Construction and quality evaluation for heterogeneous hierarchical topic modeling

Belyy Anton Vladimirovich
ITMO University

Advisor: Filchenkov A. A., PhD, ITMO University
Consultant: Vorontsov K. V., PhD, CC RAS

2018

Problem statement

Goal

Develop an aggregation method for large collection of documents from heterogeneous sources.

Tasks

1. Create a hierarchical topic model of popular scientific texts, collected from **heterogeneous** sources.
2. Propose an automated way to evaluate quality of topical hierarchy of a given model.
3. Implement an exploratory search engine to demonstrate the proposed algorithm.

Method

Additive regularization of topic models (ARTM).

Datasets

- PostNauka: 2976 documents, 43196 words, 1799 tags
- Habrahabr: 81076 documents, 588400 words (35640 are common with PostNauka), 77102 tags (673 are common with PostNauka)

Heterogeneity of sources

- The size of Habrahabr (# of documents, words and tags) is much larger than of PostNauka.
- Topical structure of collections differ a lot: PostNauka contains more different topics.

Problem relevance

State of the art

1. Topic models are successfully applied for visualization of scientific corpora.
2. Hierarchical TM provide additional tools for visualizing larger heterogeneous collections and are more suited for creating **exploratory search** systems (iris.ai, paperscape.org).

Challenges

1. There is **no common evaluation** measure for hierarchical topic models.
2. Existing methods for building topic models **do not consider heterogeneity** of sources.

The long-term research goal

We want to build a **topical exploratory search engine** for popular scientific first, and then for scientific articles.

This system should have the following properties:

- A convenient hierarchical "knowledge map" can be built with little or no human supervision.
- Each new source can extend the map both "in breadth" and "in depth".
- The search can be performed by using text queries or documents (abstracts, essays, or articles).
- Search results should be displayed as a set of "regions" on the map.

Current results: Russian knowledge map

Humanitarian topics consist mostly of PostNauka documents, and Habrahabr documents are added into technical topics. First level topics are mainly influenced by PostNauka.

Current results: symbiosis of different sources

Below are subtopics of a topic *psychology, internet, and intellection.*

A subtopic *psychology, internet, and social networks* is created after the proposed algorithm is applied to the base PostNauka model and contains a large amount of Habrahabr documents.

Presentation plan

1. Literature review
2. Quality of topical edges
3. Construction of heterogeneous models
4. Demonstration of exploratory search engine

Probabilistic topic modeling

Given: W — dictionary of tokens w
D — collection of documents $d = \{w_1, ..., w_{n_d}\}$
Matrix $F = \{n_{dw}\}_{W \times D}$
n_{dw} — amount of times w occurred in d
T — a set of topics

Find: Matrices $\Phi = \{\phi_{wt}\}_{W \times T}$, $\Theta = \{\theta_{td}\}_{T \times D}$
$\phi_{wt} = p(w|t)$ — probability of token w in topic t
$\theta_{td} = p(t|d)$ — probability of topic t in document d

From Bayes' formula $p(w|d) = \sum_{t \in T} \phi_{wt}\theta_{td}$

This is a matrix factorization problem $F = \Phi\Theta$!

Probabilistic topic modeling

- This problem has infinitely many solutions of kind $\Phi\Theta = (\Phi S)(S^{-1}\Theta) = \Phi'\Theta'$, where S is a matrix of rank $|T|$
- We can introduce regularization of Φ and Θ

PLSA: $R(\Phi, \Theta) = 0$

LDA: $R(\Phi, \Theta) = \sum_{t,w} \beta_w \ln \phi_{wt} + \sum_{d,t} \alpha_t \ln \theta_{td}$

ARTM: $R(\Phi, \Theta) = \sum_{i=1}^{n} \tau_i R_i(\Phi, \Theta)$

ARTM optimization task

$$\sum_{d \in D} \sum_{w \in W} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta},$$

$$\text{w.r.t. } \phi_{wt} \geq 0, \ \theta_{td} \geq 0, \ \sum_{w \in W} \phi_{wt} = 1, \ \sum_{t \in T} \theta_{td} = 1.$$

Hierarchical ARTM

Given: Φ^l, Θ^l — parameters of $l-$th hierarchy level
A – a set of topics of $l-$th level
Matrix factorization problem:
$\Phi^l = \Phi^{l+1}\Psi^{l+1}$, where
$\Phi^{l+1} = \{p(w|t)\}_{W\times T}$, $\Psi^{l+1} = \{p(t|a)\}_{T\times A}$

Hierarchical ARTM regularizer

$$R(\Phi,\Psi) = \sum_{a\in A}\sum_{w\in W} n_{wa} \ln \sum_{t\in T} \phi_{wt}\psi_{ta} \to \max_{\Phi,\Psi}.$$

NB: Applying this regularizer is equivalent to adding $|A|$ pseudodocuments into the collection. Columns of Ψ form $|A|$ additional columns in Θ.

Quality of flat topic models

Quality evaluation for a topic t in a flat model

$$\text{Quality}(t) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(w_i^{(t)}, w_j^{(t)}),$$

$f(w_i^{(t)}, w_j^{(t)})$ is a cooccurrence measure of top tokens $w_i \in t$ and $w_j \in t$.

Different versions of $f(w_i, w_j)$ are presented in the literature:

Newman et al, 2010: $\ln \frac{p(w_i,w_j)}{p(w_i)p(w_j)}$

Mimno et al, 2011: $\ln \frac{d(w_i,w_j)+\epsilon}{d(w_i)}$

Nikolenko et al, 2015: $\ln \frac{\sum_d \text{tfidf}(w_i,d)\text{tfidf}(w_j,d)+\epsilon}{\sum_d \text{tfidf}(w_i,d)}$

Nikolenko et al, 2016: $\langle v_{w_i}, v_{w_j} \rangle$

Presentation plan

1. Literature review
2. Quality of topical edges
3. Construction of heterogeneous models
4. Demonstration of exploratory search engine

Quality of hierarchical topic models

Quality evaluation for a topic t in a flat model

$$\text{Quality}(t) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(w_i^{(t)}, w_j^{(t)}),$$

$f(w_i^{(t)}, w_j^{(t)})$ is a cooccurrence measure of top tokens $w_i \in t$ and $w_j \in t$.

There is no standard quality measure for hierarchical models.
We propose quality measures for topical edges and ways to aggregate them in a hierarchical model.

Quality evaluation for an edge (a, t) in a hierarchical model

$$\text{Quality}_e(a, t) = \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} f(w_i^{(a)}, w_j^{(t)}),$$

$f(w_i^{(a)}, w_j^{(t)})$ is a cooccurrence measure of top tokens $w_i \in a$ and $w_j \in t$.

Proposed quality measures for an edge (a, t)

EmbedSim: $$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\langle v(w_i^{(a)}), v(w_j^{(t)})\rangle,$$

$v(w)$ — vector representation of a token w.

CoocSim: $$\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\ln\frac{d(w_i^{(a)}, w_j^{(t)})+\varepsilon}{d(w_j^{(t)})},$$

$d(w_i, w_j)$ — coccurrence of tokens w_i and w_j.

HellingerSim: $$1-\frac{1}{\sqrt{2}}\|\sqrt{p(w|a)}-\sqrt{p(w|t)}\|_2$$

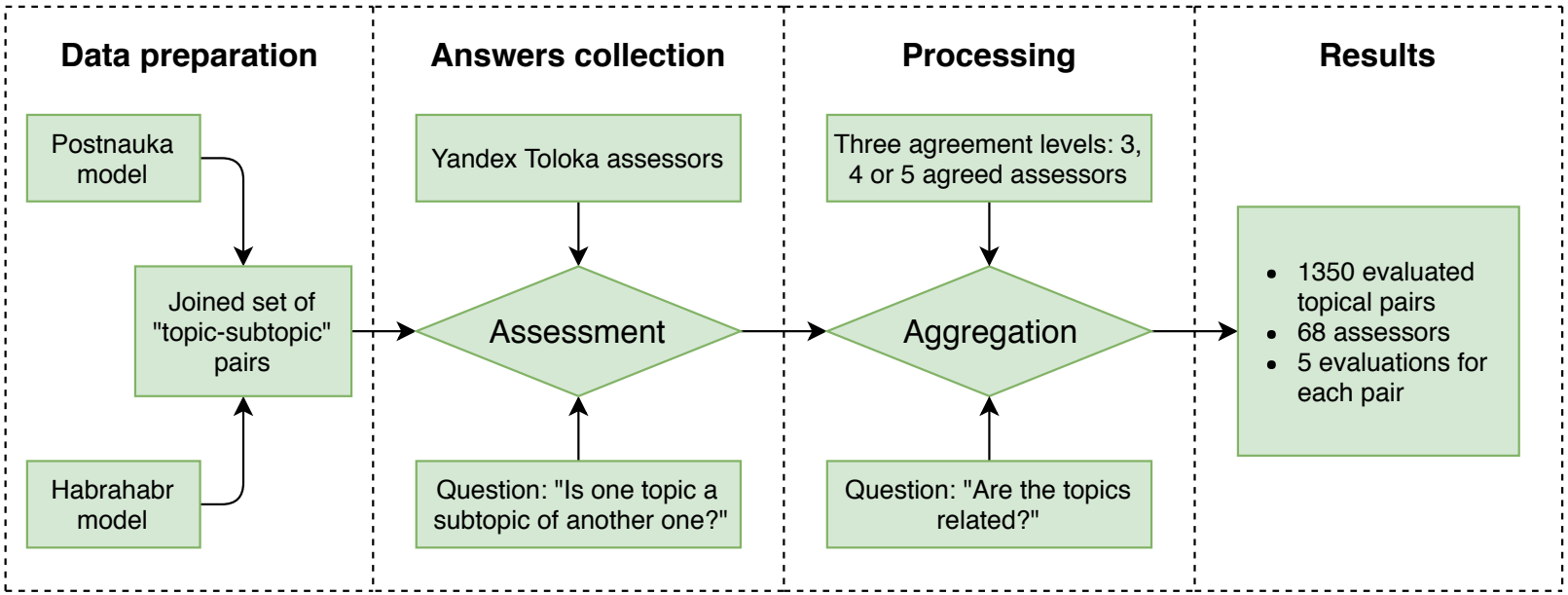
KLSim: $$-D_{KL}(p(w|a)\|p(w|t))$$

Experimentation plan

1. Assessment evaluation
2. Validation of proposed measures
3. Baseline algorithm
4. Proposed algorithm
5. Comparison of algorithms

Assessment evaluation

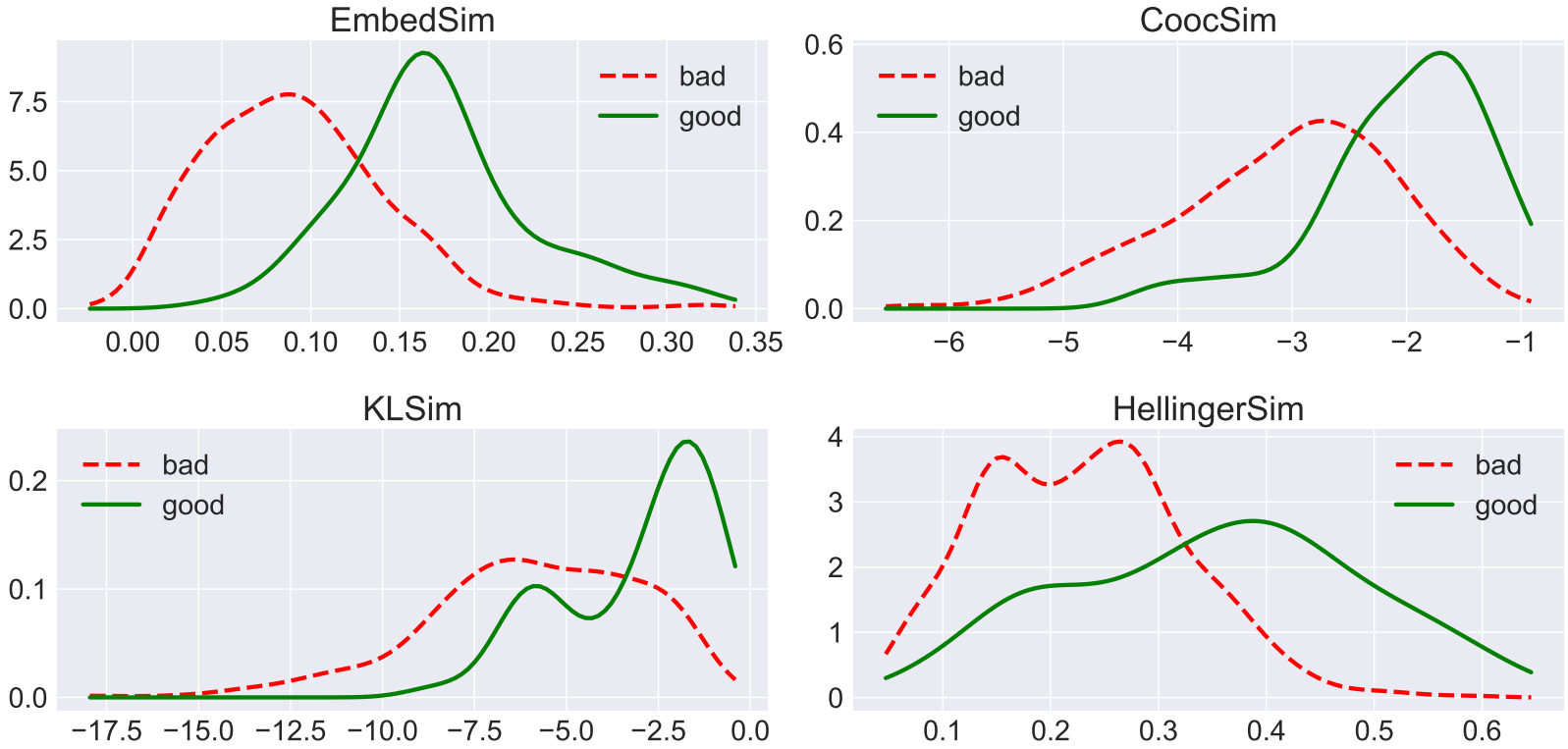
We have collected assessment evaluation using the following scheme:



We have obtained the set of edges (pairs of topics from neighboring levels), labeled as “good” (four or more assessors think that a pair of topics is connected) or “bad” (otherwise).

Validation of proposed measures

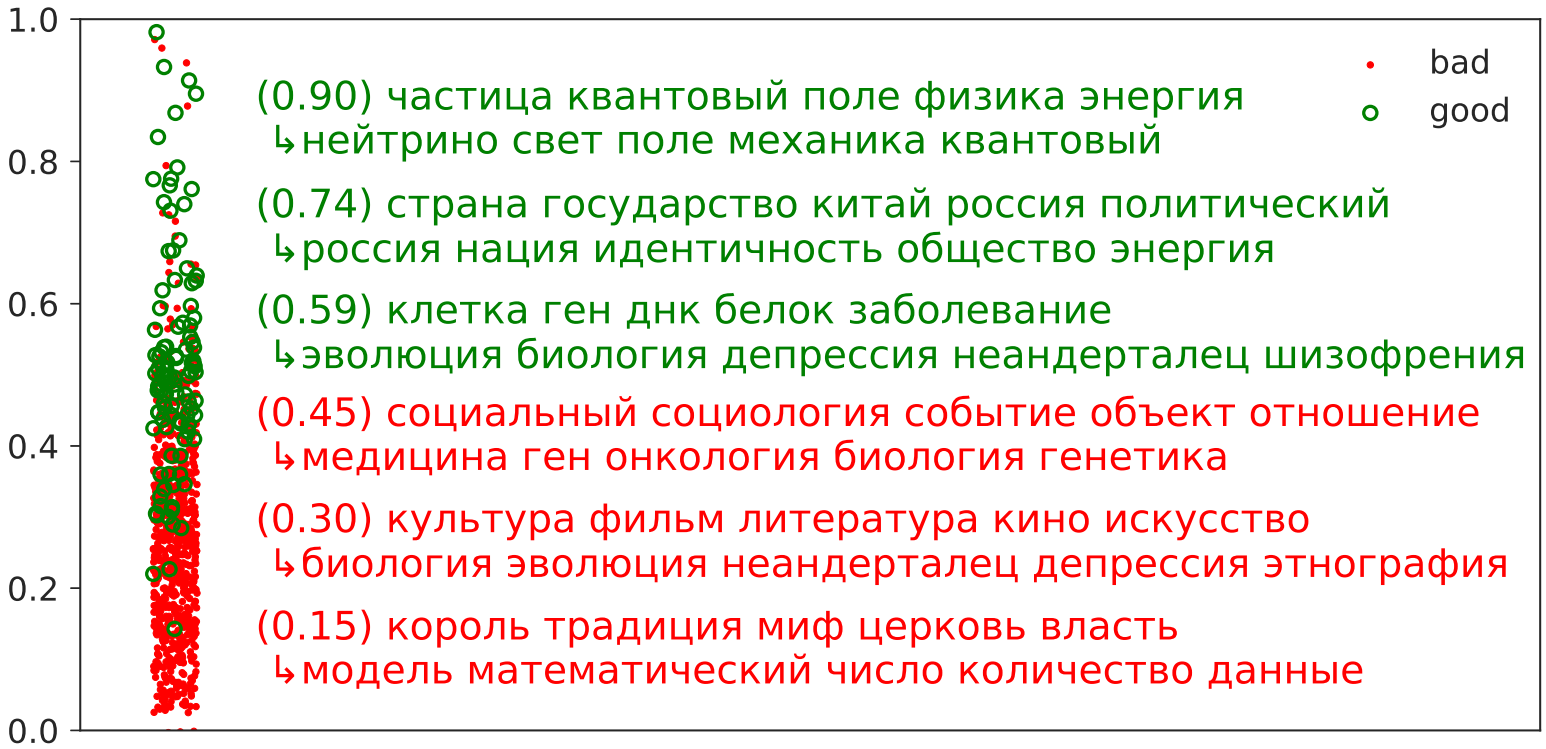
On the figures below are probability distributions of "good" and "bad" edges for the proposed measures.



Best separation of "good" and "bad" edges is achieved with EmbedSim measure (AUC=0.878), it will be used henceforth.

Interpretability of EmbedSim measure

Several examples of topical pairs, which were labeled by assessors as “good” and “bad”, and the corresponding values of EmbedSim.



“Good” are pairs for which four or more assessors agreed that there is a connection between topics.

Presentation plan

1. Literature review
2. Quality of topical edges
3. Construction of heterogeneous models
4. Demonstration of exploratory search engine

Baseline algorithm (concat)

As a baseline algorithm we consider construction of a topic model on a concatenated collection.

Challenges of the baseline algorithm

- Almost all first-level topics contain 90% documents from Habrahabr.
- PostNauka-specific topics are not created.
- Model construction over large corpus takes a long time.

Baseline algorithm does not solve the stated problem!

Proposed algorithm (heterogeneous)

Φ_0^1 – base collection model $p(w|a)$ matrix (in our case, PostNauka model), D_1 – new collection to be added to the model (Habrahabr).

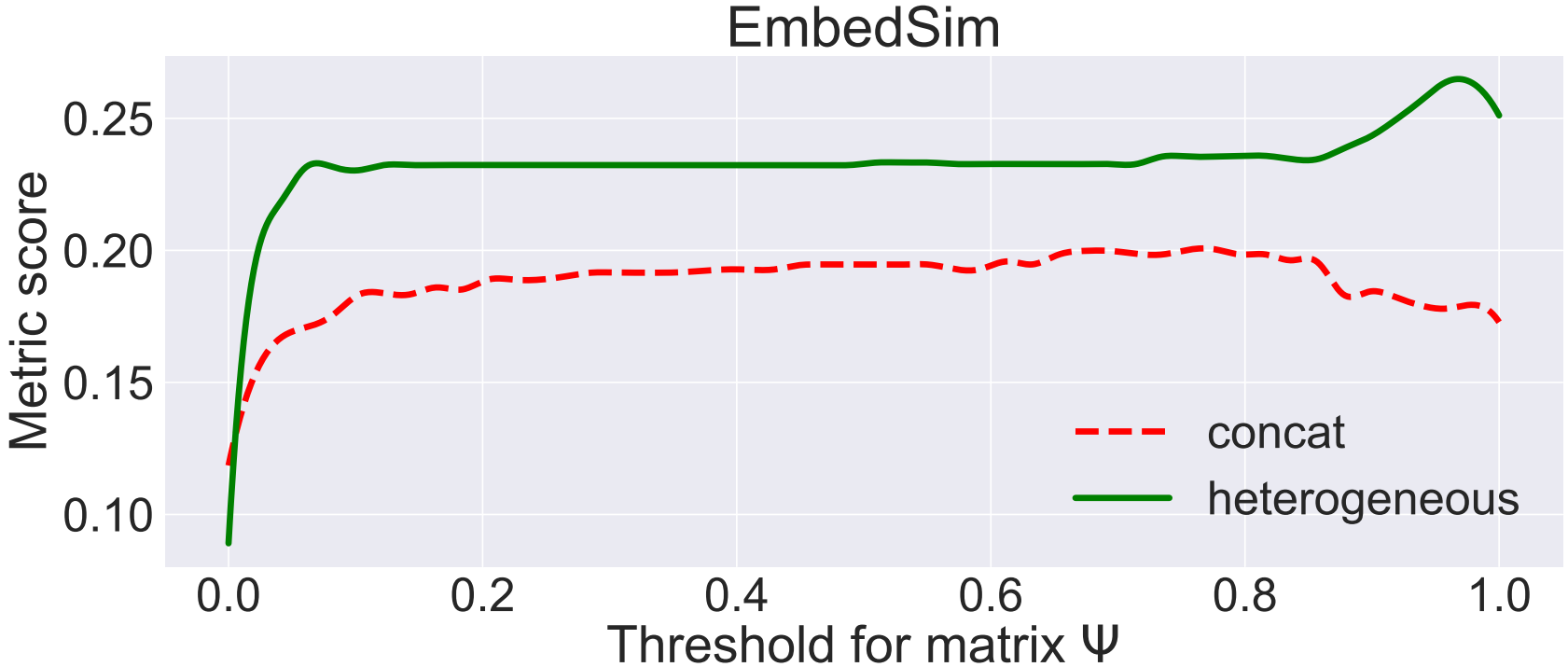
Filtration of D_1 is ranking of the new collection documents according to their similarity to the base collection. The most similar documents should be ranked first.

For i = 1, ..., N:

1. Add the documents that appeared to be on the top of the ranking list to the base collection in a quantity not exceeding 10% of the base collection size.
2. Initialize a new $p(w|a)$ matrix Φ_i^1 of the first hierarchy level with the Φ_i^1 matrix from the previous step;
3. Build a hierarchical model.

Comparison of algorithms: averaging quality

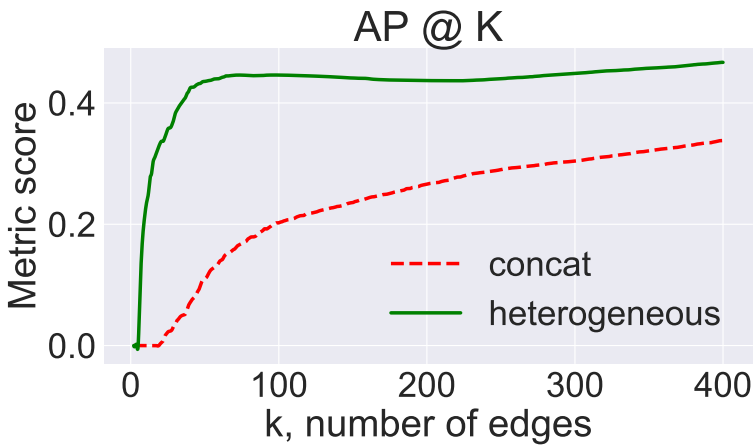
Average edge quality measured with EmbedSim for the baseline (`concat`) and the proposed (`heterogeneous`) algorithms over the $\Psi(t|a)$ threshold value needed to include an edge to hierarchy.

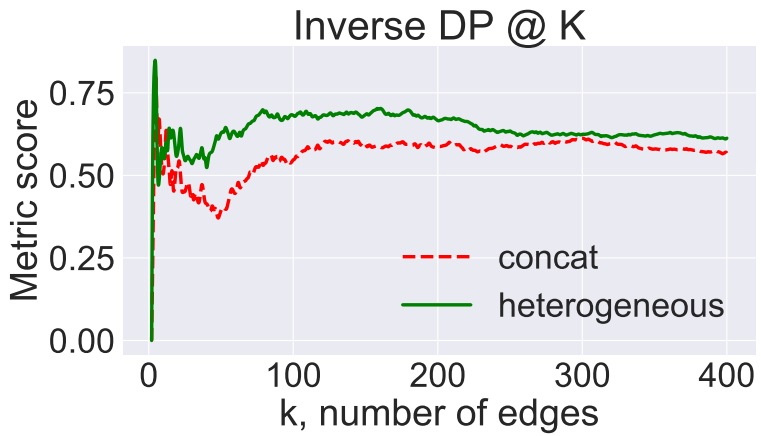


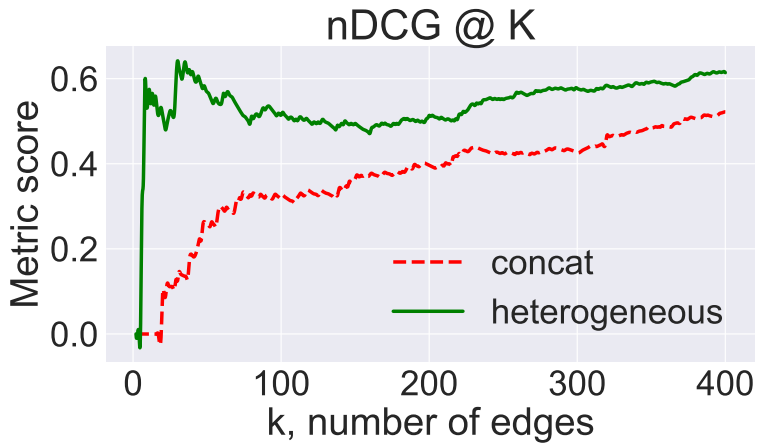
The proposed algorithm gives higher average quality values of the hierarchy edges uniformly along the Ψ threshold.

Comparison of algorithms: ranking quality

Ranking quality of Ψ values for the baseline (`concat`) and the proposed algorithm (`heterogeneous`) if the correct ranking is given by the EmbedSim measure.







Presentation plan

1. Literature review
2. Quality of topical edges
3. Construction of heterogeneous models
4. Demonstration of exploratory search engine

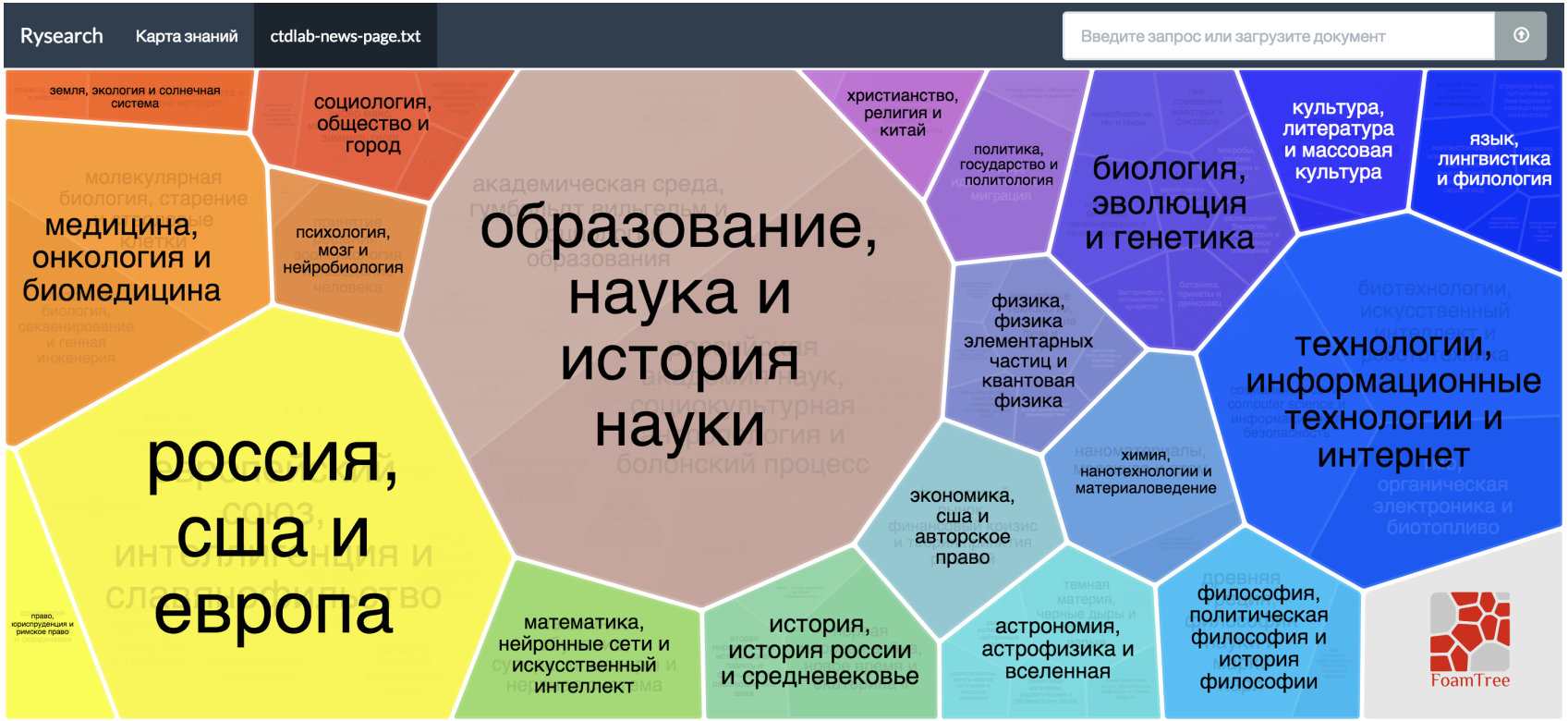
Rysearch: map view

The main screen contains hierarchical knowledge map, built using the proposed algorithm.

The user can investigate regions of interest, going deeper into the map from topics to documents.

Rysearch: document search

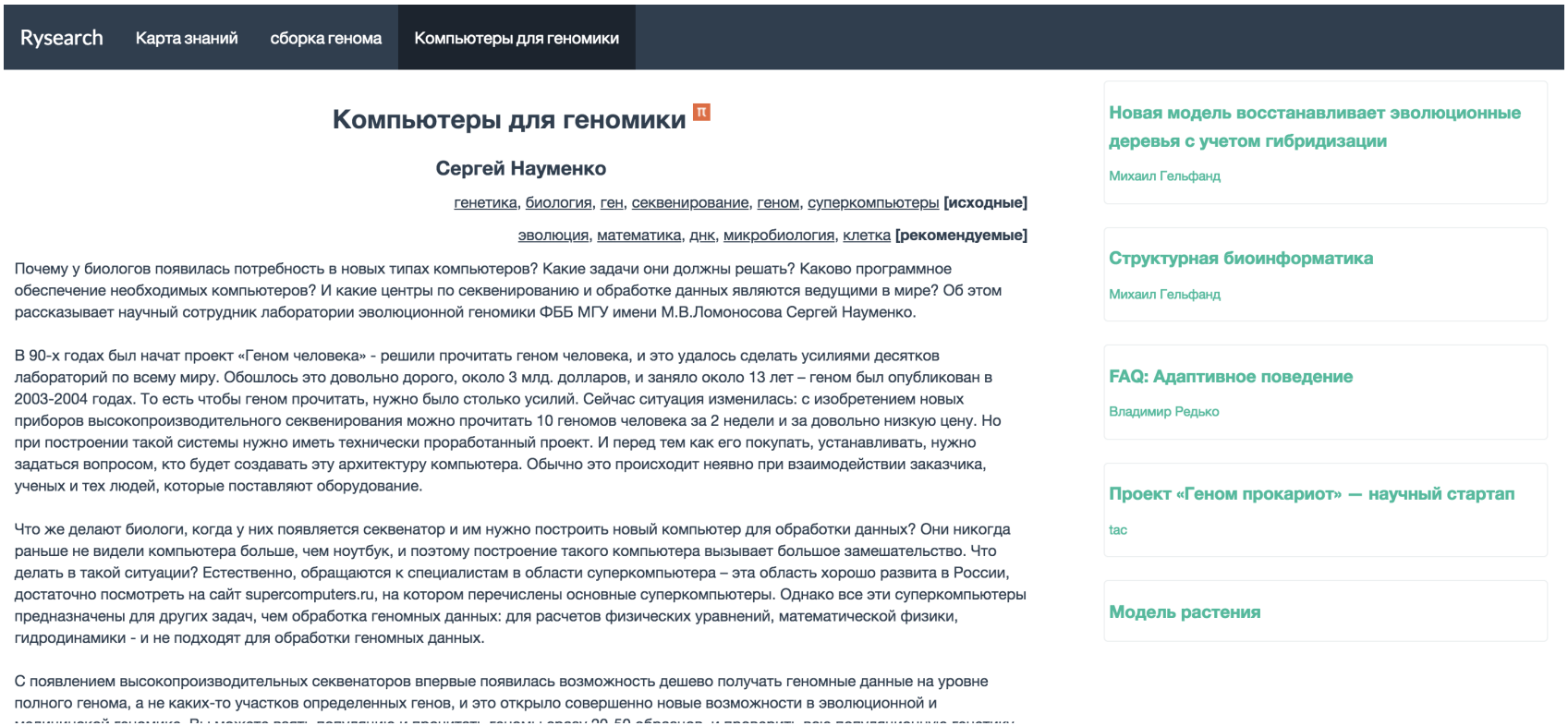
Instead of scanning through the whole map, the user can narrow down the region of interest by uploading a document.



Topics discussed in the document will be highlighted on the map.

Rysearch: document view

The lower level on the knowledge map consists of documents.



Navigation on this level becomes “horizontal” with the recommendation block of most topically related documents.

Conclusion

1. We have proposed quality measures for hierarchical topic models which are consistent with human understanding of topical relatedness.
2. We have proposed the iterative algorithm for building hierarchical TM over heterogeneous collections and have shown its advantage over the baseline approach.
3. We have implemented `Rysearch` system to demonstrate the applicability of hierarchical TM for creating exploratory search engines.

Implementation and experimentation code are available at:
`https://github.com/AVBelyy/Rysearch`

Publications

1. Belyy, A. V., Seleznova, M. S., Sholokhov, A. K., & Vorontsov, K. V. Automatic quality improvement for hierarchical topic modeling. In preparation.
2. Belyy, A. V., Seleznova, M. S., Sholokhov, A. K., & Vorontsov, K. V. (2018). Quality evaluation and improvement for hierarchical topic modeling. In *Computational Linguistics and Intellectual Technologies (pp. 110-123)*.
3. Belyy, A. V., Seleznova, M. S., Sholokhov, A. K., & Vorontsov, K. V. (2017). Aggregation of heterogeneous popular-scientific sources in hierarchical topic model. In proceedings of *60th Scientific MIPT conference (p. 90)*.
    - Best paper award.