3. LDA Intertopic Distance Map

Figure 3 depicts how LDA topics can be aligned to ASJC codes as reference scientific domains in an overlay representation based on human classification. The topics are drawn from a relatively large *Scopus World* model that was able to uncover a high percentage of scientific topics, covering all ASJC topics (i.e., setting $K = 300$, α asymmetric, $\beta = 0.01$).
Opposed to this, models trained on *Scopus FH* or *Scopus MPG* yielded many uninterpretable and fused topics with a low coverage of ASJC topics.

Topic Coherence versus Coverage

The semantic coherence of the topics 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 coherence scores for varying model size of LDA trained on *Scopus World, Scopus FH* and *Scopus MPG.* The LDA models trained on *Scopus World* reach an average UMass score between -7.53 ($K = 100$) to -11.74 ($K = 300$) that decreases as we learn more topics. Even though LDA models trained on *Scopus FH* and *Scopus MPG*, and thus less data, achieve higher UMass scores, they are inferior to the *Scopus World* LDA model in terms of coverage (see Table 3).

Table 3. Coherence and Coverage for varying model size of LDA.

Num Topics	*100*	*150*	*200*	*250*	*300*
Scopus Worlds - Average C_{UMass}	-7.53	-8.77	-9.75	-10.90	-11.74
Scopus FH - Average C_{UMass}	-0.83	-0.91	-0.95	-0.95	-0.98
Scopus MPG - Average C_{UMass}	-3.69	-3.41	-3.29	-3.26	-3.01

Experiments to assess the different dimensions of the Rao Diversity Index

Variety & 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 on *Scopus FH* used for training and testing show that Shannon Entropy ranges between 0.001 and 1 when averaged over all topics. Results confirm 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* confirms a second hypothesis: the first setting has the effect that the probability mass of the distribution will concentrate on fewer topics per document: Accordingly, entropy values are constantly lower for all topics (see Table 4). 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.* Table 5 shows that this results in high entropy values of 0.908 (when averaged over all topic settings). More crucially, however, is the fact that in all cases Spearman's correlation is weak, and Pearson indicates only moderate correlation.

Table 4. Mean Shannon Entropy and Spearman/Pearson - Setting *α=auto/asym* and *α =0.1/ 0.01*

K	*100*	*150*	*200*	*250*	*300*	*K*	*100*	*150*	*200*	*250*	*300*
α=0.1	0.127	0.118	0.114	0.118	0.098	**S.**	0.189	0.146	0.209	0.261	0.285
α=0.01	0.072	0.072	0.072	0.072	0.072	**P.**	0.587	0.638	0.667	0.642	0.724
α=auto	0.099	0.095	0.077	0.058	0.075	**S.**	0.030	0.276	0.327	0.228	0.219
α=asym	0.113	0.067	0.067	0.067	0.067	**P.**	0.393	0.688	0.712	0.738	0.662

Table 5. Mean Shannon Entropy and Spearman/Pearson – for two LDA models (α = *auto)*

K	*100*	*150*	*200*	*250*	*300*	*K*	*100*	*150*	*200*	*250*	*300*
Scopus FH	0.099	0.095	0.077	0.058	0.075	**S.**	0.071	0.243	0.28	0.194	0.238
Scopus World	0.918	0.913	0.919	0.911	0.907	**P.**	-0.02	-0.13	0.101	0.087	0.127

Disparity Scenario: 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. Figure 5 shows the degree of semantic similarity between the topics' word distributions for 100 topics and varying β, using Jensen-Shannon as the distance measure. Pairwise dissimilarity of topics is equally high for all other investigated distance metrics, i.e. Jaccard, Hellinger, Cosine. 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. Furthermore, the data is more separated when the number of topics becomes larger for both β =0.1 / 0.01.

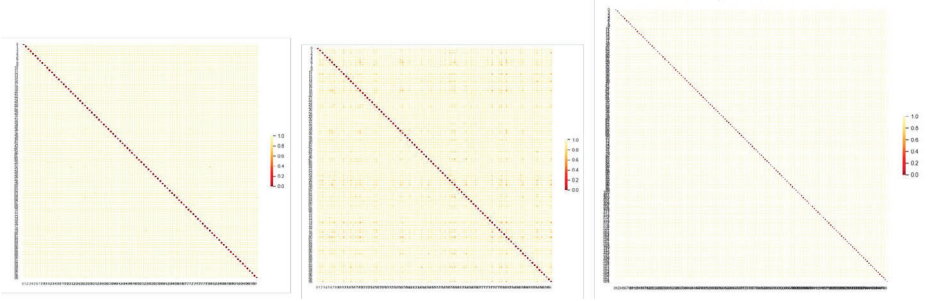


Figure 5. Pairwise Topical Distance based on JS. Topics become dissimilar when β approaches zero and the number of topics becomes larger (*K*=100, β=0.01; *K*=100, β=0.1; *K*=150, β=0.01) (left to right).

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. 6, left).

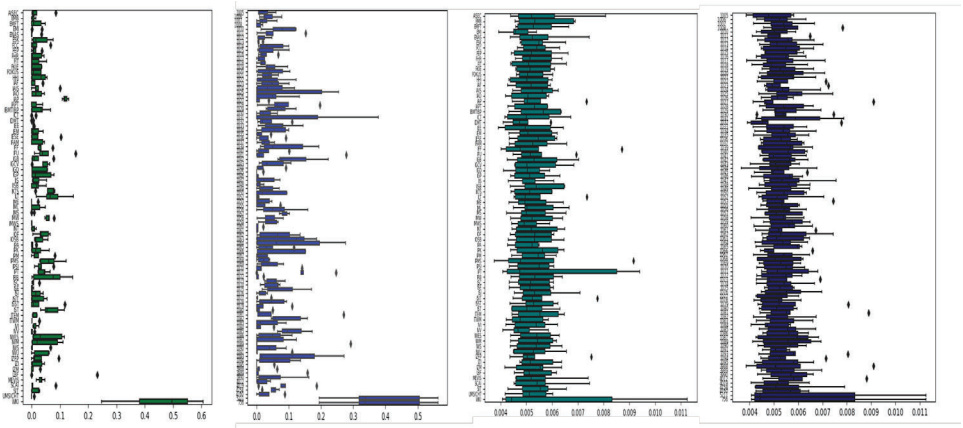


Figure 6. Rao-Index for all Fraunhofer (green) and MPG (blue) institutes;
Rao Index is computed for 100, 200, 300 topics on different LDA outputs, i.e. models are trained on *Scopus FH* versus *Scopus MPG* (left) vs. Rao Index computed on *Scopus World* (right)

Rao index values range between 0.001 to 0.605 and 0 to 0.562, with a standard deviation of 0.062 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 the *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 in section 4, 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. 6, right). The text-based Rao index thus suffers from the same limitations of low discriminating power as the bibliometric-based approach.

Last but not least, we calculated the Spearman and Pearson Rank Correlation of the Rao Index varying on the number of topic and model size. Figure 7 shows the visualization of the coefficients based on the various outputs of Rao, depicting the pairwise correlations as a heatmap. As can be seen, the choice of *K* has a great influence on the Rao results: Pairwise comparisons of Rao results vary a lot, showing that there seems to be no association between the variables. In particular, Spearman correlation is weak, showing that the general rankings amongst institutes is not preserved when varying on the number of topics.

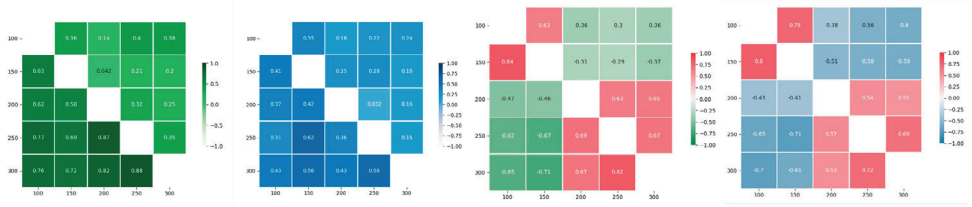


Figure 7: Spearman (upper) and Pearson (lower) Correlation of Rao Index
Computed from various LDA outputs, varying on the number of topics (on x-axis, y-axis) and model size of LDA (small models: left, large models: right), i.e. models are trained on *Scopus FH (green)* versus *Scopus MPG* (blue) vs. Rao Index computed on *Scopus World* (tested on *Scopus FH (*green-red), *Scopus MPG (*green-blue))

Conclusion

In this paper we investigated the Rao indicator for interdisciplinarity based on LDA for two German research institutes. Both institutions are specialised 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 reflecting this trend and which can be computed automatically from any data set.
Yet, our experiments show that the LDA-based Rao metrics has serious limitations and due to its instability might not be a useful indicator of interdisciplinary. Contrary to Bache (2013), who claim that the method could be applied fully automatically and would be largely invariant to the number of topics in the model, our experiments on Scopus and two major German research associations show the opposite. It results in sharp fluctuations that make it an unreliable indicator. We could not find a strong correlation between Rao results that have been generated from different settings. 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 (apart from Scientometrics).

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.

Brzezinski, M. (2015). Power laws in citation distributions: evidence from Scopus. *Scientometrics*, *103*(1), 213-228.

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).

Church, K., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational linguistics*, *16*(1), 22-29.

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 *Proc. of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 204-213).

Jónsson, E., & Stolee, J. (2015). An evaluation of topic modelling techniques for twitter.

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., & 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. (2018). Diversity and interdisciplinarity: how can one distinguish and recombine disparity, variety, and balance?. *Scientometrics*, *116*(3), 2113-2121.

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

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).

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).

Motta, E., Shum, S. B., & Domingue, J. (2000). Ontology-driven document enrichment: principles, tools and applications. *International Journal of Human-Computer Studies*, *52*(6), 1071-1109.

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

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).
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.*
Rafols, I., & Meyer, M. (2010). Diversity and network coherence as indicators of interdisciplinarity: case studies in bionanoscience. *Scientometrics*, *82*(2), 263-287.
Ramage, D., Manning, C. D., & McFarland, D. A. (2010). Which universities lead and lag? Toward university rankings based on scholarly output. In *Proc. of NIPS Workshop on Computational Social Science and the Wisdom of the Crowds*.
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).
Sievert, C., & Shirley, K. (2014). LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the workshop on interactive language learning, visualization, and interfaces* (pp. 63-70).
Stirling, A. (2007). A general framework for analysing 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, April). 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 II, B. W., Wallach, H. M., Burns, G. A., ... & 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, January). 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., ... & 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, 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.
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. In *European Conference on Information Retrieval* (pp. 787-794). 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.

[i] https://www.nsf.gov/od/oia/additional_resources/interdisciplinary_research/definition.jsp
[ii] In scopus, e.g., the classification of scientific papers is derived from the classification of journals and considers 27 top-level ASJC codes or 334 sub-level ASJC codes
[iii] Brezezinski (2015) find evidence for power laws in the citation distributions from Scopus.
[iv] The distance between categories can be calculated by the means of a matrix of citation flows between categories (Rafols and Meyer, 2010). A common measure is cosine similarity.
[v] https://stanfordnlp.github.io/CoreNLP/