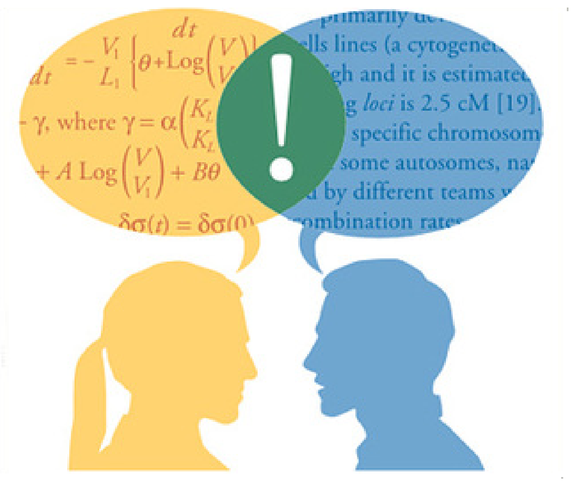
Hierarchical Re-estimation of Topic Models for Measuring Topical Diversity

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

Motivation

Quantitative notions of *topical diversity* in text documents are useful in several contexts, e.g., to assess the *interdisciplinarity* of a research proposal or to determine the *interestingness* of a document.



Diversity of a population:

- Diversity is decomposed in terms of *elements* that belong to *categories* within a *population*.
- Diversity is defined as the expected distance between two randomly selected elements of the population:

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

where p_i and p_j are the proportions of categories i and j in the population and $\delta(i,j)$ is the distance between i and j.

Measuring topical diversity:

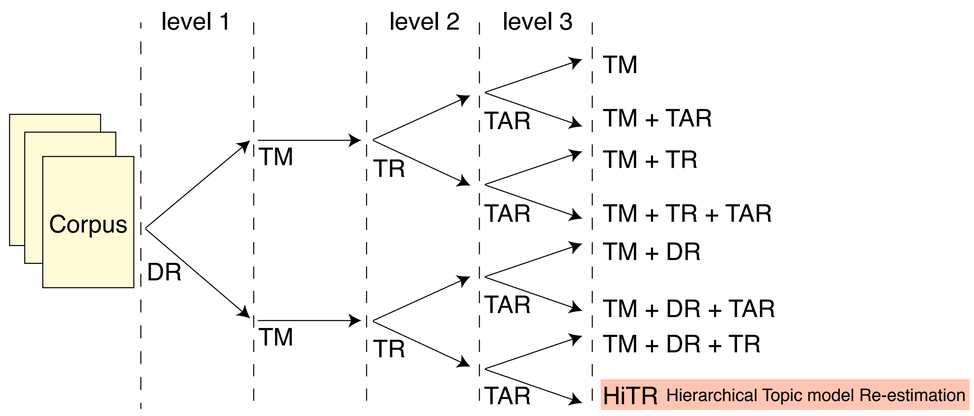
- Using *topic models* like LDA to assign words to topics and topics to documents.
- Modeling elements based on the probability of a word w given a document d, $P(w|d)$, categories based on the probability of w given a topic t, $P(w|t)$, and populations based on the probability of t given d, $P(t|d)$.
- Topic models play a central role in this approach.

Challenges:

- Generality: General topics mostly contain general words and are typically assigned to most documents in a corpus.
- Impurity: Impure topics contain words that are not related to the topic.
- Generality and impurity of topics both result in low quality of $P(t|d)$ distributions.
- The quality of diversity scores are highly dependent to the quality of $P(t|d)$ distributions.

Hierarchical Topic Model Re-estimation (HiTR)

HiTR re-estimates elements ($P(w|d)$ distributions), categories ($P(w|t)$ distributions), and populations ($P(t|d)$ distributions) to address generality and impurity issues.



- TM is a topic modeling approach like, e.g. LDA.
- Document re-estimation (DR) re-estimates the language model per document $P(w|d)$.
- Topic re-estimation (TR) addresses impurity problem by re-estimating the language model per topic $P(w|t)$.
- Topic assignment re-estimation (TAR) addresses generality problem by re-estimating the distribution over topics per document $P(t|d)$.

Probability distribution re-estimation (parsimonization):

Assumption: The language model of a document/topic is a mixture of its *specific* language model and a *general* language model:

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

Goal: To estimate the specific language model ($P(w|\tilde{\theta}_d)$) for each document/topic.

E-step:

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

M-step:

$$P(w|\tilde{\theta}_d) = \frac{e_w}{\Sigma_{w'} e_{w'}}$$

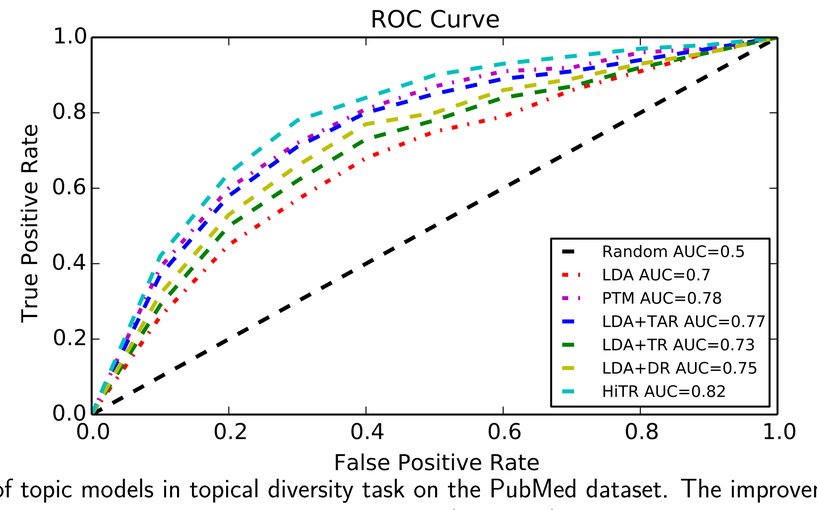
Experiments

Experimental setup

- Datasets
  - Topic models trained on 300K articles published between 2012 and 2015 in PubMed.
  - Evaluation is done on 500 diverse and 500 non-diverse documents created using articles from the PubMed dataset.

Experiment 1: Performance of HiTR

- Research Question 1: How effective is HiTR in measuring topical diversity of documents?



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

- Outcome 1: HiTR benefits from the three re-estimation approaches it encapsulates by successfully improving the quality of estimated diversity scores.

Experiment 2: Effectiveness of TR

- Research Question 2: Does TR increase the purity of topics? If so, how does using the more pure topics influence the performance in topical diversity task?

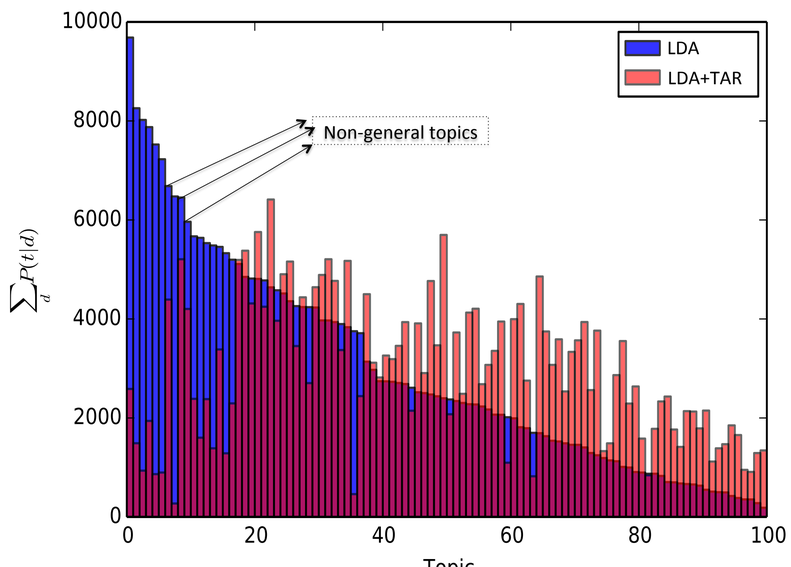
LDA	PTM	LDA+TR	LDA+DR+TR
8.17	9.89	9.46	10.29▲

Topic model coherence in terms of average normalized mutual information between top 10 words in the topics on the PubMed dataset.

- Outcome 2:TR is effective in removing impurity from topics. DR also contributes in making topics more pure.

Experiment 3: Effectiveness of TAR

- Research Question 3: How does TAR affect the sparsity of document-topic assignments?



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

- Outcome 3:TAR removes general topics from documents and increases the probability of document-specific topics for each document.

Conclusions

- General purpose topic models might fail in estimating topical diversity of documents due to the generality and impurity issues.
- HiTR is able to address impurity and generality issues with topic models.
- Re-estimation at each level helps to achieve high quality diversity scores.
- HiTR utilizes all re-estimation approaches and outperforms state-of-the-art in topical diversity task.

ExPoSe:Exploratory Political Search, Funded by NWO (project# 314-99-108)
Netherlands Organisation for Scientific Research
University of Amsterdam
39th European Conference on Information Retrieval
h.azarbonyad@uva.nl