HiTR: Hierarchical Topic Model Re-estimation for Measuring Topical Diversity of Documents

Azarbonyad, H.; Dehghani, M.; Kenter, T.; Marx, M.; Kamps, J.; de Rijke, M.

DOI
10.1109/TKDE.2018.2874246

Publication date
2019

Document Version
Final published version

Published in
IEEE Transactions on Knowledge and Data Engineering

HiTR: Hierarchical Topic Model Re-Estimation for Measuring Topical Diversity of Documents

Hosein Azarbonyad, Mostafa Dehghani, Tom Kenter, Maarten Marx, Jaap Kamps, and Maarten de Rijke

Abstract—A high degree of topical diversity is often considered to be an important characteristic of interesting text documents. A recent proposal for measuring topical diversity identifies three distributions for assessing the diversity of documents: distributions of words within documents, words within topics, and topics within documents. Topic models play a central role in this approach and, hence, their quality is crucial to the efficacy of measuring topical diversity. The quality of topic models is affected by two causes: *generality* and *impurity* of topics. General topics only include common information of a background corpus and are assigned to most of the documents. Impure topics contain words that are not related to the topic. Impurity lowers the interpretability of topic models. Impure topics are likely to get assigned to documents erroneously. We propose a hierarchical re-estimation process aimed at removing generality and impurity. Our approach has three re-estimation components: (1) *document re-estimation*, which removes general words from the documents; (2) *topic re-estimation*, which re-estimates the distribution over words of each topic; and (3) *topic assignment re-estimation*, which re-estimates for each document its distributions over topics. For measuring topical diversity of text documents, our HiTR approach improves over the state-of-the-art measured on PubMed dataset.

Index Terms—Text diversity, topic models, topic model re-estimation

1 Introduction

Quantitative notions of measuring topical diversity of text documents are useful in a number of applications, such as assessing the interdisciplinariness of a research proposal [1] and helping to determine the interestingness of a document [2], [3], [4].

Well over three decades ago, an influential formalization of diversity was introduced in biology [5]. It decomposes diversity in terms of three central concepts: *elements* that belong to *categories* within a *population* [6]. Given a set T of categories which partitions a population d, the diversity of d is then defined as

$$div(d) = \sum_{i \in T} \sum_{j \in T} p_i^d p_j^d \delta(i, j), \tag{1}$$

where p_i^d denotes the proportion of category i in d and $\delta(i, j)$ is the distance between categories i and j, which can be calculated in a chosen manner. This notion of population diversity can be interpreted as the expected distance between two randomly selected (with replacement) elements of the population.

- *H. Azarbonyad, T. Kenter, M. Marx, and M. de Rijke are with Informatics Institute, University of Amsterdam, Amsterdam 1012, WX, The Netherlands. E-mail: {h.azarbonyad, maartenmarx, derijke}@uva.nl, tom.kenter@gmail.com.*
- *M. Dehghani and J. Kamps are with the Institute for Logic, Language, and Computation, University of Amsterdam, Amsterdam 1012, WX, The Netherlands. E-mail: {dehghani, kamps}@uva.nl.*

Manuscript received 21 July 2017; revised 17 Aug. 2018; accepted 18 Sept. 2018. Date of publication 5 Oct. 2018; date of current version 4 Oct. 2019. (Corresponding author: Hosein Azarbonyad).
Recommended for acceptance by J.-W. Wen.
For information on obtaining reprints of this article, please send e-mail to: reprints@ieee.org, and reference the Digital Object Identifier below.
Digital Object Identifier no. 10.1109/TKDE.2018.2874246

Bache et al. [1] have adapted the biological notion of population diversity to quantify the topical diversity of a text document. For measuring the topical diversity of a text document, words are considered elements, topics are categories, and a document is a population. When using topic modeling for measuring topical diversity of a text document d, [1] estimate elements based on the probability of a word given the document ($P(w|d)$), categories based on the probability of a word given a topic ($P(w|t)$), and populations based on the probability of a topic given the document ($P(t|d)$).

In probabilistic topic modeling, at estimation time, these distributions are usually assumed to be sparse. First, the main content of documents is assumed to be generated by a small subset of words from the vocabulary (i.e., $P(w|d)$ is sparse). Second, each topic is assumed to contain only some topic-specific related words (i.e., $P(w|t)$ is sparse). Finally, each document is assumed to deal with a few topics only (i.e., $(P(t|d)$ is sparse). When approximated using currently available methods, however, $P(w|t)$ and $P(t|d)$ often turn out to be dense rather than sparse [7], [8], [9]. Dense distributions cause two important problems for the quality of topic models: *generality* and *impurity*. General topics mostly contain general words. They are typically assigned to most of the documents in a corpus. In other words, the $P(t|d)$ distributions are not document-specific. Impure topics contain words that are not related to the topic. These impure words are mostly general words. Generality and impurity of topics are problematic when estimating topical diversity of text documents since they both result in low quality $P(t|d)$ distributions. Recall that these are core to the topical diversity score based on the biological notion of diversity (Equation (1)).

To improve the measurement of topical diversity of text documents we propose a hierarchical way of making the three distributions $P(w|d)$, $P(w|t)$ and $P(t|d)$ more sparse.

To this end we re-estimate the parameters of these distributions so that general, collection-wide items are removed and only salient items are kept. For the re-estimation, we use the concept of parsimony [10] to extract only essential parameters of each distribution.

We evaluate the performance of the proposed hierarchical re-estimation method for measuring topical diversity of text documents and compare our approach against the state-of-the-art [7]. In doing so, we answer our main *research question*:

How effective is our hierarchical re-estimation approach in measuring topical diversity of documents? How does its effectiveness compare to the state-of-the-art in addressing the general and impure topics problem? Are the thus improved topic models also successfully applicable in other tasks?

Our main contributions are:

1) We propose a hierarchical re-estimation process for topic models to address the two main problems in estimating the topical diversity of text documents, using a biologically inspired definition of diversity.
2) We study each level of re-estimation individually in terms of efficacy in solving the general topics problem, the impure topics problem, and improving the accuracy of estimating the topical diversity of documents.
3) We study the impact of re-estimation parameters on the statistics of documents and its relation to the quality of trained topic models and recommend effective settings of these parameters.

As an additional contribution, we also make the source code of our topic model re-estimation method available to the research community to further advance research in this area.[1]

2 Related Work

Our hierarchical topic model re-estimation touches on research in multiple areas. We review work in four directions: improving the quality of topic models, measuring text diversity, evaluating topic models, and parsimonization.

2.1 Improving the Quality of Topic Models

Topic models are effective for modeling text documents and expressing the contents of text documents in a low-dimensional space [11]. Although topic models like Latent Dirichlet Allocation (LDA) are powerful tools for modeling data in an unsupervised fashion, they suffer from different issues, especially when dealing with noisy data [12]. As mentioned already, the two most important issues with topic models are the *generality problem* and the *impurity problem* [7], [8], [9], [12]. These problems with topic models have a negative influence on the performance of tasks in which topic models are applied besides document diversity, namely document clustering, document classification, document summarization, information retrieval, sentiment analysis (see [12] for an overview).

1. The source code is available here: https://github.com/HoseinAzarbonyad/HiTR

Wallach et al. [8] propose asymmetric Dirichlet priors to construct a general topic and assign general terms to this general topic in the learning process. Similar ideas to improve the quality of topic models have been employed by others [13], [14]. Similar to [8], [13], [14], one of our goals is to address the generality problem. The main difference, however, is that they do not aim to address the two issues mentioned with topic models directly and the topic representations and topic word distributions that they arrive at are neither parsimonious nor sparse. That is, in their approach, each topic could still have a non-zero assignment probability to each document. We hypothesize that parsimony is essential in topic modeling, since it is expected that each document only focuses on a few topics [7] and in contrast to the work cited above our goal is to achieve this parsimony.

Soleimani and Miller [7] propose parsimonious topic models (PTM) to address the generality and impurity problems. A shared topic is created and general words are assigned to this topic. PTM achieves state-of-the-art results compared to existing topic models. We also address the generality and impurity problems with topic models. The background language model in our model and the shared topic in PTM have similar functionalities. They both are used to handle and remove generality from topic-word distributions. However, in PTM, the shared topic is more complicated as for each word there are a few more parameters to be estimated:

1) whether a word is topic-specific for each topic and
2) probability of being topic-specific under each topic for each word.

In our approach, we model all this using a background language model with much fewer parameters. Moreover, we model and remove the generality in three different levels: document-word distribution, topic-word distribution, and document-topic distribution. PTM handles the generality in topic-word and document-topic distributions and does not handle the generality in document-word distribution explicitly.

2.2 Evaluating Topic Models

Topic models are usually evaluated either intrinsically, for example, in terms of their generalization capabilities, or extrinsically in terms of their contribution to external tasks [15]. We focus on extrinsic evaluations of the effectiveness of our re-estimation approach. Our main evaluation concerns its effectiveness in measuring the topical diversity of text documents. In addition, in Section 7, we analyze the effectiveness of our re-estimation approach in removing impurity from documents in terms of purity in document clustering and document classification tasks.

Specifically, in the document classification task, topics are used as features of documents with values $P(t|d)$. These features are used for training a classifier [7], [16], [17]. In the document clustering task, each topic is considered a cluster and each document is assigned to its most probable topic [16], [18]. For the analyses in Section 7, following common practice (e.g., [16], [19], [20]), we use Purity and Normalized Mutual Information in the clustering task, and Accuracy as our prime evaluation metric in the classification task. Furthermore, the quality of topic models can be measured by the quality of the term distributions per topic, in terms of

topic coherence [16], [20], and by having their interpretability judged by humans [21], [22].

2.3 Text Diversity and Interestingness

Prior to [1], measuring topical diversity of documents had not been studied comprehensively from a text mining perspective. Bache et al. [1] use Rao's diversity score (Equation (1)) [5] to quantify diversity of text documents by means of LDA topic models [11]. In their framework, the diversity of a document is proportional to the number of dissimilar topics it covers. Similar to [1], [4] define the diversity of documents by means of topic models, but instead of Rao's measure they use an information theoretic diversity measure based on the Kullback Leibler divergence. Azarbonyad et al. [2] also use Rao's diversity measure to quantify the diversity of political documents and analyze the correlation of topical diversity and interestingness over political documents. Their main finding, however, is different from [4]'s conclusion, as they conclude that although in general topical diversity and interestingness of political documents are somehow correlated, a text's topical diversity does not necessarily reflect its interestingness.

2.4 Model Parsimonization

Parsimonization refers to the process of extracting essential elements of a distribution and removing superfluous, general information. Parsimonization can be considered an unsupervised feature selection approach. The idea is to extract features containing information about samples and remove features that are not informative for explaining the samples [23], [24]. Because our hierarchical re-estimation process builds on parsimonious language models (PLMs) [10], we briefly review them.

PLMs were introduced in an information retrieval setting, in which language models are used to model documents as distributions over words. The goal of parsimonization in this context is to extract words that reflect the content of documents and remove collection-specific general words [25]. To extract salient document-specific words for each document, some studies define a layered language model of documents where the language model of a document is composed of a general background model and a document-specific language model [26], [27], [28]. The Expectation-Maximization (EM) algorithm is employed to estimate the parameters of such models. Using this idea, [10] propose a method for parsimonizing document language models with the aim of removing general words by pushing the probabilities of the words that are well explained by the background model toward zero. We employ this approach for re-estimating and refining topic models.

Here we briefly recall the formal principles underlying PLMs. The main assumption is that the language model of a document is a mixture of its own specific language model and the language model of the collection:

$$P(w|d) = \lambda P(w|\tilde{\theta}_d) + (1-\lambda)P(w|\theta_C), \tag{2}$$

where w is a term, d a document, $\tilde{\theta}_d$ the document specific language model of d, θ_C the language model of the collection C, and λ is a mixing parameter ($0 \leq \lambda \leq 1$). The main goal is to estimate $P(w|\tilde{\theta}_d)$ for each document. Language model parsimonization is an iterative EM algorithm in which the initial parameters of the language model are the parameters of the standard language model, estimated using maximum likelihood:

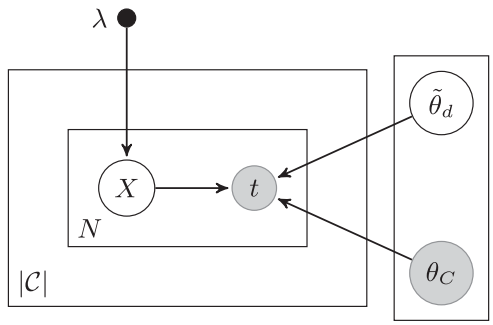


Fig. 1. Plate diagram of PLM. X corresponds to e_w in Equation (3).

Initialization:

$$P(w|\tilde{\theta}_d) = \frac{tf_{w,d}}{\sum_{w' \in d} tf_{w',d}},$$

where $tf_{w,d}$ is the frequency of w in d. The following steps are done in each iteration of the algorithm:

E-step:

$$e_w = tf_{w,d} \cdot \frac{\lambda P(w|\tilde{\theta}_d)}{\lambda P(w|\tilde{\theta}_d) + (1-\lambda)P(w|\theta_C))}, \tag{3}$$

M-step:

$$P(w|\tilde{\theta}_d) = \frac{e_w}{\sum_{w' \in d} e_{w'}}, \tag{4}$$

where $\tilde{\theta}_d$ is the parsimonized language model of document d, which is initialized by the language model of d, C is the background collection, $P(w|\theta_C)$ is estimated using maximum likelihood estimation, and λ is a parameter that controls the level of parsimonization. A low value of λ will result in a more parsimonized model while $\lambda = 1$ yields a model without any parsimonization. The E-step gives high probability values to terms that occur relatively more frequently in the document than in the background collection, while terms that occur relatively more frequently in the background collection get low probability values. In the M-step the parameters are normalized to form a probability distribution again. After this step, terms that receive a probability lower than a predefined *threshold* are removed from the model. The EM process will stop after a fixed number of iterations or when the models $\tilde{\theta}_d$ do not change significantly anymore.

PLM is a two-topic mixture model (the graphical model is shown in Fig. 1, as can be seen θ_C is considered as an external observation and the goal is to estimate $\tilde{\theta}_d$ given θ_C and λ). In that sense, PLM is similar to an LDA model with two topics (general and specific topics). However, its mechanism is different than LDA. In LDA, all topics are shared among documents and only the proportions of topics (document-topic distributions) are different for different documents. In PLM, there is a general topic which is shared among all documents, but there is a specific topic for each document which is not shared with other documents.

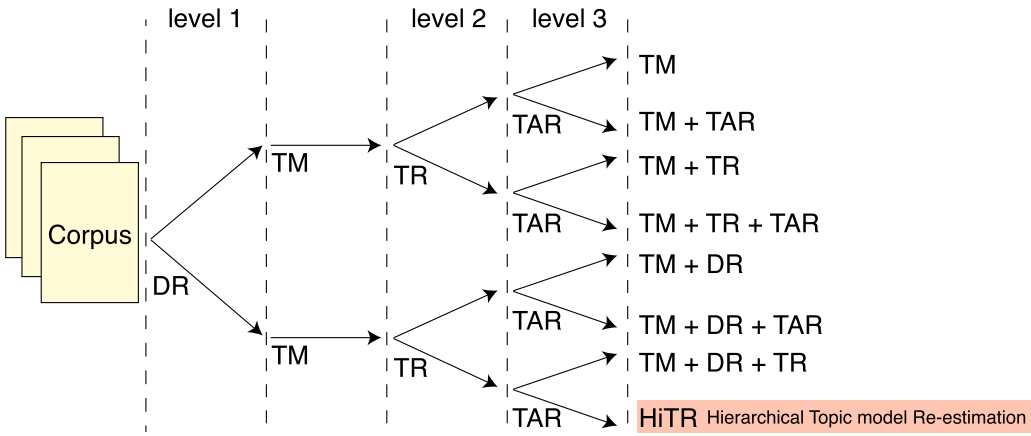


Fig. 2. Different topic re-estimation approaches. TM is a topic modeling approach like, e.g., LDA. DR is document re-estimation, TR is topic re-estimation, and TAR is topic assignment re-estimation.

Moreover, in PLM, the λ controls the proportion of general and specific topics in documents and it is fixed.

3 Hierarchical Topic Model Re-Estimation

In this section, we describe HiTR (*hi*erarchical *to*pic model *r*e-estimation). HiTR can be applied on top of any topic modeling approach that has two main components, $P(w|t)$ and $P(t|d)$ distributions.

3.1 Overview

The input of HiTR is a corpus of text documents. The output is a probability distribution over topics for each document in the corpus.

As explained in the introduction, the quality of topic models such as LDA is highly dependent on the quality of the $P(w|d)$, $P(w|t)$, and $P(t|d)$ distributions. However, generality and impurity of these distributions cause the poor quality of topic models. To solve these issues, we propose to apply re-estimation at three levels:

document re-estimation (DR) re-estimates the language model per document $P(w|d)$

topic re-estimation (TR) re-estimates the language model per topic $P(w|t)$

topic assignment re-estimation (TAR) re-estimates the distribution over topics per document $P(t|d)$

Based on applying or not applying re-estimation at different levels, there are 8 possible re-estimation approaches. Fig. 2 gives a graphical overview of the different levels of re-estimation and how they are combined. *Hierarchical topic model re-estimation* (HiTR) refers to the model that uses all three re-estimation techniques, i.e., DR+TR+TAR that can be applied to any topic model TM.

To summarize, HiTR works as follows: we first do the DR step, then train a topic model (TM step) on top of the re-estimated documents. Afterwards, we apply the TR step on the trained topic model and use the re-estimated topic model (the topic model achieved after TR step) to assign topics to documents. Finally, we apply the TAR step to topics assigned to the documents using the re-estimated topic model. We follow this order of re-estimation for two reasons: first, for the topical diversity task we only use the document-topic distributions. And second, this order provides the maximum amount of re-estimation in the final document-topic distribution because at each step of re-estimation impurity and generality is removed from document-word and topic-word distributions and finally the remaining impurity and generality is removed using TAR. Next, we describe each of the re-estimation steps in more detail.

3.2 Document Re-Estimation

The first level of re-estimation is *document re-estimation*, which re-estimates $P(w|d)$. The main intuition behind this level of re-estimation is to remove unnecessary information from documents before training topic models. This is comparable to pre-processing steps such as removing stopwords and high and low frequency words, that are typically carried out prior to applying topic models [11], [16], [19], [20], [29]. Proper pre-processing of documents, however, takes lots of effort and involves tuning several parameters, such as the number of high-frequent words to remove, if stopwords should be removed or not, whether rare words should be removed or not, whether IDF values should be considered in removing general/rare words. When dealing with a large document collection, finding optimum values for all of these parameters is non-trivial, while blindly removing words from documents without considering the distribution of them over documents could lead to missing important words and losing important information.

To solve this issue and pre-process documents automatically, we propose *document re-estimation*. After document re-estimation, we can train any standard topic model on the re-estimated documents. If general words are absent from (re-estimated) documents, we expect that the trained topic models will not contain general topics. Moreover, document re-estimation removes impure elements (general words) from documents, which will lead to more pure topics. Hence, document re-estimation is expected to address both the general topic and the impure topic problem.

Document re-estimation uses the parsimonization method described in Section 2.4. The parsimonized model $P(w|\tilde{\theta}_d)$ in Equation 4 is used as the language model of document d, and after removing unnecessary words from d, the frequencies of the remaining words (words with $P(w|\tilde{\theta}_d) > 0$) are re-calculated for d using the following equation:

$$tf(w,d) = \lfloor P(w|\tilde{\theta}_d) \cdot |d| \rfloor,$$

where $|d|$ is the document length in words. Topic modeling is then applied on the recalculated document-word frequency matrix.

3.3 Topic Re-Estimation

The second level of re-estimation is *topic re-estimation*, which re-estimates $P(w|t)$ by removing general words from it. The re-estimated distributions from this step are used to assign topics to documents.

The goal of this step is to increase the purity of topics by removing general words that have not yet been removed by document re-estimation. It is known from the literature [7], [8], [9], [12] that some topics extracted by means of topic models are impure and contain general words.

The two main advantages of applying TR are that

1) it results in more pure topics which are more interpretable by human, and
2) after getting pure, topics are less likely to be assigned to documents erroneously.

A topic is modeled as a distribution over words, which is itself a language model. Our main assumption is that each topic's language model is a mixture of its topic-specific language model and the language model of the background collection. The goal of TR is to extract a topic-specific language model for each topic and remove the part which can be explained by the background model. Given a set of topics T, background language model θ_T, and for each $t \in T$, a topic-specific language model $\tilde{\theta}_t$, we initialize $P(w|\tilde{\theta}_t)$ and $P(w|\theta_T)$ as follows:

$$P(w|\tilde{\theta}_t) = P(w|\theta_t^{\mathcal{TM}})$$
$$P(w|\theta_T) = \frac{\sum_{t \in T} P(w|\theta_t^{\mathcal{TM}})}{\sum_{w' \in V_T} \sum_{t' \in T} P(w'|\theta_{t'}^{\mathcal{TM}})},$$

where $P(w|\theta_t^{\mathcal{TM}})$ is the probability of w belonging to topic t estimated by a topic model $\mathcal{TM}$, and V_T is the set of all words occurring in all topics. Having these estimations, the steps of TR are similar to the steps of PLM, except that in the E-step we estimate $tf_{w,t}$ (the frequency of w in t) using $P(w|\theta_t^{\mathcal{TM}})$.

3.4 Topic Assignment Re-Estimation

The third and final level of re-estimation is *topic assignment re-estimation* which re-estimates $P(t|d)$.

In topic modeling, most topics are usually assigned with a non-zero probability to most of documents. When documents are typically focused on just a few topics, this is an incorrect assignment, as topics should only be assigned to documents that deal with them. General topics assigned to a majority of documents are uninformative. The goal of TAR is to address the general topics problem and achieve more document specific topic assignments.

To re-estimate topic assignments, a topic model is first trained on the document collection. This model is used to assign topics to documents based on the proportion of words in common between them. We then model the distribution over topics per document as a mixture of its document-specific topic distribution and the topic distribution of the entire collection. The goal of TAR is to extract the document-specific topic distribution for each document and remove general collection-wide topics from them.

We initialize the document-specific topic distribution $P(t|\tilde{\theta}_d)$ and the distribution of topics in the entire collection C, $P(t|\theta_C)$ as follows:

$$P(t|\tilde{\theta}_d) = P(t|\theta_d^{\mathcal{TM}})P(t|\theta_C) = \frac{\sum_{d \in C} P(t|\theta_d^{\mathcal{TM}})}{\sum_{t' \in T} \sum_{d' \in C} P(t'|\theta_{d'}^{\mathcal{TM}})}.$$

Here $P(t|\theta_d^{\mathcal{TM}})$ is the probability of assigning topic t to document d estimated by the topic model $\mathcal{TM}$. The remaining steps of TAR follow the ones of PLM. The only difference is that in the E-step, we estimate $f_{t,d}$ using $P(t|\theta_d^{\mathcal{TM}})$.

4 Evaluating HiTR

To evaluate the performance of our approach to topical diversification, we follow the evaluation setup introduced in [1]. Our main research question is:

RQ1 How effective is our hierarchical re-estimation approach in measuring topical diversity of documents? How does its effectiveness compare to the state-of-the-art in addressing the general and impure topics problem? Are the thus improved topic models also successfully applicable in other tasks?

To address RQ1 we run our models on a binary classification task. We generate a synthetic dataset of documents with high and low topical diversity (the process is detailed in Section 5.2), and the task for every model is to predict whether a document belongs to the high or low diversity class. We employ HiTR to re-estimate topic models and use the re-estimated models for measuring topical diversity of documents. We compare our method to LDA (as also used in [1] for the same purpose) and to the state-of-the-art parsimonious topic models PTM [7]. The results of experiments regarding RQ1 are discussed in Section 6.1. Moreover, we evaluate the performance of HiTR in document clustering and classification tasks and analyze its effectiveness in these tasks. The results of these experiments are described in Section 7.

Additionally, to gain deeper insights into how HiTR performs, we conduct a separate analysis of each level of re-estimation, DR, TR and TAR and answer the following research questions:

RQ2 What is the effect of DR on the quality of topic models? Can DR replace manual pre-processing?
RQ3 Does TR increase the purity of topics? And if so, how does using the more pure topics influence the performance in topical diversity task?
RQ4 How does TAR affect the sparsity of document-topic assignments? And what is the effect of the achieved parsimonized document-topic assignments on the topical diversity task?

RQ2 concerns the effectivenes of DR in removing general words from documents and its effect on the quality of topic models. To answer RQ2, we train LDA models with and without manual pre-processing and with and without DR. We compare the quality of models achieved using different combinations. This will show how effective is DR in pre-

processing documents automatically. Moreover, we measure corpus statistics such as vocabulary size, average type-token ratio, average document length after running DR with different parameters. We train LDA models on the corpora achieved with different parameters and measure the quality of trained models. Then, we analyze the correlation of corpus statistics achieved from DR with different parameters and the quality of models trained on them. In Section 6.2.1, the results regarding RQ2 are described.

To answer RQ3, we first evaluate the performance of TR on the topical diversity task and compare its performance to DR and TAR. We focus on its effectiveness in removing impure words from topics and perform a qualitative analysis on topic models before and after running TR. The results of experiments regarding RQ3 are discussed in Section 6.2.2.

To answer RQ4, we first evaluate TAR together with LDA in a topical diversity task and analyze its effect on the performance of LDA to study how successful TAR is in removing general topics from documents. The results of this experiment are presented in Section 6.2.3.

5 Topical Diversity with HiTR

In this section, we discuss the experimental setup for the topical diversity test.

5.1 Topical Diversity Measure

After re-estimating words distributions in documents, topics, and document-topic distributions using HiTR, we use the final distributions over topics per document for measuring topical diversity. Diversity of texts is computed using Rao's coefficient (Equation (1)). For each topic x, observed in corpus C, we construct a vector V_x of length $|C|$ (the number of documents in the corpus). Each entry of this vector corresponds to a document d_y and its value is assigned as: $V_x[y] = p_x^y$. We use the normalized angular distance for measuring the distance between topics, since it is a proper distance function [2]:

$$\delta(i,j) = \frac{ArcCos(CosineSim(V_i, V_j))}{\pi},$$

where $CosineSim(\cdot,\cdot)$ is the cosine similarity between two vectors, and $ArcCos(\cdot)$ is the arc cosine. We use the distributions over topics per document for calculating the distance between topics. There are two possible approaches for measuring the topic distance: based on document-topic distributions or topic-word distributions. From a diversity perspective, document-topic distributions are more suitable for this task. For example, consider two topics which co-occur frequently in documents but have different topic-word distributions. In principle, if a document contains these topics, it should not be diverse, but since the topic-word similarity of these two topics is low the document will have a high diversity.

5.2 Dataset

We use the PubMed abstracts dataset [30] in our experiments. This dataset contains articles published in bio-medical journals. We use the articles published between 2012 to 2015 for training topic models. This subset contains about 300,000 documents. Following [1], we generate 500 documents with a high value of diversity and 500 documents with a low value of diversity. We create high diversity documents as follows: we first randomly select 10 pairs of journals. Each pair contains two journals that are relatively unrelated to each other (we select 20 journals in total). For each pair of journals A and B, we select 50 articles to create 50 new probability distributions over topics as follows: we randomly select one article from A and one article from B and generate a document by averaging the selected articles' bag of topic counts. In this way, for each pair of journals we generate 50 documents with a high diversity value. We create low diversity documents as follows: for each of the chosen 20 journals, we perform a similar procedure but instead of choosing articles from two different journals, we select them from the same journal and generate 25 non-diverse documents. In the final set we have 500 diverse and 500 non-diverse documents.

5.3 Baselines

Our baseline for the topical diversity task is the method proposed in [1], which uses LDA for measuring topical diversity of documents. As an additional baseline, we use PTM [7] instead of LDA for measuring topical diversity. PTM is the state-of-the-art in topic modeling approaches, and based on our results PTM is more effective than the method proposed in [1]. Thus, PTM is our main baseline in this task.

5.4 Metrics

To measure the performance of topic models in the topical diversity task, we follow [1] and report ROC curves and AUC values. As another evaluation measure, we report the *sparsity* of topic models: the average number of topics assigned to the documents of a corpus [7]. This measure reflects the ability of topic models to achieving sparse $P(t|d)$ distributions. We also measure the *coherence* of the extracted topics. This measure indicates the purity of $P(w|t)$ distributions and a high value of coherence implies high purity within topics. For estimating the coherence of a topic model we use a reference corpus. As our reference corpus, we use a version of English Wikipedia.[2] We estimate the coherence of a topic model using normalized pointwise mutual information between the top N words within a topic using the following equation [16], [20]:

$$NPMI(T) = \sum_{t \in T} \sum_{w_i, w_j \in topN(t) \wedge i < j} \frac{\log \frac{P(w_i,w_j)}{P(w_i)P(w_j)}}{-\log\left(P(w_i,w_j)\right)}, \quad (5)$$

where T is the set of extracted topics, $topN(t)$ is the top N most probable words within topic t. w_i is a word, $P(w_i, w_j)$ is estimated based on the number of documents in which w_i and w_j co-occur divided by the number of documents in the reference corpus. $P(w_i)$ is estimated similarly, using maximum likelihood estimation.

5.5 Preprocessing

We first lowercase all the text in the corpus. Then, we remove the stopwords included the standard stop word list from Python's NLTK package. In addition, we remove the

2. We use a dump of June 2, 2015, containing 15.6 million articles.

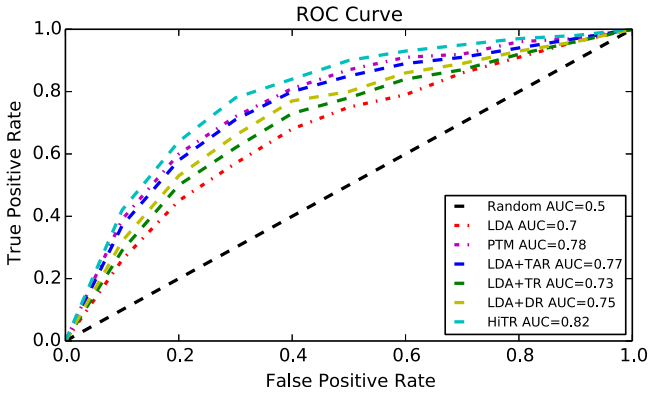


Fig. 3. Performance of topic models in topical diversity task on the PubMed dataset. The improvement of HiTR over PTM is statistically significant (p-value < 0.05) in terms of AUC.

100 most frequent words in the collection and words with fewer than five occurrences.

5.6 Model Parameters

As noted above, the topic modeling approach used in our experiments with HiTR is LDA. Following [1], [7], [31] we set the number of topics to 100. We set the two hyperparameters to $\alpha = 1/T$ and $\beta = 0.01$, where T is the number of topics, following [16]. In the re-estimation process, at each step of the EM algorithm, we set the threshold for removing unnecessary components from the model to 0.01 and remove terms with an estimated probability less than this threshold from the language models, as in [10].

We perform 10-fold cross validation, using 8 folds as training data, 1 fold as development set to tune the parameters, and 1 fold for testing.

5.7 Statistical Significance

For statistical significance testing, we compare our methods to PTM using paired two-tailed t-tests with Bonferroni correction. To account for multiple testing, we consider an improvement significant if: $p \leq \alpha/m$, where m is the number of conducted comparisons and α is the desired significance. We set $\alpha = 0.05$. In Section 6, $^{\blacktriangle}$ and $^{\blacktriangledown}$ indicate that the corresponding method performs significantly better and worse than PTM, respectively.

6 Results

In this section, we first present the results of HiTR in topical diversity task. Then, we analyze each individual level of re-estimation.

6.1 Topical Diversity Results

Fig. 3 plots the performance of our topic models across different levels of re-estimation, and the models we compare to, on the PubMed dataset. We plot ROC curves and compute AUC values. To plot the ROC curves we use the diversity scores calculated for the generated pseudo-documents with diversity labels. HiTR improves the performance of LDA by 17 percent and PTM by 5 percent in terms of AUC. From Fig. 3 two observations can be made.

First, HiTR benefits from the three re-estimation approaches it encapsulates by successfully improving the quality of estimated diversity scores. Second, the performance of LDA+TAR, which tries to address the generality problem, is higher than the performance of LDA+TR, which addresses impurity. General topics have a stronger negative effect on measuring topical diversity than impure topics. Also, LDA +DR outperforms LDA+TR. So, removing impurity from $P(t|d)$ distributions is the most effective approach in the topical diversity task, and removing impurity from $P(w|d)$ distributions is more effective than removing impurity from $P(w|t)$ distributions. Table 1 illustrates the difference between LDA and HiTR with the topics assigned by the two methods for a non-diverse document that is combined from two documents from the same journal, entitled "Molecular Neuroscience: Challenges Ahead" and "Reward Networks in the Brain as Captured by Connectivity Measures", using the procedure described in Section 5.2. As only a very basic stopword list was applied, words like *also* and *one* still appear. We expect to have a low diversity value for

the combined document. However, using Rao's diversity measure, the topical diversity of this document based on the LDA topics is 0.97. This is due to the fact that there are three document-specific topics—topics 1, 2 and 4—and four general topics. Topics 1 and 2 are very similar and the δ between them is 0.13. The δ between, the other, more general topics is high; the average δ value between pairs of topics is as high as 0.38. For the same document, HiTR only assigns three document-specific topics and they are more pure and coherent. The average δ value between pairs of topics assigned by HiTR is 0.19. The diversity value of this document using HiTR is 0.16, which indicates that this document is non-diverse.

Next, Table 2 shows the sparsity of $P(t|d)$ using different topic models. All topic models that have TAR level of re-estimation achieve very sparse topic models. Thus, TAR contributes more to the sparsity achieved by HiTR. TAR increases the sparsity of LDA by more than 80 percent. This sparsity leads to improvements over the performance of LDA on the topical diversity task, which indicates that TAR is able to remove general topics from documents. Topic models achieved by PTM are slightly more sparse than those achieved by HiTR. However, the difference is not

TABLE 1
Topic Assignments for a Non-Diverse Document Using LDA and HiTR

LDA		
Topic	$P(t\|d)$	Top 5 words
1	0.21	brain, anterior, neurons, cortex, neuronal
2	0.14	channel, neuron, membrane, receptor, current
3	0.10	use, information, also, new, one
4	0.08	network, nodes, cluster, functional, node
5	0.08	using, method, used, image, algorithm
6	0.08	time, study, days, period, baseline
7	0.07	data, values, number, average, used
HiTR		
Topic	$P(t\|d)$	Top 5 words
1	0.68	brain, neuronal, neurons, neurological, nerve
2	0.23	channel, synaptic, neuron, receptor, membrane
3	0.09	network, nodes, cluster, community, interaction

Only topics with $P(t|d) > 0.05$ are shown.

TABLE 2
Sparsity of Topic Models Trained on PubMed for the Topical Diversity Task

Method	Sparsity
LDA	13.77
PTM	**1.78**
LDA+DR	13.17▼
LDA+TR	12.35▼
LDA+TAR	2.12
LDA+DR+TR	11.46▼
LDA+DR+TAR	2.01
LDA+TR+TAR	1.92
HiTR	1.80

For significance tests, we consider p-value $< 0.05/7$.

statistically significant. The fact that HiTR outperforms PTM indicates that PTM extremely parsimonizes documents and throws away essential information from documents while HiTR removes mostly non-essential information from documents.

6.2 HiTR Results

In this section we analyze different levels of re-estimation to get insights on how different levels on re-estimation work individually and how much they are successful in removing non-necessary information from documents, topics, and topic-assignments.

6.2.1 Document Re-Estimation Results

In this section we focus on answering our second research question: What is the effect of DR on the quality of topic models? Can DR replace manual pre-processings?

DR outperforms LDA by 7 percent in measuring documents' topical diversity in terms of AUC. It also outperforms TR in this task but the difference is not significant. In fact, DR and TR are addressing the same problem with topic models. Both are successful in addressing *impure topics*. However they are not successful in addressing the *general topics* problem, since they have high value of sparsity.

To analyze the effectiveness of DR in re-estimating documents and addressing the problems with topic models, we design an experiment in which no manual pre-processing is done and topic models are trained on these not-pre-processed documents. Our expectation is that even without doing any pre-processing a method that addresses the generality problems with topic models should still be able to achieve a good performance and do the pre-processing implicitly and automatically. Since DR tries to pre-process documents automatically, it should achieve a high quality topic model on these datasets. Table 3 shows the performance of LDA, DR, and LDA+DR+TR in terms of their coherence. As expected, the coherence of LDA decreases by more than 23 percent when no pre-processing is done on documents. More interestingly, adding DR scores better, both in terms of coherence and AUC, than manual pre-processing.

Next, we analyze the effect of the amount of document re-estimation on the quality of topic models. We control the amount of re-estimation by the values of the parameters of DR: λ and $threshold$. Fig. 4 shows the effect of different values of the parameters on documents and its impact on the quality of trained topic models. Two conclusions can be drawn. First, λ does not have a great impact on the documents' statistics as even with very different values of λ documents have similar statistics. The threshold has a bigger impact on the documents. Second, although the statistics of documents are similar for different values of λ, the thresholds for which the best coherence is achieved for them, are very different. For $\lambda = 0.5$ the best coherence is achieved for $threshold = 0.01$, while for $\lambda = 0.8$ the best coherence is achieved for $threshold = 0.05$. This indicates that there is a correlation between these parameters. As expected, when λ is high, which corresponds to less re-estimation, the threshold should be high to remove unnecessary words from documents.

TABLE 3
The Effect of Document Pre-Processing on the Quality of Topic Models Measured in Terms of Coherence and AUC Achieved in the Topical Diversity Task

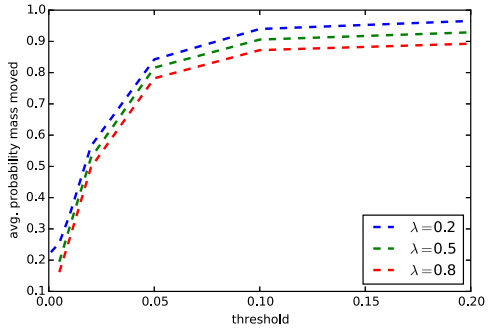
Method	*Coherence*	*AUC*
LDA (without pre-processing)	6.23	0.54
LDA+pre-processing	8.45	0.73
LDA+DR	8.95	0.75
LDA+DR+TR	10.29	0.79

6.2.2 Topic Re-Estimation Results

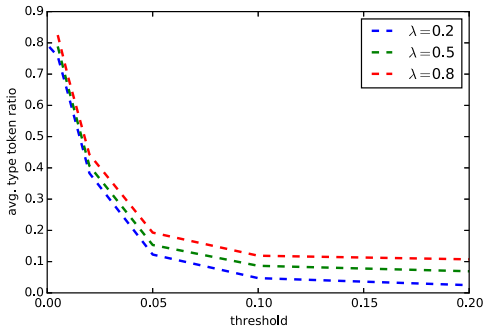
To answer our third research question, we now focus on the TR level of HiTR. Since TR tries to remove the impurity from topics, we expect TR to increase the coherence of the topics by removing unnecessary words from topics. Table 4 shows the top five words for some example topics calculated from the PubMed dataset, before and after applying TR. These examples indicate that TR can successfully remove general words from topics.

We measure the purity of topics based on the coherence of words within $P(w|t)$ distributions. Table 5 shows the coherence of topics according to different topic modeling approaches, in terms of average mutual information. More coherent topics are beneficial, because they are an indicator of more pure topics, which are essential to achieving a good performance in topical diversity task. TR increases the coherence of topics by removing the impure parts from topics. The coherence of PTM is higher than the coherence of TR. However, when we first apply DR, train LDA, and finally apply TR, the coherence of the extracted topics is significantly higher than the coherence of topics extracted by PTM. From these findings we conclude that TR is effective in removing impurity from topics. Moreover, DR also contributes in making topics more pure.

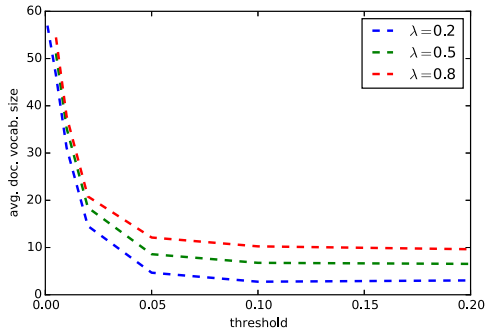
To see how much impurity is being removed from topics by using TR, we investigate the effect of TR on the distribution of words within topics and we measure the number of words and the re-allocated probability mass within topics before and after TR. Fig. 5 shows the probability mass of the words left after TP is applied to the topics of the original LDA model. The average number of words within extracted topics from the PubMed dataset is about 337 without TR,



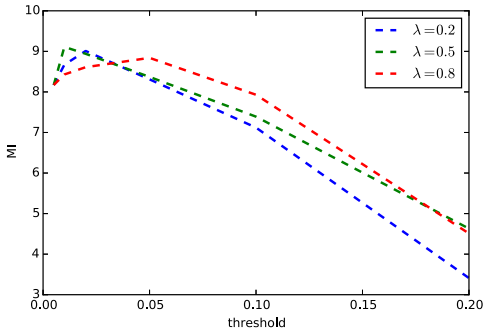
(a) Probability mass moved from removed words to the remaining words



(b) Average type-token ratio of documents



(c) Average document vocabulary size



(d) Coherence of topic models estimated using Equation 5

Fig. 4. The effect of different values of the parameters of DR on the documents in terms of their probability mass moved, type-token ratio, and vocabulary size and its effect on the quality of trained topic models in terms of their coherence.

and about 181 after performing TR. On average, the words that are not removed by TR take 41 percent of the probability mass in the LDA topic models (the dotted red line in Fig. 5). In the re-estimated topic model, they occupy the full 100 percent of the probability mass. Thus, after applying TR, the topic models become more sparse, and the remaining topic-specific words receive higher probabilities. As shown in the figure, over all topics, after applying TR, the probability mass is re-allocated and some words are removed.

TABLE 4
Examples of Topics before and after Applying Topic Re-Estimation on the PubMed Dataset

	Before TR		After TR	
Topic t	w	$p(w\|t)$	w	$p(w\|t)$
1	women	0.07	women	0.06
	men	0.02	men	0.05
	costs	0.02	health	0.05
	per	0.02	costs	0.03
	total	0.02	economic	0.02
2	using	0.01	algorithm	0.04
	method	0.01	method	0.03
	used	0.01	data	0.03
	algorithm	0.01	performance	0.02
	data	0.01	system	0.01
3	sequences	0.02	genome	0.05
	genome	0.02	sequences	0.04
	genes	0.02	genes	0.03
	using	0.01	genomic	0.03
	two	0.01	gene	0.02

TABLE 5
The Coherence of Different Topic Models in Terms of Average Mutual Information between Top 10 Words in the Topics Calculated Using Equation (5) on the PubMed Dataset

Method	Coherence
LDA	8.17
PTM	9.89
LDA+TR	9.46
LDA+DR+TR	10.29▲

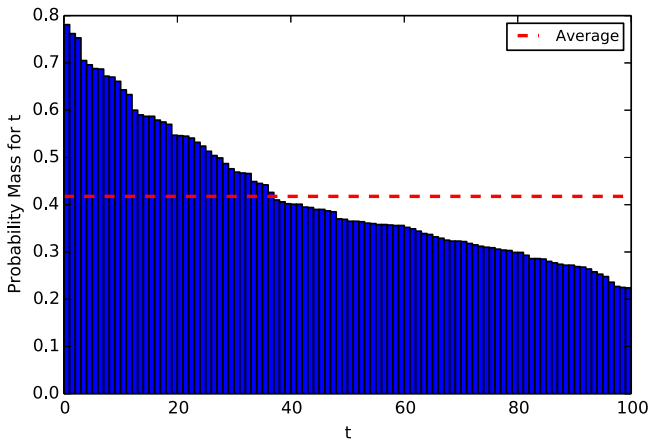


Fig. 5. Probability mass of the words left after TR in the topics of the original LDA model. The y-axis shows $\sum_{\{w \mid P_{LDA+TR}(w|t)>0\}} P_{LDA}(w|t)$ for a topic t.

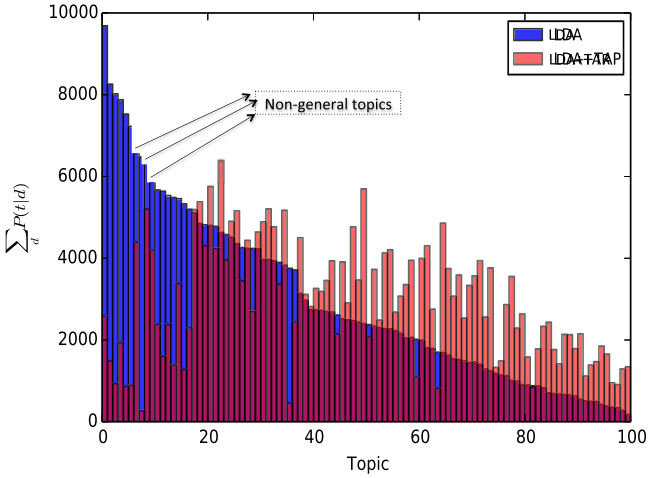


Fig. 6. The total probability of assigning topics to the documents in the PubMed dataset estimated using LDA and LDA+TAR. (The two areas are both equal to the number of documents ($N \approx 300K$)).

6.2.3 Topic Assignment Re-Estimation Results

To answer our fourth research question, we now turn to the TAR level of HiTR. We are interested in seeing how HiTR deals with the issue of general topics. General topics are topics that, for many documents, have a high probability of being assigned. To gain insight in how LDA and HiTR perform in this respect, we sum the probability of assigning a topic to a document, over all documents: for each topic t, we calculate $\sum_{d \in C} P(t|d)$, where C is the collection of all documents. Fig. 6 shows the distribution of probability mass before and after applying TAR. General topics naturally have high values as they are assigned to most of the documents with high probability. In Fig. 6 the topics are sorted based on the topic assignment probability of LDA. As we can see from Fig. 6, LDA assigns a vast portion of the probability mass to a relatively small number of topics. These topics are mostly general topics that are assigned to most of documents. We expect, however, that many topics are represented in some documents, while relatively few topics will be relevant to all documents. When TAR is applied, the distribution is less skewed and the probability mass is more evenly distributed. There are some topics that have high $\sum_d P(t|d)$ value in LDA's topic assignments and high $\sum_d P(t|d)$ value after applying TAR as well (they are marked as "non-general topics" in Fig. 6). Table 6 shows the top five words for these topics. Although these topics contain some general words such as "used", they are not general topics. TAR is able to find these three non-general topics and their assignment probabilities to documents in the $P(t|d)$ distributions is not changed as much as the actual general topics.

TABLE 6
Top Five Words for the Topics Marked As "Non-General Topics" in Fig. 6

Topic	Top 5 words
1	health, services, public, countries, data
2	surgery, surgical, postoperative, patient, performed
3	cells, cell, treatment, experiments, used

TABLE 7
Top Five Words for the Topics Detected by TAR As General Topics

Topic	Top 5 words
1	use, information, also, new, one
2	ci, study, analysis, data, variables
3	time, study, days, period, baseline
4	group, control, significantly, compared, groups
5	study, group, subject, groups, significant
6	may, also, effects, however, would
7	data, values, number, average, used

To further investigate whether TAR really removes general topics, in Table 7 we show the top five words for the first 10 topics in Fig. 6, excluding the topics marked as "non-general topics" in the figure. These seven topics have the highest decrease in $\sum_d P(t|d)$ values when we apply TAR. As can be seen from Table 7, the topics contain general words and are not informative. In the figure, we can see that after applying TAR, the $\sum_d P(t|d)$ values are decreased dramatically for these topics and that the mass is re-distributed across other topics, without creating new general topics that apply to nearly all documents. We can conclude that TAR can correctly distinguish general from specific topics and re-assign probability mass accordingly.

6.3 Parameter Analysis

In this section we analyze the effect of the λ parameter on the performance of DR, TR, and TAR in the topical diversity task. Fig. 7 displays the performance at different levels of re-estimation based on a range of values for λ. Recall that with $\lambda = 1$, no re-estimation takes place, and all methods equal LDA. The following interesting observations can be made from this figure.

First, DR reaches its best performance with moderate values of λ ($0.4 \leq \lambda \leq 0.45$). This reflects that documents contain a moderate amount of general information and that DR is able to successfully deal with it. For $\lambda \geq 0.8$ the performance of DR and LDA is the same and for these values of λ DR does not increase the quality of LDA.

Second, the best performance of TR is achieved with high values of λ ($0.65 \leq \lambda \leq 0.75$). This indicates that topics usually only need a small amount of re-estimation. With this slight re-estimation, TR is able to improve the quality of LDA. However, for the values of $\lambda \geq 0.75$ the accuracy of TR degrades.

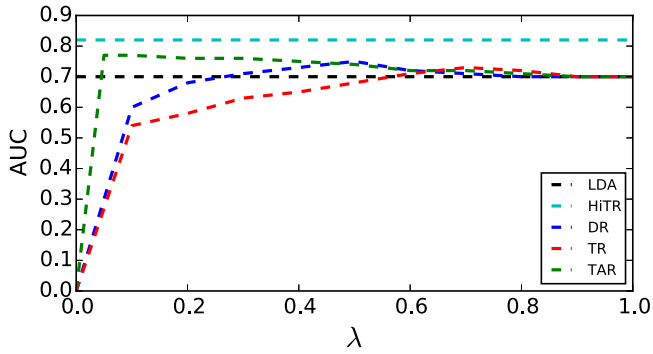


Fig. 7. The effect of the λ parameter on the performance of topics models in topical diversity task on PubMed dataset.

TABLE 8
The Performance and Sparsity of HiTR Using PTM As the Underlying Topic Model in the Topical Diversity Task

Method	AUC	Sparsity
PTM	0.78	1.78
PTM+DR	0.79	1.73
PTM+TR	0.77	1.71
PTM+TAR	0.78	1.65
PTM+HiTR	0.79	1.63

Third, TAR achieves its best performance with very low values of λ ($0.02 \leq \lambda \leq 0.05$). These low values of λ correspond to more re-estimation. From this result, we conclude that most of the noise is in the $P(t|d)$ distributions, and that aggressive re-estimation allows TAR to remove most of this noise. The best values of λ optimized for HiTR using the development set are close to the best values of λ according to Fig. 7.

6.4 Impact of Underlying Topic Model on the Performance of HiTR

In this section, we analyze the effect of using PTM as the underlying topic model for HiTR on the performance of HiTR. We apply HiTR on top of PTM and compare the results with the results of applying HiTR on top of LDA. Table 8 shows the results of this experiment. The results show that:

1) Applying HiTR on top of PTM does not improve PTM's performance significantly. We believe, the reason is that PTM already removes a lot of general information from topics/documents, but in some cases it also removes non-general information. LDA is in the other side of the spectrum, it keeps all information (general and non-general), and HiTR removes general information and keeps only the non-general information which leads to a higher performance.
2) PTM benefits the most from the DR step. It shows that PTM is already effective in removing generality/impurity from topic-word and document-topic distributions, however it does not have a mechanism to remove generality/impurity from document-word distributions.
3) The performance of HiTR with LDA is significantly better than the performance of PTM and PTM with HiTR. As we mentioned, this shows that HiTR is more effective when the underlying topic model contains all information (general and non-general) and it can remove the non-general part.
4) In terms of sparsity, HiTR makes PTM more sparse, however the difference is not significant. Thus, applying HiTR on an already sparse topic model does not have a big influence on its sparsity.

7 Analysis

In this section, we want to gain additional insights into HiTR and its effects on topic estimation. Purity of topic assignments to documents based on $P(t|d)$ distributions has the highest effect on the quality of estimated diversity scores for documents. Therefore, it is important to measure how pure the estimated topic assignments are using HiTR. In this section, we measure how much impurity is removed by HiTR from topic distributions. Then, we analyze the efficiency of HiTR.

Based on the topics assigned by HiTR, LDA and PTM, we perform document clustering and document classification. For clustering, following [16], we consider each topic as a cluster. Each document d is assigned to the topic that has the highest probability value in $P(t|d)$. For classification, we use all topics assigned to the document and consider them as features for a supervised classification algorithm. As the classification algorithm we use SVM. High accuracy achieved in document classification is then an indicator of high purity of topic distributions.

We note that our focus in this section is not on achieving a top performance in document clustering and classification tasks: we only consider these tasks as a means to assess the purity of topic distributions using different topic models.

7.1 Datasets

We use three datasets: 20-NewsGroups,[3] Reuters [32] and Ohsumed.[4] The Reuters dataset contains 806,791 documents with category labels for 126 categories. For clustering and classification of documents, we use the 55 categories in the second level of the category hierarchy. 20-NewsGroups contains 20 categories and around 1,000 documents in each category, so in total there are about 20,000 documents. The Ohsumed dataset contains 50,216 documents grouped into 23 categories.

7.2 Purity Metrics

For measuring the purity of clusters, two standard evaluation metrics are used: *purity* and *normalized mutual information* (NMI) [33].

7.3 Settings

We evaluate document clustering and classification using 10-fold cross validation and perform the same document pre-processing as described in Section 5.5.

7.4 Purity Results

Table 9 shows the purity of HiTR in the document clustering task. For all 3 datasets, on both measures, the purity of topics created by HiTR is significantly higher than with PTM. As expected, TAR is mostly responsible for the purity of $P(t|d)$: all runs which include TAR either improve or do not differ significantly from PTM. The different combinations show that also DR and TR yield additional purity, indicating that each of the three address different issues and contribute in a different way.

Table 10 shows the performance of different topic models on the document classification task. Again HiTR significantly outperforms PTM on all three datasets. We see the same trend as with clustering, but amplified: here all runs

3. Available at http://www.ai.mit.edu/people/~jrennie/20Newsgroups/
4. Available at http://disi.unitn.it/moschitti/corpora.htm

TABLE 9
Purity of Topic Models Estimated in Terms of Purity Achieved in Document Clustering

	Reuters (N = 806, 791, C = 55)		20-Newsgroups (18, 846, C = 20)		Ohsumed (N = 50, 216, C = 23)	
Method	Purity	NMI	Purity	NMI	Purity	NMI
LDA	0.55	0.40	0.52	0.36	0.50	0.30
PTM	0.61	0.43	0.57	0.38	0.55	0.33
LDA+DR	0.57▼	0.41▼	0.56	0.39	0.53▼	0.32▼
LDA+TR	0.57▼	0.42▼	0.56	0.38	0.53▼	0.31▼
LDA+TAR	0.60	0.43	0.57	0.39	0.54	0.33
LDA+DR+TR	0.58	0.42▼	0.57	0.38	0.54	0.32
LDA+DR+TAR	0.60	0.43	0.58	0.40	0.55	0.35▲
LDA+TR+TAR	0.61	0.43	0.58	0.40▲	0.56▲	0.34▲
HiTR	**0.64**▲	**0.45**▲	**0.60**▲	**0.42**▲	**0.57**▲	**0.35**

For significance tests, we consider p-value $< 0.05/7$.

without TAR perform significantly worse than PTM. Note that on the smallest dataset, LDA and PTM performs already well, and so are harder to improve. Where in document clustering only the topics with the highest probability are considered, in document classification the classifiers use the entire $P(t|d)$ distributions to classify documents. Performance of all methods in document classification is more closer to the perfect classifier than their performance in document clustering, as the maximum value of both accuracy and purity is 1. This indicates that the most probable topic does not necessarily contain all information about the content of a document. In the cases that a document is about more than one topic, the classifier utilizes all $P(t|d)$ information and performs better. Therefore, the higher accuracy of HiTR in this task is an indicator of its ability to assigning document-specific topics to documents.

7.5 HiTR's Efficiency

Table 11 shows the execution times of HiTR, LDA, and PTM. The reported execution time for HiTR is the time took to run HiTR once, given the corpus as input and topic assignments to documents as output. All models were run on machines with 6-core 3.0 GHz processors. The results show that, even on large datasets, HiTR does not add much complexity to LDA and the difference between the execution times of LDA and HiTR are reasonable. The execution times of PTM grow much faster than those of LDA and HiTR when the number of documents increase.

TABLE 11
The Execution Time of HiTR, LDA, and PTM in Hours

Dataset	Method	Hours
Reuters	LDA	6.18
N = 807 K	PTM	26.00
#w =1,5 M	HiTR	9.17
20-NewsGroups	LDA	1.13
N = 19 K	PTM	0.93
#w = 5,2 M	HiTR	1.45
Ohsumed	LDA	1.42
N = 50 K	PTM	3.88
#w = 10 M	HiTR	2.45

N and #w are the number of documents and tokens in the corpus, respectively

8 Conclusion

We have proposed Hierarchical Topic model Re-estimation, an approach for re-estimating topic models and applied them to measure topical diversity of text documents.

We have shown by experimental means that our approaches are able to remove general topics from topic models and increase the purity of topics. The results show that the estimated diversity scores for documents using HiTR are more accurate than those extracted using topic models created by LDA and PTM. Our three main findings are as follows. First, general topics have the largest negative impact on the quality of topic models when they are used for measuring topical diversity. This indicates that purity of topic assignments is more important than purity of the distribution of words in topics and the distribution of words in documents in topical diversity task. The topic assignment re-estimation that is designed to address this problem successfully detects general topics and removes them from documents. Second, re-estimation at each level helps to improve the quality of estimated diversity scores. We have shown that these "cleaned document topic models" yield better results when applied to measure topical diversity of documents. However, to achieve a highly accurate diversity scores, re-estimation at all three levels is needed to improve

TABLE 10
Purity of Topic Models Estimated in Terms of Accuracy Achieved in Document Classification

	Reuters (N = 806,791, C = 55)		20-Newsgroups (N = 18,846, C = 20)		Ohsumed (N = 50,216, C = 23)	
Method	Accuracy	Imp. over LDA	Accuracy	Imp. over LDA	Accuracy	Imp. over LDA
LDA	0.76	–	0.81	–	0.50	–
PTM	0.82	8%	0.87	7%	0.56	12%
LDA+DR	0.79▼	4%	0.83▼	2%	0.52▼	4%
LDA+TR	0.78▼	3%	0.83▼	2%	0.53▼	1%
LDA+TAR	0.82	8%	0.85▼	5%	0.54	8%
LDA+DR+TR	0.80▼	5%	0.84▼	4%	0.53▼	6%
LDA+DR+TAR	0.83	9%	0.86	6%	0.56	12%
LDA+TR+TAR	0.82▲	8%	0.87	7%	0.58▲	16%
HiTR	**0.85**▲	12%	**0.89**▲	10%	**0.60**▲	20%

For significance tests, we consider p-value $< 0.05/7$.

on the state-of-the-art PTM approach. Third, we analyzed the effectiveness of HiTR in two other tasks: document clustering and document classification. We found that HiTR can achieve higher performances in these tasks compared to LDA and PTM. This finding suggest that although HiTR is originally designed for better estimation of topical diversity, it can be applied in a wider variety of tasks.

Our proposed approach has some limitations. First, HiTR is most effective at removing general information from the probability distributions mentioned. However, to train a more accurate topic model which has a good performance in topical diversity task it is also important to remove very specific words from documents. Current approaches, including HiTR, are not able to address this problem adequately. Second, the experiments on the topical diversity task are conducted in an artificially created dataset. More robust datasets are needed for evaluating HiTR in this task.

There are several future directions. In principle, HiTR is a re-estimation method that can be applied to any topic model to enhance its quality. In this paper, we have applied HiTR to LDA and PTM. In our future work, we plan to examine the effect of HiTR on a wide range of topic models besides LDA and PTM such as PLSA. In this research we adapted and used Rao's diversity measure for estimating diversity of documents. There are several other diversity measures proposed in biology such as Functional Divergence and Functional Attribute Diversity.

Acknowledgments

This research was supported by the Netherlands Organization for Scientific Research (ExPoSe project, NWO CI # 314.99.108; DiLiPaD project, NWO Digging into Data # 600.006.014), Nederlab (340-6148-t1-6), and by the European Community's Seventh Framework Program (FP7/2007-2013) under grant agreement ENVRI, number 283465.

References

[1] K. Bache, D. Newman, and P. Smyth, "Text-based measures of document diversity," in *Proc. 19th ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2013, pp. 23–31.

[2] H. Azarbonyad, F. Saan, M. Dehghani, M. Marx, and J. Kamps, "Are topically diverse documents also interesting?" in *Proc. 6th Int. Conf. Exp. IR Meets Multilinguality Multimodality Interaction*, 2015, pp. 215–221.

[3] H. Azarbonyad, M. Dehghani, T. Kenter, M. Marx, J. Kamps, and M. de Rijke, "Hierarchical re-estimation of topic models for measuring topical diversity," in *Proc. 39th Eur. Conf. IR Res.*, 2017, pp. 68–81.

[4] M. Derzinski and K. Rohanimanesh, "An information theoretic approach to quantifying text interestingness," in *Proc. Neural Inf. Process. Syst. MLNLP Workshop*, 2014, pp. 1–6.

[5] C. Rao, "Diversity and dissimilarity coefficients: A unified approach," *Theoretical Population Biol.*, vol. 21, no. 1, pp. 24–43, 1982.

[6] A. Solow, S. Polasky, and J. Broadus, "On the measurement of biological diversity," *J. Environ. Econ. Manage.*, vol. 24, no. 1, pp. 60–68, 1993.

[7] H. Soleimani and D. Miller, "Parsimonious topic models with salient word discovery," *IEEE Trans. Knowl. Data Eng.*, vol. 27, no. 3, pp. 824–837, Mar. 2015.

[8] H. M. Wallach, D. M. Mimno, and A. McCallum, "Rethinking lda: Why priors matter," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 1973–1981.

[9] T. Lin, W. Tian, Q. Mei, and H. Cheng, "The dual-sparse topic model: Mining focused topics and focused terms in short text," in *Proc. 23rd Int. Conf. World Wide Web*, 2014, pp. 539–550.

[10] D. Hiemstra, S. Robertson, and H. Zaragoza, "Parsimonious language models for information retrieval," in *Proc. 27th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2004, pp. 178–185.

[11] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent dirichlet allocation," *J. Mach. Learn. Res.*, vol. 3, pp. 993–1022, 2003.

[12] J. Boyd-Gaber, D. Mimno, and D. Newman, "Care and feeding of topic models: Problems, diagnostics, and improvements," in *Handbook of Mixed Membership Models and Their Applications*. Boca Raton, FL, USA: CRC Press, 2014.

[13] C. Wang and D. M. Blei, "Decoupling sparsity and smoothness in the discrete hierarchical dirichlet process," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 1982–1989.

[14] S. Williamson, C. Wang, K. A. Heller, and D. M. Blei, "The IBP compound Dirichlet process and its application to focused topic modeling," in *Proc. 27th Int. Conf. Mach. Learn.*, 2010, pp. 1151–1158.

[15] H. M. Wallach, I. Murray, R. Salakhutdinov, and D. Mimno, "Evaluation methods for topic models," in *Proc. 26th Annu. Int. Conf. Mach. Learn.*, 2009, pp. 1105–1112.

[16] D. Q. Nguyen, R. Billingsley, L. Du, and M. Johnson, "Improving topic models with latent feature word representations," *Trans. Assoc. Comput. Linguistics*, vol. 3, pp. 299–313, 2015.

[17] S. Lacoste-Julien, F. Sha, and M. I. Jordan, "DiscLDA: Discriminative learning for dimensionality reduction and classification," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 897–904.

[18] P. Xie and E. P. Xing, "Integrating document clustering and topic modeling," in *Proc. 29th Conf. Uncertainty Artif. Intell.*, 2013, pp. 694–703.

[19] R. Mehrotra, S. Sanner, W. Buntine, and L. Xie, "Improving LDA topic models for microblogs via tweet pooling and automatic labeling," in *Proc. 36th Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2013, pp. 889–892.

[20] J. H. Lau, D. Newman, and T. Baldwin, "Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality," in *Proc. 14th Conf. Eur. Chapter Assoc. Comput. Linguistics*, 2014, pp. 530–539.

[21] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-graber, and D. M. Blei, "Reading tea leaves: How humans interpret topic models," in *Proc. Neural Inf. Process. Syst.*, 2009, pp. 288–296.

[22] D. Newman, E. V. Bonilla, and W. Buntine, "Improving topic coherence with regularized topic models," in *Proc. Neural Inf. Process. Syst.*, 2011, pp. 496–504.

[23] M. Law, M. Figueiredo, and A. Jain, "Simultaneous feature selection and clustering using mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 26, no. 9, pp. 1154–1166, Sep. 2004.

[24] C. Constantinopoulos, M. Titsias, and A. Likas, "Bayesian feature and model selection for gaussian mixture models," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 28, no. 6, pp. 1013–1018, Jun. 2006.

[25] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx, "On horizontal and vertical separation in hierarchical text classification," in *Proc. ACM Int. Conf. Theory Inf. Retrieval*, 2016, pp. 185–194.

[26] C. Zhai and J. Lafferty, "Model-based feedback in the language modeling approach to information retrieval," in *Proc. 10th Int. Conf. Inf. Knowl. Manage.*, 2001, pp. 403–410.

[27] Y. Zhang, J. Callan, and T. Minka, "Novelty and redundancy detection in adaptive filtering," in *Proc. 25th Annu. Int. ACM SIGIR Conf. Res. Develop. Inf. Retrieval*, 2002, pp. 81–88.

[28] M. Dehghani, H. Azarbonyad, J. Kamps, and M. Marx, "Two-way parsimonious classification models for evolving hierarchies," in *Proc. 7th Int. Conf. CLEF Assoc.*, 2016, pp. 69–82.

[29] D. M. Blei, "Probabilistic topic models," *Commun. ACM*, vol. 55, no. 4, pp. 77–84, 2012.

[30] "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/

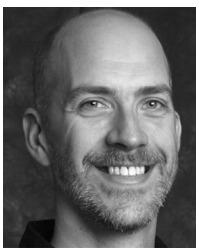
[31] M. Röder, A. Both, and A. Hinneburg, "Exploring the space of topic coherence measures," in *Proc. 8th ACM Int. Conf. Web Search Data Mining*, 2015, pp. 399–408.

[32] D. D. Lewis, Y. Yang, T. G. Rose, and F. Li, "Rcv1: A new benchmark collection for text categorization research," *J. Mach. Learn. Res.*, vol. 5, no. Apr, pp. 361–397, 2004.

[33] C. D. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*. Cambridge, United Kingdom: Cambridge Univ. Press, 2008.

Hosein Azarbonyad received the MSc degree in information technology from the University of Tehran. He is working toward the PhD degree at the University of Amsterdam. His research interests include information retrieval, machine learning, and natural language processing. He is a member of the Information and Language Processing Systems (ILPS) group. He has served as an organizing committee member of WSDM 2017 and of ICTIR 2017. He has received the best paper award at ICTIR 2016 and the best poster award at ECIR 2015.

Mostafa Dehghani working toward the PhD degree at the University of Amsterdam. His research interests include intersection of machine learning, in particular, deep learning, and information retrieval. He has published several papers in conferences like SIGIR, CIKM, ICTIR, ECIR, and CHIIR. He has got the best poster award at ECIR2015, best doctoral consortium award at SIGIR 2016, and the best paper award at ICTIR2016.



Tom Kenter received the PhD degree from the Information and Language Processing Systems group, University of Amsterdam, supervised by Maarten de Rijke. He is currently doing research with Google United Kingdom, on the topic of text-to-speech and natural language understanding. He received two internships with Google Research in Mountain View, was part of the organizing committees of ECIR 2014 and BNAIC 2016 and is an editorial board member of *Information Processing & Management* (Elsevier). His research interests include natural language understanding, machine reading, and text-to-speech. He has published at ACL, CIKM, SIGIR, and AAAI.

Maarten Marx received the master's degree in political science and the PhD degree in mathematical logic, both from the University of Amsterdam, in 1990 and 1995, respectively. His current research interest is integration of large amounts of semi-structured, text-centric, and data. He co-authored three books and more than 75 scientific articles. Since 2002, his main research topic is XML, in particular XPath dialects. In 2004, he received the ACM Principles of Database Systems best paper award for his Codd-completeness result for "Conditional XPath". His work on the parliamentary proceedings was recognized with the XML Holland Award 2008 and the Dutch Data prize (awarded by DANS-KNAW) 2012.

Jaap Kamps is an associate professor of information retrieval with archives and information studies, Faculty of Humanities, University of Amsterdam. His research interests include information storage and retrieval, big data, linked data, structure and semantic annotation, digital humanities, e-humanities, digital heritage, evaluation and user studies, interactive search, task based search, exploratory search, and sense making. He is PI of a range of externally funded research projects on the search and exploration of domain specific collections from libraries, archives, and museums. He is an active organizer in DL and IR conferences and workshops, in particular focusing on richly annotated corpora (e.g., INEX, CLEF, TREC, and ESAIR).

Maarten de Rijke is an university professor of artificial intelligence and information retrieval. His research concerns technology to connect people to information: search engines, recommender systems, and conversational agents. He is the editor-in-chief of the *ACM Transactions on Information Systems* and co-editor-in-chief of the *Foundations and Trends in Information Retrieval*. He is a member of the Royal Netherlands Academy of Arts and Science and director of the Innovation Center for Artificial Intelligence (ICAI).