ADDITIVELY REGULARIZED HIERARCHICAL TOPIC MODELS

NADIA CHIRKOVA, NADIINCHI@GMAIL.COM

WHAT IS TOPIC HIERARCHY

Topic hierarchy is an oriented multipartite graph of topics that characterises the topical structure of document collection.

Each topic t is represented by a set of terms.

The hierarchy helps to navigate over document collection and to understand how big topics are divided into smaller ones.

BASE TOPIC MODEL

Given:

D is a set of documents, W is a set of terms, n_{dw} is a frequency of term w in document d.

Find model:

$$p(w|d) = \sum_{t\in T} p(w|t)p(t|d) = \sum_{t\in T} \varphi_{wt}\theta_{td}$$

with parameters φ_{wt}, θ_{td}:

- $\varphi_{wt} = p(w|t)$ is a distribution over terms in topic t;
- $\theta_{td} = p(t|d)$ is a distribution over topics in document d.

Optimization task is regularized likelihood maximization:

$$\underbrace{\sum_{d\in D}\sum_{w\in d} n_{dw} \ln \sum_{t\in T} \phi_{wt}\theta_{td}}_{L(\Phi,\Theta)} + \underbrace{\sum_{i=1}^{n} \tau_i R_i(\Phi,\Theta)}_{R(\Phi,\Theta)} \to \max_{\Phi,\Theta}$$

where $R(\Phi,\Theta)$ is a weighted sum of regularization criteria.

HIERARCHY REGULARIZER

The hierarchy is built level by level. First level is a plain flat model. To construct next levels, hierarchy regularizer is used. Denote T is a set of previous level (parent) topics, S is a set of current level topics, $\psi_{ts} = p(t|s)$. The aim is to decompose already built parent topic-doc matrix: $\Theta^{par} \approx \Psi\Theta$.

Hierarchy regularizer:

$$R_1(\Phi,\Theta,\Psi) = \lambda \sum_{t\in T}\sum_{d\in D} \theta^{par}_{td} \ln \sum_{s\in S} \psi_{ts}\theta_{sd}$$

MODEL TRAINING

Applying Karush–Kuhn–Tucker theorem, one can obtain the following EM-algorithm:

E-step:

$p(s|d,w) \propto \varphi_{ws}\theta_{sd}$

$p(s|t,d) \propto \psi_{ts}\theta_{sd}$

M-step:

$$n_{ws} = \sum_{d\in D} n_{dw} p(s|d,w)$$

$$n^1_{sd} = \sum_{w\in W} n_{dw} p(s|d,w)$$

$$n_{ts} = \sum_{d\in D} \theta^{par}_{td} p(s|t,d)$$

$$n^2_{sd} = \sum_{t\in T} \theta^{par}_{td} p(s|t,d)$$

$$\phi_{ws} \propto \left(n_{ws} + \phi_{ws}\frac{\partial R}{\partial \phi_{ws}}\right)_+$$

$$\psi_{ts} \propto \left(n_{ts} + \psi_{ts}\frac{\partial R}{\partial \psi_{ts}}\right)_+$$

$$\theta_{sd} \propto \left(n^1_{sd} + \lambda n^2_{sd} + \theta_{sd}\frac{\partial R}{\partial \theta_{sd}}\right)_+$$

SPARSING REGULARIZERS

On the each level topics are divided into two groups: *domain* and *background* topics. Second group collects common lexis for current level or for the whole collection; first group is for domain-specific lexis.

- Φ *Sparsing*. Each domain topic contains small number of domain-specific terms, while background topics contain the majority of terms:

  $$R_2(\Phi) = -\sum_{s\in S^{dom}} KL(\alpha\|\varphi_s) + \sum_{s\in S^{bcg}} KL(\alpha\|\varphi_s)$$

  α is uniform or prior collection distribution over terms.

- Φ *Decorrelating*. All topics on one level are significantly different:

  $$R_3(\Phi) = -\sum_{s\in S}\sum_{s'\in S\setminus s}\sum_{w\in W} \phi_{ws}\phi_{ws'}$$

- Θ *Sparsing*. Each document is related to a few number of domain topics, but it must be related to the background topic:

  $R_4(\Theta) = -\sum_{d\in D} KL(\beta\|\theta_d)$.

QUALITY MEASURES

The quality of hierarchy is measured per level. Criteria:

1. **Size of topic kernel**:

   size = $|W_t|$, $\quad W_t = \{w \colon p(t|w) > 0.25\}$

2. **Topic contrast**: $\frac{1}{|W_t|}\sum_{w\in W_t} p(t|w)$

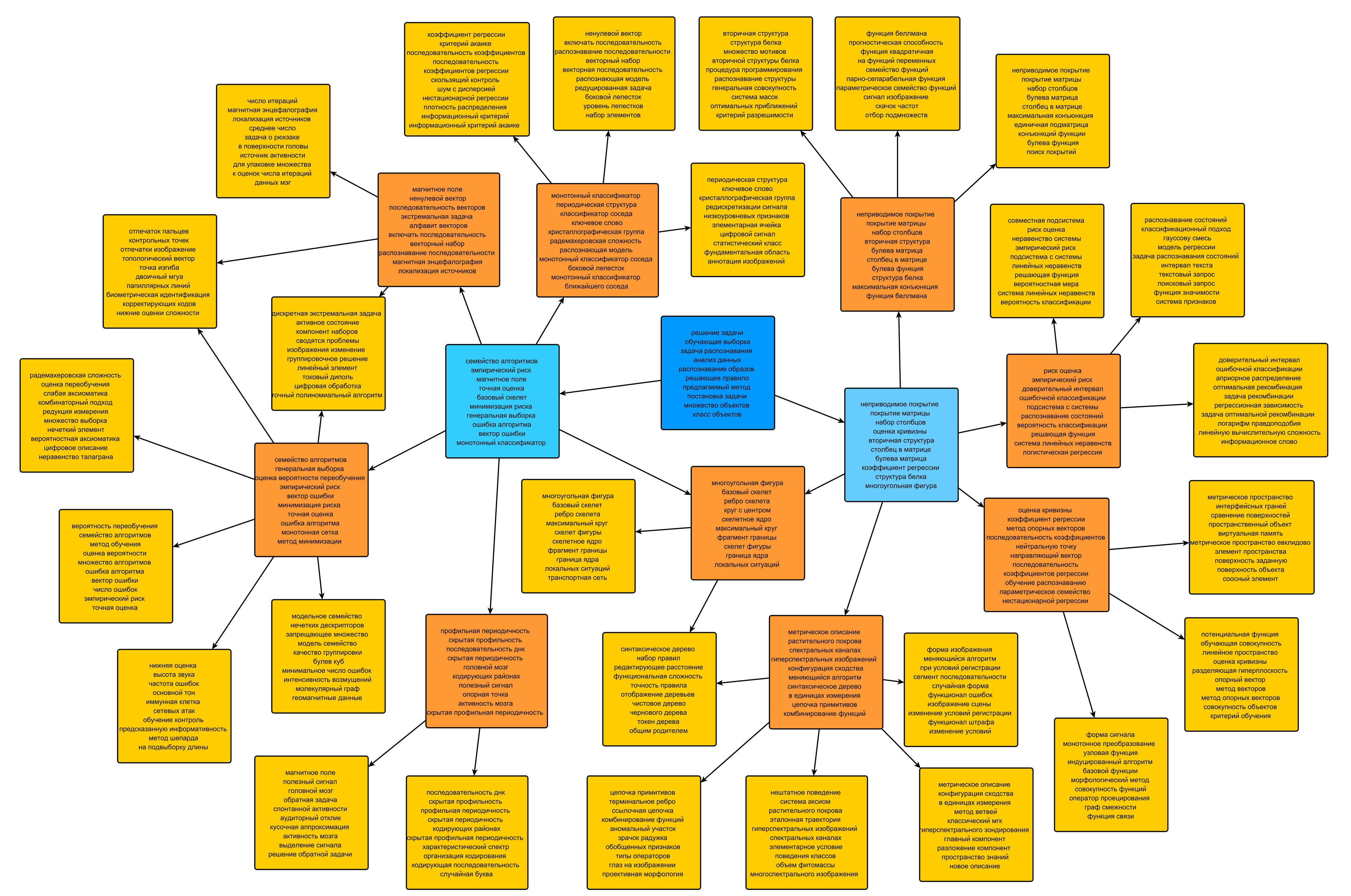
3. **Topic purity**: $\sum_{w\in W_t} p(w|t)$

4. **Topic coherence**:

   $$\frac{2}{k(k-1)}\sum_{i=1}^{k}\sum_{j=1}^{i-1} PMI(w_i,w_j),$$

   where terms in t are sorted by $p(w|t)$.

HIERARCHY OF MMPR&IIP CONFERENCES (WWW.MMRO.RU)



EXPERIMENTS

Text dataset came from two Data Analysis conferences: *Mathematical Methods of Pattern Recognition* and *Intellectualization of Information Processing*, $|D| = 850$, $|W| = 42000$. To increase interpretability n-grams are used (they are collected using external software).

Comparison of flat and hierarchical models

model	purity	contrast	coherence
flat	0.999	0.961	1.063
hier	0.998	0.959	1.211

REFERENCES

Vorontsov K. V., Potapenko A. A. Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization. — Analysis of Images, Social Networks, and Texts (AIST-2014). — LNCS, Springer.