Text-Based Measures of Document Diversity

Kevin Bache
Department of Computer Science
University of California, Irvine
kbache@ics.uci.edu

David Newman
Department of Computer Science
University of California, Irvine
newman@uci.edu

Padhraic Smyth
Department of Computer Science
University of California, Irvine
smyth@ics.uci.edu

ABSTRACT

Quantitative notions of diversity have been explored across a variety of disciplines ranging from conservation biology to economics. However, there has been relatively little work on measuring the diversity of text documents via their content. In this paper we present a text-based framework for quantifying how diverse a document is in terms of its content. The proposed approach learns a topic model over a corpus of documents, and computes a distance matrix between pairs of topics using measures such as topic co-occurrence. These pairwise distance measures are then combined with the distribution of topics within a document to estimate each document's diversity relative to the rest of the corpus. The method provides several advantages over existing methods. It is fully data-driven, requiring only the text from a corpus of documents as input, it produces human-readable explanations, and it can be generalized to score diversity of other entities such as authors, academic departments, or journals. We describe experimental results on several large data sets which suggest that the approach is effective and accurate in quantifying how diverse a document is relative to other documents in a corpus.

Keywords

Diversity; Interdisciplinarity

Categories and Subject Descriptors

I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.7 [**Artificial Intelligence**]: Learning

1. INTRODUCTION

The quantification of diversity has been widely studied in areas such as ecology [9], genetics [12], linguistics [8], and sociology [5]. The typical context is where one wishes to measure the diversity of a population, where a population consists of a set of individual elements that have been categorized into T types (such as species), with proportions $\pi = \{p_1, \ldots, p_T\}$ and $\sum_{i=1}^{T} p_i = 1$.

A relatively simple measure of diversity is *variety*, how many different species are present in a population, or the number of non-zero proportions in π. One can alternatively measure diversity as a function of the relative *balance* among the proportions (also referred to as 'evenness' in ecology [13] or 'concentration' in economics [4]), using measures such as Shannon entropy $H(\pi) = -\sum_{i=1}^{T} p_i \log p_i$ or variance-based quantities such as $\sum_{i=1}^{T} p_i(1-p_i) = 1 - \sum_{i=1}^{T} p_i^2$ (e.g., [20]). The intuition is that higher entropy or variance implies greater population diversity (e.g., see [19]).

From a more general perspective, Stirling [22] proposed that there are three distinct aspects to diversity: *variety*, *balance*, and *disparity*. *Disparity* is the extent to which the categories that are present are different from each other, based for example on distance within a known taxonomy [21]. For example, a population with 5 beetles and 5 elephants would be considered more diverse than a population with 5 beetles and 5 spiders, given that beetles and elephants are more taxonomically distant than beetles and spiders. Stirling argued that each of these three properties is a necessary (but non-sufficient) component in any quantitative characterization of diversity, arriving at a relatively simple mathematical formulation for diversity, a formulation originally proposed in earlier work by Rao [18]:

$$div = \sum_{i=1}^{T} \sum_{j=1}^{T} p_i p_j \;\delta(i,j) = \pi^t \Delta \pi \qquad (1)$$

where p_i, p_j are the proportions of category i and j in the population, $\delta(i,j)$ is the distance between categories i and j, Δ is a $T \times T$ matrix of such distances, and π^t is the transpose of the $T \times 1$ vector of proportions π.

This diversity measure div has a simple and intuitive interpretation as the expected distance between two randomly selected elements of the population. The probability of selecting a pair of elements with replacement from categories i and j is $p_i p_j$. Thus, div can be interpreted as the expected value of the categorical distance, $E[\delta(i,j)]$, where the expectation is with respect to the distribution of pairs of elements.

The contribution of this present paper is to investigate diversity in the context of text documents, using Rao's measure a starting point. In particular, we will use words as elements, topics as word categories, and documents as collections (or "populations") of words. Specifically, we address the following task: given a corpus of documents, assign a

KDD'13, August 11–14, 2013, Chicago, Illinois, USA.
ACM 978-1-4503-2174-7/13/08 ...$15.00.

diversity score to each document, where this diversity score can be used to rank documents from most to least diverse.

There are a number of different practical problems where quantifying the topical diversity of documents in this manner is potentially useful. One specific area of application is in science policy. There is broad interest among science policy experts in diversity and interdisciplinarity in scientific research. In particular, there is interest in the hypothesis that interdisciplinary research can lead to new discoveries at a rate faster than that of traditional research projects conducted within single disciplines. Indeed, the United States National Science Foundation (NSF) encourages interdisciplinary proposals, and has put out solicitations for proposals that include specific combinations of disciplines. One such example was the recent NSF program "Collaboration in Mathematical Geosciences" (CMG), which was focused on research at the intersection of mathematics and geoscience. In this context an automated diversity measure would be potentially helpful in evaluating the diversity of submitted proposals during the review process. Furthermore, being able to quantify the diversity of papers that resulted from funding under such a program, compared to papers funded by traditional single-discipline programs, would be useful as a component in overall evaluation of the effectiveness of interdisciplinary research programs.

Similarly in scientometrics and bibliometrics, there is significant interest in developing quantitative measures of interdisciplinarity for both individual scientific articles as well as collections of articles such as journals (e.g., [23]). Further afield, one can envision tools that allow researchers to explore and rank the diversity of individual papers and journals, and for administrators (such as department chairs, deans, and heads of research labs) to quantify the diversity of the research in their departments and labs relative to other institutions.

We begin in Section 2 by discussing related work. Section 3 outlines a number of possible diversity measures based on topic models. Section 4 describes the text corpora and the topic modeling approach we use in the paper. In Section 5 we describe a set of experiments based on pseudo-documents which serve as a proxy for ground truth and allow us to evaluate the performance of different text-based diversity measures. Section 6 discusses several examples of both high and low diversity scientific articles and grant abstracts detected by our approach, and Section 7 concludes the paper.

2. RELATED WORK

2.1 Interdisciplinarity in Scientometrics

There has been a significant amount of work in the field of scientometrics on quantifying notions of *interdisciplinarity* as reflected in the output of scientific research (e.g., via published scientific articles). The 2005 *National Academies Committee on Facilitating Interdisciplinary Research* defined interdisciplinarity from an operational viewpoint as a "mode of research that integrates .... concepts ... tools ... data ... from two or more bodies of knowledge or research practice" [15]. Diversity in this context (e.g., diversity of citations or diversity of text content) can be thought of as a broader construct than interdisciplinarity, but one which serves as a useful proxy for it. Indeed, diversity as defined via cocitation counts is the most widely-used approach to quantify interdisciplinarity in practice, based on the notion that disciplines that are co-cited more often by the same article are "closer" than disciplines that are less frequently co-cited. *Journal subject categories* are typically used to capture the notion of a *discipline*, typically using the manually-defined 244 ISI subject categories from Thomson Reuters, with articles being assigned to a subject category associated with the journal the article is published in (e.g., [15, 14, 17, 23]).

Rafols and Porter [14] used journal subject categorizations of citations to analyze how interdisciplinarity has changed between 1975 and 2005 for six specific subject-categories. They concluded that although the number of citations and co-authors per paper was increasing significantly over time, the degree of interdisciplinarity was increasing at a much slower rate, as reflected by citation patterns between subject categories. As a component in their analysis, Rafols and Porter used Rao's diversity index based on a count matrix of D documents by T categories derived from citations: p_i was the proportion of citations made by an article to other articles that were published in journals belonging to subject category i, and $\delta(i, j)$ was defined as 1 minus the cosine distance between citation count vectors (across documents) of subject categories i and j.

Our work differs from this earlier work and related threads in scientometrics in two specific ways. First, in our approach the categories and distances, $\delta(i, j)$, are learned directly from the text content, rather than being based on manually predefined schema such as the ISI subject categories. There are obvious limitations to relying on pre-defined taxonomies, as pointed out by Rafols and Porter [15]. Subject categories can change over time and no longer necessarily reflect current disciplinary boundaries. In addition, in some contexts such as analysis of proposals and grants, there may be very limited or no categorizations available. For analysis of narrow domains (say the field of data mining and machine learning) existing categorization schemes may be too coarse-grained to be useful. In this context, a corpus-driven approach to learning the categories, such as the topic-based method we describe here, is a useful alternative, and in some cases may be the only option.

The second major difference in our approach is our use of word counts rather than citation counts (which are the basis of most prior work in scientometrics on quantifying interdisciplinarity). We expect that using text content will complement citation-based approaches, as both words and citations carry useful signal. There has long been debate over whether citations accurately reflect the content of a scientific article [2, 1]—arguably the words in an article may provide a more accurate reflection of the author's intentions than the citations the author uses. A systematic approach to the use of *both* word-based and citation-based measures of diversity would also be worth exploring in future work—in this paper, however, we limit our attention to the exploration of word-based measures.

2.2 Diversity as Outlier Detection

Another field which is related to our current work is that of outlier detection. If we consider documents as being represented by T-dimensional vectors of counts, then one approach to quantifying diversity is to look for documents that are outliers in this T-dimensional space, using a multivariate outlier detection algorithm. Typically these algorithms rely on a notion of global or local density, e.g., by finding data points that have low-probability under a global distribution or that are relatively distant from their nearest neighbors.

In addition to the usual issues associated with estimating distances and densities in high dimensions, a further complication in diversity characterization is that we are seeking low-probability data points with the constraint that we are not interested in solutions where all of the probability mass is on a single component, i.e., where $p_i \approx 1, p_j \approx 0, j \neq i$. Equivalently, since the p_i are the components of a probability vector in a $T-1$ dimensional simplex, we can think of high diversity documents as points that lie in the interior of the simplex (in at least 2 of the dimensions) rather than at the edge.

Although it might be possible to develop a principled approach to characterizing diversity in this way, e.g., by a constraint-based approach to outlier detection, the use of Rao's measure bypasses both the problem of estimating a high-dimensional distribution and the problem of constraining points of interest to lie in the interior of the simplex. In particular, we can view Rao's measure as a form of outlier detection based on second-order information, focusing on pairwise dependencies among the columns of the count matrix, via the $\delta(i,j)$ term, combined with a term $p_i p_j$ that penalizes count vectors consisting of a single dominant component.

2.3 Diversity in Information Retrieval

A third potentially relevant source of prior work is in information retrieval and search where one wishes to generate a diverse list of search results in response to a user query (e.g., to avoid showing similar items in a list of search results). This work has a somewhat different motivation than the one we pursue in this paper. In the typical search context, diversity is closely aligned with making inferences about users' goals, i.e., trying to find a diverse group of documents such that the probability is maximized that at least one of the documents matches a user's implicit goals (e.g., [24]) or maximizing some notion of coverage (e.g., [6]). In contrast, the focus in this paper is on characterizing the inherent topical diversity of single documents, rather than finding a group of documents that best fulfill a user's information need.

3. DEFINING TOPIC-BASED DIVERSITY

In the general case we consider a count-matrix representation for a corpus of D documents, where each row indexed by $d, 1 \leq d \leq D$, represents a document, each column $j, 1 \leq j \leq T$, represents a category, and each entry indexed by (d,j) in the matrix represents how many elements in document d belong to category j. In particular, in this paper we focus on word tokens as the elements of a document, and a learned set of topics as the categories to which elements have been assigned.

We use the Latent Dirichlet Allocation (LDA) topic model with collapsed Gibbs sampling to learn T topics for the D documents in the corpus [7]. A single iteration of the collapsed Gibbs sampler consists of iterating through the word tokens in the corpus, sequentially sampling topic assignments for each word token in each document while keeping all other topic-word assignments fixed. Using the topic-word assignments from the final iteration of the Gibbs sampler[1] , we create a $D \times T$ *document-topic* count matrix with entries n_{dj} corresponding to the number of word tokens in document d that are assigned to topic j.

In this context we can define Rao's diversity measure for each document d as

$$div^{(d)} = \sum_{i=1}^{T} \sum_{j=1}^{T} P(i|d)P(j|d)\delta(i,j) \tag{2}$$

where $P(j|d)$ is the proportion of word tokens in document d that are assigned to topic j (estimated as $\frac{n_{dj}}{n_d}$ where n_d is the number of word tokens in d) and $\delta(i,j)$ is a measure of the distance between topic i and topic j. Note that $\delta(i,j)$ is constant across all documents, and $P(i|d)$ and $P(j|d)$ vary from document to document.

The interpretation of Equation 2 is intuitive: if we randomly select a pair of words from document d (with replacement), then $div^{(d)}$ is the *expected topical distance between a pair of words in document d.* Thus, a document that has two topics that are far away from one another, each with a large proportion of the word tokens assigned to them, will have a high diversity score. Conversely, documents whose word tokens are assigned to topics that are all relatively close to one another, or whose word tokens predominantly fall into a single topic, will earn a lower diversity score.

There are a number of possible approaches to defining distances between topics $\delta(i,j)$. We explore below a number of different pairwise measures of similarity between topics, $s(i,j)$, as well as different methods of transforming these similarities into distances. We begin with topic similarity functions based on *topic co-occurrence* in documents, as defined by the $D \times T$ matrix of *document-topic* counts. An alternative approach that we also explore is topic similarity based on the similarity of *topic-word* distributions using the $W \times T$ *word-topic* count matrix.

3.1 Topic Co-occurrence Similarity

A straightforward measure of topic similarity based on co-occurrence within documents is the cosine distance of columns in the $D \times T$ matrix of *document-topic* counts. This is defined as

$$s(i,j) \equiv \frac{\sum_d n_{di} n_{dj}}{\sqrt{\sum_d n_{di}^2}\sqrt{\sum_d n_{dj}^2}} \tag{3}$$

where i and j represent two column indices (two topics) and $\sum_d$ is a sum over all documents indexed by d.

Other similarity measures can also be used. For example, consider randomly selecting two word tokens with replacement from within a randomly selected document d in the corpus. Let $s(i,j) = P(w_1 = i, w_2 = j)$ be the probability that the first word token w_1 is assigned to topic i and the second word token w_2 is assigned to topic j:

$$\begin{aligned} P(w_1 = i, w_2 = j) &= \sum_d P(w_1 = i, w_2 = j|d)P(d) \\ &= \sum_d P(j|d)P(i|d)P(d) \end{aligned} \tag{4}$$

where $P(d)$ is the probability of a random word belonging to document d and is estimated using $\frac{n_d}{N}$ where N is the number of word tokens in the corpus. In estimating $P(j|d)$ and

[1] An alternative approach would be to average over multiple samples and use expected counts in the document-topic count matrix rather than actual counts from the final sample.

$P(i|d)$ above we use smoothed maximum a posteriori estimates, with hyperparameter values from the Dirichlet prior on the document-topic multinomials in the topic model. The use of smoothed estimates produces non-zero similarities $P(w_1 = i, w_2 = j)$ for all pairs of topics i and j, avoiding singularities in the corresponding distances $\delta(i, j)$ and diversity measures. The conditional version of the expression above, $P_C(w_2 = j|w_1 = i)$ can be viewed as a topic-based version of the contextual word distribution defined by Dillon et al. [3], defined as the probability that one word is present in a document given that another word is also in the same document.

3.2 Topic-Word Similarity

An alternative strategy to using topic co-occurrence is to consider topic similarity based on topic-word distributions. Similarity can be defined in exactly the same manner as above, but now using the $W \times T$ *word-topic* count matrix instead of the $D \times T$ *document-topic* count matrix, where W is the number of words in the model's vocabulary. In the context of measuring diversity, it is interesting to consider whether the *document-topic* or *topic-word* similarity is likely to be more useful. One can imagine situations where two topics have relatively different distributions over words (low similarity in *topic-word* distributions), yet the same two topics co-occur relatively frequently across documents (high similarity in *document-topic*). From a diversity perspective, documents that contain these two topics should in principle not be diverse, yet the *word-topic* similarity measure would indicate that they are since their word distributions are different. In our experimental results we explore this further and report results using diversities computed from both the *document-topic* (DT) and *word-topic* (WT) matrices.

3.3 From Similarity to Distance

We empirically investigated two different transformations to convert each similarity measure into a distance measure: $\delta(i, j) = 1 - s(i, j)$ and $\delta(i, j) = 1/s(i, j)$. We also investigated the effectiveness of $\delta(i, j) = -\log s(i, j)$ but found that it did not provide a performance gain over the other transformations.

4. DATA SETS AND TOPIC MODELS

4.1 Data Sets

The PubMed Central Open Access dataset (PubMed) is comprised of articles published in biomedical journals which are freely available under a creative commons license [11]. We collected approximately 228k articles which were published between the dataset's inception in 1996 and our collection date in mid-2010. We focused our efforts on a subset of approximately 165k articles for which full text was available. Each document contained a title, the name of the journal in which it was published, its year of publication, and names of its authors. We eliminated approximately 20k documents which had either fewer than 600 words or more than 10,000 words, yielding a collection of approximately 145k documents.

Our second data set is a collection of 74k NSF Awards from 2007 to 2012 gathered from www.nsf.gov/awardsearch. Each record includes the title and abstract of the award, as well as various metadata such as the NSF Directorate, Division and Program that funded the award. We eliminated approximately 12k documents which had duplicate titles, followed by an additional 10k which had fewer than 70 words or more than 1,000, resulting in a final set of 52k documents.

As a third data set we used the Association of Computational Linguistics Anthology Network (ACL) [16], consisting of papers published in selected computational linguistics conferences. This corpus contains the full-text of approximately 19k papers appearing at these conferences over a time span of more than four decades, in addition to each document's title, year, and conference of publication. We eliminated approximately 7k documents which were published as workshop papers, and an additional 1k which had fewer than 600 words or more than 10,000 words, yielding a collection of approximately 11k documents.

4.2 Topic Modeling

We performed simple tokenization and topic modeling on each of the three text corpora using MALLET [10]. This involved splitting on whitespace, removing punctuation and lowercasing, and converting into a bag-of-words representation using MALLET's default stopword list.

We then learned an LDA topic model with a fixed symmetric prior β over the word-topic distributions, and optimized the prior α over the document-topic distributions. The β prior was set to 0.01 and we initialized the α prior over the document-topic distributions at $0.05\frac{N}{DT}$, where N is the number of tokens in the dataset, D is the number of documents in the dataset, and T is the number of topics defined in the model. We enabled hyperparameter optimization every 10 iterations, and ran each Gibbs sampler for a total of 5,000 iterations, keeping only the final sample in the chain. For each dataset, we learned models with $T = 10$, 30, 100 and 300 topics.

5. PSEUDO-DOCUMENT EXPERIMENTS

5.1 Pseudo-Documents

A significant challenge in evaluation is that there is no ground-truth measure for a document's diversity. To address this problem, we created artificial 'pseudo-documents,' half of which were designed to have *high* diversity and half of which were designed to have *low* diversity.

We create each pseudo-document by combining two actual documents into one pseudo-document in the following fashion. We begin by manually selecting two journals A and B with relatively unrelated (e.g., *The Journal of Cell Biology* and *The Journal of Foot and Ankle Research*). A pseudo-document is created by randomly selecting one article from journal A and one article from journal B, which we denote as *parent documents*. A *child pseudo document* is then created by computing the average of each *parent document's* bag of topic counts, rounded to the nearest count. If the parent journals, A and B, are relatively dissimilar in content, we expect the resulting pseudo-documents to be relatively diverse. We can also create low-diversity pseudo-documents by repeating the above process but now selecting both parent articles from the same journal. By labeling pseudo-documents as having *high* or *low* diversity in this manner, we can create a proxy for ground truth diversity for evaluation purposes. This approach will not necessarily be perfect: for example, it is possible that if one of the journals contains documents that span diverse topics (relative to the corpus as a whole) some of the pseudo-documents la-

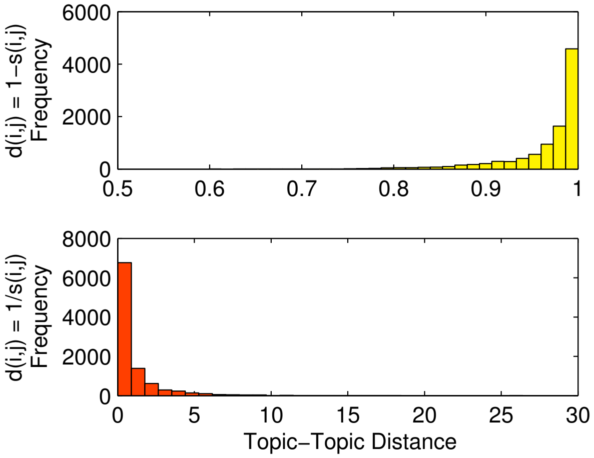


Figure 1: Histograms of topic-topic distances for $\delta(i,j) = 1 - s_c(i,j)$ and $\delta(i,j) = 1/s_c(i,j)$.

beled as low-diversity by this method could have relatively high actual diversity. However, even though such mislabeling could occur in theory, our assumption is that this pseudo-document approach will allow us to accurately measure *relative* performance across different diversity measures.

We manually selected ten pairs of journals from PubMed, where each pair appeared to have unrelated content (see Table 2 for a list of journal pairs). Using the process outlined above, for each pair of journals, we generated 50 high-diversity pseudo-documents and for each individual journal in the pair generated an additional 25 low-diversity pseudo-documents. Each parent document was drawn without replacement, meaning that no real document served as a parent of more than one pseudo-document across the entire set. This process yielded a total of 1,000 pseudo-documents, half of which were designed to have high diversity, and half of which were designed to have low diversity.

5.2 Experiments

We first tested whether our diversity scores could be used to differentiate the two classes of pseudo-documents.

We started by learning a set of topic distances on the document-topic count matrix for the 145k PubMed documents. We then used this distance matrix to assign a diversity score to each pseudo-document using the method described in section 3. We computed an area under the curve (AUC) value for the ROC curve generated from the set of diversity scores produced by our method based on the designed ground truth 'high' and 'low' diversity values for each pseudo-document.

Table 1 lists AUC values for multiple diversity formulas across topic models with 10, 30, 100, and 300 topics. Chance performance will yield AUC values of 0.50, and perfect classification accuracy will yield an AUC of 1.

First, it is clear from these results that different distance measures yield significantly different results. For example, distance measures with $\delta(i,j) = 1/s(i,j)$ perform significantly better than distance measures with $\delta(i,j) = 1-s(i,j)$ (see Table 1).

This is because $s(i,j)$ is close to 0 for most pairs of topics, with large values being on the order of 0.2. As a result, most distances are ≈ 1 when $\delta(i,j) = 1 - s(i,j)$ (see figure 1), making this method more akin to a "balance method" than Rao's diversity (as discussed in Section 1). On the other hand, when $\delta(i,j) = 1/s(i,j)$, small similarity values create very large distances, making the distance term appropriately dominant.

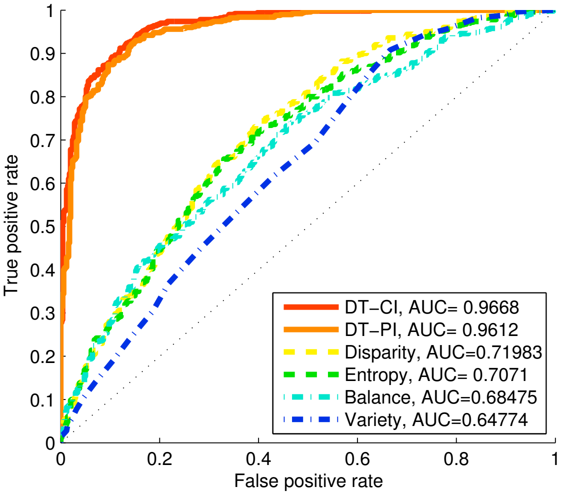


Figure 2: Pseudo-document ROC curves for PubMed data with 100 topics comparing Rao diversity to alternate methods. See also Table 1.

A second general observation from Table 1 is that distance formulas based on the *document-topic* matrix outperform distance formulas based on the *word-topic* matrix (see Table 1). This may indicate that topic co-occurrences in documents are generally more useful in characterizing diversity than are similarities in *topic-word* distributions. As mentioned in section 3.2, two topics with very different word distributions may still frequently co-occur within documents in the corpus, which is one possible explanation for why similarity based on *topic-word* distributions performs relatively poorly on this task.

A third observation is that Rao diversity significantly outperforms alternative approaches (see Figure 2 and Table 1). This supports Stirling's arguments [22] that taking each of *balance*, *variety*, and *distance* is important for measuring diversity, compared to methods such as entropy which don't take all three aspects into account.

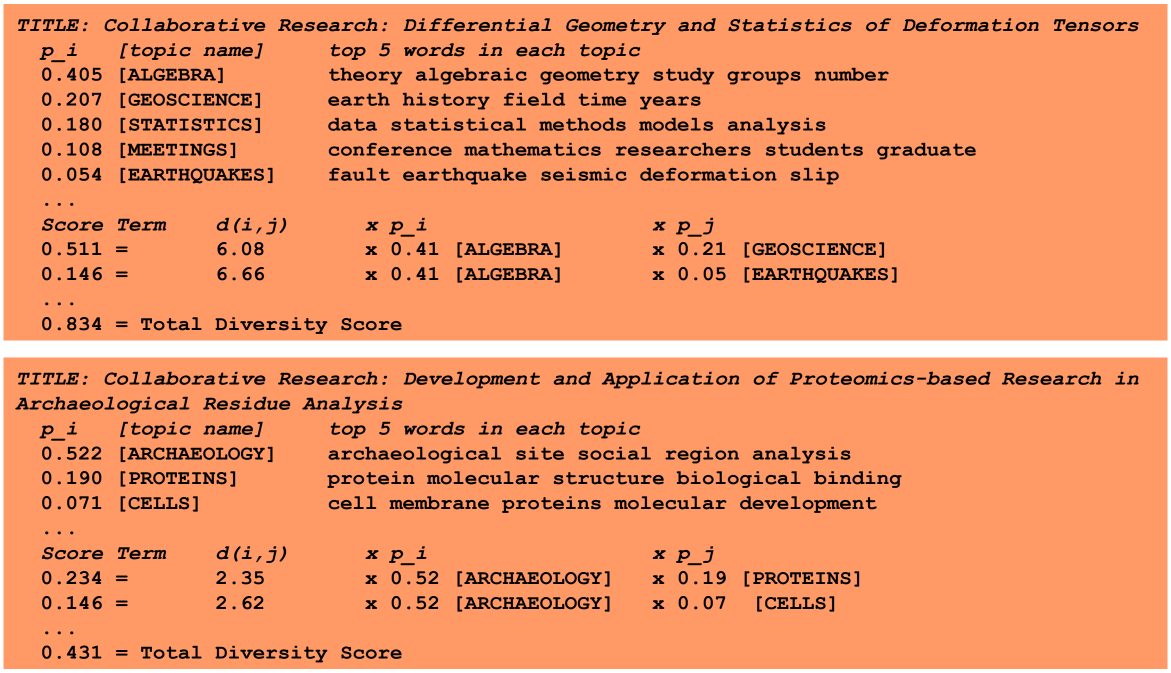
Overall, Rao diversity with the distance measures we have termed 'DT-PI' or 'DT-CI' perform the best, where DT refers to a *document-topic* based similarity measure, P to probability-based similarity, C to cosine-based similarity, and I to the inverse transformation of similarity. In addition to yielding high pseudo-document classification accuracies, these methods also appear to be largely invariant to the number of topics in the model (see Table 1), and show consistent performance across pseudo-documents drawn from different pairs of journals (Table 2). Since the 'DT-PI' and 'DT-CI' methods are very close in performance overall, we use 'DT-CI' as our default measure of diversity from this point forward.

Abbreviation	Data Matrix	$s(i,j)$	$\delta(i,j)$	10 Topics	30 Topics	100 Topics	300 Topics
DT-PI	*Document-Topic*	Probabilistic	$1/s(i,j)$	0.923	0.911	0.955	0.950
DT-CI	*Document-Topic*	Cosine	$1/s(i,j)$	**0.926**	**0.929**	**0.964**	**0.964**
DT-P	*Document-Topic*	Probabilistic	$1-s(i,j)$	0.799	0.710	0.685	0.608
DT-C	*Document-Topic*	Cosine	$1-s(i,j)$	0.842	0.770	0.772	0.716
WT-PI	*Word-Topic*	Probabilistic	$1/s(i,j)$	0.828	0.722	0.801	0.771
WT-CI	*Word-Topic*	Cosine	$1/s(i,j)$	0.856	0.805	0.814	0.689
WT-P	*Word-Topic*	Probabilistic	$1-s(i,j)$	0.798	0.709	0.685	0.608
WT-C	*Word-Topic*	Cosine	$1-s(i,j)$	0.838	0.779	0.762	0.659
Abbreviation	**Diversity Formula for Document d**			**10 Topics**	**30 Topics**	**100 Topics**	**300 Topics**
Variety	$\sum_{i=1}^{T} 1_{[p(i\|d)>0]}$			0.681	0.667	0.648	0.643
Balance	$\sum_{i,j=1}^{T} p(i\|d)p(j\|d)$			0.797	0.709	0.685	0.608
Entropy	$-\sum_{i=1}^{T} p(i\|d)\log p(i\|d)$			0.812	0.738	0.707	0.646
Disparity	$\sum_{i,j=1}^{T} 1_{[p(i\|d),p(j\|d)>0]}\delta(i,j)$; $\delta(i,j)$ as in DT-CI			0.706	0.706	0.720	0.724

Table 1: AUC scores for different diversity measures based on 1000 pseudo-documents from PubMed.

Journal Name Abbreviations	DT-PI	DT-CI	WT-PI	WT-CI	Variety	Bal	Ent	Disp
All Journal Pairs	0.955	**0.964**	0.801	0.814	0.648	0.685	0.707	0.720
Neuroimage || BMC Public Health	0.961	**0.967**	0.894	0.770	0.669	0.654	0.703	0.658
Eplasty || Plant Mthds	**0.963**	0.962	0.817	0.810	0.616	0.657	0.660	0.712
Clinical Orthp || J Nucleic Acids	**0.972**	**0.972**	0.892	0.854	0.621	0.616	0.642	0.735
J Cell Biol || J Foot, Ankle Rsrch	**0.996**	0.993	0.908	0.962	0.631	0.684	0.718	0.805
BMC Med Ethics || BMC Immnlgy	0.989	**0.997**	0.822	0.974	0.654	0.758	0.750	0.756
Intl J Emrgy Med || Intl J Nanomed	0.955	**0.978**	0.796	0.809	0.690	0.723	0.758	0.743
J Ethnbio, Ethnmed || J Expl Botny	0.962	**0.969**	0.781	0.825	0.744	0.666	0.712	0.786
Tbcco Indced Dis || Neurl Devt	0.960	**0.966**	0.840	0.888	0.713	0.723	0.735	0.812
Frntrs in Neuro || Prtcle, Fibr Txclgy	**0.888**	0.887	0.764	0.610	0.631	0.754	0.778	0.611
Thromb J || Evlnry Bioinf Online	0.984	**0.988**	0.849	0.828	0.643	0.758	0.785	0.706

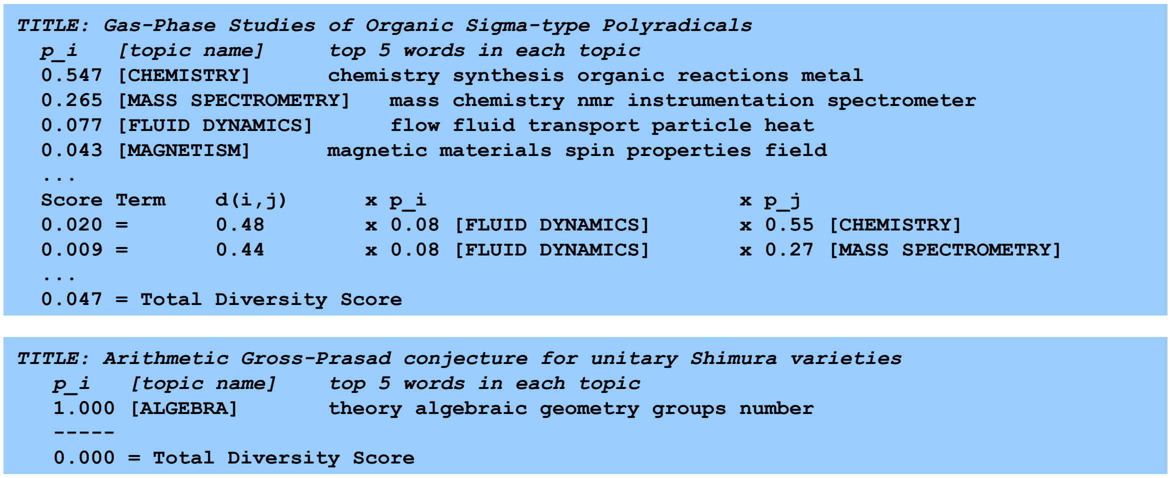
Table 2: AUC scores for pseudo-documents from specific journal pairs from PubMed.



```
TITLE: Collaborative Research: Differential Geometry and Statistics of Deformation Tensors
  p_i   [topic name]      top 5 words in each topic
  0.405 [ALGEBRA]         theory algebraic geometry study groups number
  0.207 [GEOSCIENCE]      earth history field time years
  0.180 [STATISTICS]      data statistical methods models analysis
  0.108 [MEETINGS]        conference mathematics researchers students graduate
  0.054 [EARTHQUAKES]     fault earthquake seismic deformation slip
  ...
  Score Term     d(i,j)       x p_i                   x p_j
  0.511 =        6.08         x 0.41 [ALGEBRA]        x 0.21 [GEOSCIENCE]
  0.146 =        6.66         x 0.41 [ALGEBRA]        x 0.05 [EARTHQUAKES]
  ...
  0.834 = Total Diversity Score
```

```
TITLE: Collaborative Research: Development and Application of Proteomics-based Research in
Archaeological Residue Analysis
  p_i   [topic name]      top 5 words in each topic
  0.522 [ARCHAEOLOGY]     archaeological site social region analysis
  0.190 [PROTEINS]        protein molecular structure biological binding
  0.071 [CELLS]           cell membrane proteins molecular development
  ...
  Score Term     d(i,j)       x p_i                   x p_j
  0.234 =        2.35         x 0.52 [ARCHAEOLOGY]    x 0.19 [PROTEINS]
  0.146 =        2.62         x 0.52 [ARCHAEOLOGY]    x 0.07  [CELLS]
  ...
  0.431 = Total Diversity Score
```

Figure 3: Two of the most diverse NSF grant proposals.



```
TITLE: Gas-Phase Studies of Organic Sigma-type Polyradicals
  p_i    [topic name]      top 5 words in each topic
  0.547 [CHEMISTRY]       chemistry synthesis organic reactions metal
  0.265 [MASS SPECTROMETRY]    mass chemistry nmr instrumentation spectrometer
  0.077 [FLUID DYNAMICS]       flow fluid transport particle heat
  0.043 [MAGNETISM]       magnetic materials spin properties field
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.020 =       0.48        x 0.08 [FLUID DYNAMICS]        x 0.55 [CHEMISTRY]
  0.009 =       0.44        x 0.08 [FLUID DYNAMICS]        x 0.27 [MASS SPECTROMETRY]
  ...
  0.047 = Total Diversity Score
```

```
TITLE: Arithmetic Gross-Prasad conjecture for unitary Shimura varieties
   p_i    [topic name]     top 5 words in each topic
   1.000  [ALGEBRA]        theory algebraic geometry groups number
   -----
   0.000 = Total Diversity Score
```

Figure 4: Two of the least diverse NSF grant proposals.

6. DETECTING DIVERSE DOCUMENTS

In this section we show examples of the most diverse and least diverse documents detected by our algorithm for each of our three corpora: PubMed Open Access, NSF Grant Awards, and the ACL Anthology. For each corpus we built a topic model with 100 topics, and computed diversity scores using Rao diversity with the DT-CI distance measure as defined in Table 1. We scaled the distances $\delta(i, j)$ to have a mean value of 1 within each corpus, putting the distances and diversity scores on roughly the same scale across corpora. We also manually assigned names to topics to aid in interpreting the results.

Figure 3 shows two of the most diverse NSF awards (from a corpus of approximately 52k abstracts of awards) detected by the algorithm. The first award is a collaborative research project between mathematicians and geoscientists. As shown in Figure 3, the releatively large distances (6 times larger than the mean pairwise topic distance) between ALGEBRA and each of the GEOSCIENCE and EARTHQUAKE topics drive a significant portion of the total score. The distances between these topics is reflected in the description of the project in the abstract:

> This vast mathematical theory has been applied to geology in only a few instances. This project represents collaboration between two structural geologists and a mathematician.... [It] opens the door to further cross-fertilization among geology, mathematics, and other fields.

The second of the two awards in Figure 3 is considered diverse because of the combination of the topic ARCHAEOLOGY and the two biology-related topics PROTEINS and CELLS. Again, the relatively large distances (2.4 and 2.6) between these topics and their relative strength within the document yield a particularly high diversity score for this document.

The two examples of low-diversity documents in Figure 4 tell a different story. The first grant is somewhat narrowly focused, dominated by topics that are relatively close such as CHEMISTRY, MASS SPECTROMETRY, and FLUID DYNAMICS. The second grant is an example of a document that gets a topical diversity score of 0 because all of its words are assigned to the single topic of ALGEBRA.

Figure 5 shows two the most diverse articles from the PubMed corpus. The diversity score for the first article is dominated by the combination of the PSYCHIATRY and FUNGI topics, which have a distance of 16.91 times the mean topic distance. The diversity score of the second document is largely driven by the fact that the BONES/JOINTS topic is relatively distant from each of the HIV/AIDS and VIRUSES topics. Low diversity PubMed documents showed similar patterns to low diversity NSF grants.

Finally, Figure 6 shows examples of one *high* diversity document and one *low* diversity document from the ACL corpus. The *high* diversity document achieves its score because the SUMMARIZATION topic is usually associated with text, but here it co-occurs with a set of topics related to SPEECH RECOGNITION. Thus, this paper is unusual in that it applies summarization techniques to non-text data (as indicated in the title). The other paper in Figure 6 is a typical example of a low-diversity document which is composed of a combination of topics that are very close together.

7. CONCLUSIONS

We presented an approach for quantifying the diversity of individual documents in a corpus based on their text content. Empirical results illustrated the effectiveness of the method on multiple large corpora. This text-based approach for assigning diversity scores has several potential advantages over previous alternatives, such as methods that define diversity based on citations categorized into predefined journal subject categories. The text-based approach is more data-driven, performing the equivalent of learning journal categories by learning topics from text, and can be run on any collection of text documents, even without a prior categorization scheme. In addition, it produces human-readable explanations and can be easily generalized to score the diversity of other entities such as authors, departments, or journals (e.g., by aggregating counts across such entities).

A possible direction for future work is that of temporal document diversity, for example, using topics and topic-based distance measures that only depend on documents

```
TITLE: Neuropsychiatric manifestation of confusional psychosis due to Cryptococcus
neoformans var. grubii in an apparently immunocompetent host: a case report
   p_i    [topic name]     top 5 words in each topic
  0.314  [CLINICAL MEDICINE]  patient case diagnosis lesions examination
  0.195  [PSYCHIATRY]      depression patients disorder symptoms mental
  0.131  [FUNGI]           fungal species albicans amp cbs
  0.120  [INFECTIOUS DISEASE] isolates infection tuberculosis strains resistance
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.432 =       16.91       x 0.20 [PSYCHIATRY]            x 0.13 [FUNGI]
  0.045 =       0.44        x 0.08 [PSYCHIATRY]            x 0.27 [INFECTIOUS DISEASE]
  ...
  0.598 = Total Diversity Score
```

```
TITLE: Operations about Hip in Human Immunodeficiency Virus-Positive Patients
   p_i    [topic name]     top 5 words in each topic
  0.264 [BONES/JOINTS]     bone patients joint knee fracture
  0.234 [HIV/AIDS]         hiv aids sexual infection drug
  0.189 [SURGERY]          surgery patients procedure postoperative patient
  0.043 [VIRUSES]          virus infection replication hiv influenza
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.404 =       6.53        x 0.26 [BONES/JOINTS]          x 0.23 [HIV/AIDS]
  0.047 =       4.11        x 0.26 [BONES/JOINTS]          x 0.04 [VIRUSES]
  ...
  0.541 = Total Diversity Score
```

Figure 5: Two of the most diverse PubMed OA articles.

```
TITLE: Summarizing Speech Without Text Using Hidden Markov Models
  p_i    [topic name]      top 5 words in each topic
  0.248 [SUMMARIZATION]   summary document rouge sentences content
  0.132 [SPEECH RECOGNITION]   speech recognition speaker training models
  0.089 [FINITE STATE MACHINES] state finite transducer transition automaton
  0.078 [EVALUATION]      results set precision performance score
  0.073 [PROSODY]         prosodic pitch speech phrase cue 0.417
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.136 =       7.55        x 0.25 [SUMMARIZATION]         x 0.07 [PROSODY]
  0.135 =       4.13        x 0.25 [SUMMARIZATION]         x 0.13 [SPEECH RECOGNITION]
  0.044 =       1.99        x 0.25 [SUMMARIZATION]         x 0.09 [FINITE STATE MACHINES]
  ...
  0.431 = Total Diversity Score
```

```
TITLE: Less is More: Significance-Based N-gram Selection for Smaller, Better Language Models
   p_i    [topic name]     top 5 words in each topic
   0.507 [LANGUAGE MODELS]    model language data training gram
   0.254 [PROBABILITY]    probability distribution number estimate entropy
   0.077 [ALGORITHMS]     algorithm time search number size
  ...
  Score Term    d(i,j)      x p_i                          x p_j
  0.005 =       0.04        x 0.51 [LANGUAGE MODELS]       x 0.25 [PROBABILITY]
  0.003 =       0.07        x 0.51 [LANGUAGE MODELS]       x 0.08 [ALGORITHMS]
  ...
  0.021 = Total Diversity Score
```

Figure 6: High diversity (top) and low diversity (bottom) ACL articles.

in the corpus with earlier time stamps. This would allow for distances and diversities that change over time and the detection of documents that are highly diverse relative to the time-period they were published in. An example would be early papers in bioinformatics, combining machine learning and biological concepts, which co-occur relatively frequently in the current literature but far less so 20 years ago.

Acknowledgements

KB was supported by an NSF Graduate Research Fellowship, DN was supported by NSF awards 1158699 and 1250452, and PS was supported by a Google Faculty Research Award. PS and DN were also supported by the Intelligence Advanced Research Projects Activity (IARPA) via Dept. of Interior National Business Center contract #D11PC20155. The U.S. government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

8. REFERENCES

[1] R. N. Broadus. An investigation of the validity of bibliographic citations. *Journal of the American Society for Information Science*, 34(2):132–135, 2007.

[2] D. Davies. Citation idiosyncrasies. *Nature*, 228:1356, 1970.

[3] J. Dillon, Y. Mao, G. Lebanon, and J. Zhang. Statistical translation, heat kernels and expected distances. In *Proceedings of the Uncertainty in AI Conference (UAI 2007)*, pages 93–100, 2007.

[4] M. O. Finkelstein and R. M. Friedberg. The application of an entropy theory of concentration to the Clayton act. *Yale Law Journal*, 76:677, 1966.

[5] J. Gibbs and W. Martin. Urbanization, technology, and the division of labor: International patterns. *American Sociological Review*, pages 667–677, 1962.

[6] J. Gillenwater, A. Kulesza, and B. Taskar. Discovering diverse and salient threads in document collections. In *Proceedings of the 2012 Conference on Empirical Methods in Machine Learning (EMNLP-CoNLL)*, pages 710–720, 2012.

[7] T. L. Griffiths and M. Steyvers. Finding scientific topics. *Proceedings of the National Academy of Sciences of the United States of America*, 101(Suppl 1):5228–5235, 2004.

[8] S. Lieberson. Measuring population diversity. *American Sociological Review*, pages 850–862, 1969.

[9] A. Magurran and A. Magurran. *Ecological Diversity and its Measurement*, volume 168. Princeton University Press, Princeton, NJ, 1988.

[10] A. K. McCallum. Mallet: A machine learning for language toolkit. http://www.cs.umass.edu/∼mccallum/mallet, 2002.

[11] National Center for Biotechnology Information, U.S. National Library of Medicine. Pubmed Central Open Access Initiative. 2010. http://www.ncbi.nlm.nih.gov/pmc/tools/openftlist/.

[12] M. Nei. Analysis of gene diversity in subdivided populations. *Proceedings of the National Academy of Sciences*, 70(12):3321–3323, 1973.

[13] E. C. Pielou. *An Introduction to Mathematical Ecology*. Wiley-Interscience, 1969.

[14] A. L. Porter and I. Rafols. Is science becoming more interdisciplinary? measuring and mapping six research fields over time. *Scientometrics*, 81(3):719–745, 2009.

[15] A. L. Porter, D. J. Roessner, and A. E. Heberger. How interdisciplinary is a given body of research? *Research Evaluation*, 17(4):273–282, 2008.

[16] D. Radev, P. Muthukrishnan, V. Qazvinian, and A. Abu-Jbara. The ACL anthology network corpus. *Language Resources and Evaluation*, pages 1–26, 2013.

[17] I. Rafols and M. Meyer. Diversity and network coherence as indicators of interdisciplinarity: Case studies in bionanoscience. *Scientometrics*, 82(2):263–287, 2010.

[18] C. Rao. Diversity and dissimilarity coefficients: a unified approach. *Theoretical Population Biology*, 21(1):24–43, 1982.

[19] C. Ricotta and L. Szeidl. Towards a unifying approach to diversity measures: bridging the gap between the Shannon entropy and Rao's quadratic index. *Theoretical Population Biology*, 70(3):237–243, 2006.

[20] E. Simpson. Measurement of diversity. *Nature*, page 688, 1949.

[21] A. Solow, S. Polasky, and J. Broadus. On the measurement of biological diversity. *Journal of Environmental Economics and Management*, 24(1):60–68, 1993.

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

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

[24] M. J. Welch, J. Cho, and C. Olston. Search result diversity for informational queries. In *Proceedings of the 20th International Conference on the World Wide Web (WWW)*, pages 237–246. ACM, 2011.