Additive regularization for hierarchical multimodal topic modeling*

***N. A. Chirkova*[1,2] *and K. V. Vorontsov*[3]**
nadiinchi@gmail.com; vokov@forecsys.ru
[1]JSC Antiplagiat, 33 Nagatiskaya Str., Moscow, Russia
[2]Lomonosov Moscow State University, 1 Leninskie Gory, Moscow, Russia;
[3]Federal Research Center "Computer Science and Control" of RAS
44/2 Vavilova Str., Moscow, Russia

Probabilistic topic models uncover the latent semantics of text collections and represent each document by a multinomial distribution over topics. Hierarchical models divide topics into subtopics recursively, thus simplifying information retrieval, browsing and understanding of large multidisciplinary collections. The most of existing approaches to hierarchy learning rely on Bayesian inference. This makes difficult the incorporation of topical hierarchies into other types of topic models. The authors use non-Bayesian multicriteria approach called Additive Regularization of Topic Models (ARTM), which enables to combine any topic models formalized via log-likelihood maximization with additive regularization criteria. In this work, such formalization is proposed for topical hierarchies. Hence, the hierarchical ARTM (hARTM) can be easily adapted to a wide class of text mining problems, e. g., for learning topical hierarchies from multimodal and multilingual heterogeneous data of scientific digital libraries or social media. The authors focus on topical hierarchies that allow a topic to have several parent topics which is important for multidisciplinary collections of scientific papers. The regularization approach allows one to control the sparsity of the parent–child relation and automatically determine the number of subtopics for each topic. Before learning the hierarchy, it is necessary to fix the number of topics for each layer. The additive regularization does not complicate the learning algorithm; so, this approach is well scalable on large text collections.

Keywords: *topic modeling; ARTM; topic hierarchies; regularization*

DOI: 10.21469/22233792.2.2.05

1 Introduction

Topic modeling is a popular technique for semantic analysis of text collections. A probabilistic topic model defines each topic by a probability distribution over words and describes each document by a probability distribution over topics. In large text collections such as digital libraries or social media archives, the topics are usually organized in a hierarchy. Topic hierarchy helps user to navigate through the collection: going down the hierarchy, user chooses subtopics and finds a small subset of documents to read.

In last years, a lot of research was done about topic hierarchies learning. There is no common definition and common quality measure of topic hierarchy in the literature. Also, there is still no common hierarchy learning approach [1].

It is difficult to combine existing approaches with other modifications of topic models: spatiotemporal [2], short text [3], multilingual [4], multimodal [5], semisupervised [6], decorrelated [7], sparse [8], etc. On the other hand, there is a general approach for combining different types of topic models called additive regilarization of topic models [9, 10]. This framework is

*The research was supported by the Russian Foundation for Basic Research (grants 16-37-00498, 14-07-00847, and 14-07-00908).

well scalable for large collections [10] and is implemented in open-source topic modeling library BigARTM.

The goal of this work is to propose a method of learning topic hierarchies via topic model regularization and integrate it with ARTM.

Let us focus on hierarchies as multipartite (multilevel) directed acyclic graph of topics rather than a topic tree. While the last definition is a mainstream in literature, an assumption that a topic can inherit from several parent topics looks more reasonable. It is a common case in any field of knowledge when specific topic occurs on the edge of two or even more parent topics. For example, bioinformatics combines applied mathematics and computer science to solve the biology problems. This situation is called multiple inheritance. The presented approach supports multiple inheritance and controllable sparsing of topic graph and automatically determines the number of subtopics for each topic.

The reminder of the paper is organized as follows. In section 1, an overview of existing approaches for learning hierarchies is presented. In section 2, a formal problem statement is given and then, in section 3, the present authors' approach is described and in section 4, its implementation in BigARTM is presented. The last two sections are about experiments and discussion.

2 Related Work

Two basic topic modeling techniques are probabilistic latent semantic analysis (PLSA) [11] and its Bayesian extension latent Dirichlet allocation (LDA) [12]. A lot of LDA modifications were developed to meet applications tasks [13].

Additive regularization of topic models [10] is non-Bayesian extension that allows to impose additional, problem-specific criteria on topic model parameters. Many of LDA expansions can be interpreted as regularization criteria, this allows to combine several modifications in a single model.

In hierarchal models, the topics are linked by parent–child relations. Topic hierarchies are usually constructed in two ways: via generative model complication or as a combination of several tied flat models. Hierarchical LDA (hLDA) [14] and hierarchical Pachinko allocation model (hPAM) [15] are the examples of generative models. As other LDA extensions, these models are trained using time consuming Gibbs sampling that limits available collection size [16] and integration with other types of topic models. Hierarchical LDA is a tree structure and hPAM is a directed acyclic multilevel graph with no tools for edges number reduction.

The second group is split into top-down and bottom-up approaches. Tree structured hierarchies are often learned top-down recursively: firs, a flat model with few topics is learned and then, process repeats for each subtopic. SplitLDA splits documents between topics accordingly to the distribution over topics for each document–word pair [17]. Constructing A Topical HierarchY (CATHY) approach [18] operates with phrases rather than with words and divides them between subtopics. In Scalable and Robust Construction of Topic Hierarchies (STROD) [16], each topic distribution over words can be expanded to a mixture of suptopics distributions using tensor decomposition algorithm. The drawback of recursive approaches is that they need heuristics to determine the number of subtopics in each topic. On the other hand, recursive learning is usually fast, STROD is proven [16] to be the fastest on large collections.

Multiple inheritance supporting hierarchies are usually learned level by level. In [19], the hierarchy is learned in two steps: first, flat LDA models are learned for each level and next, topics between levels are linked using special subsumption criteria. An advantage is that changing a threshold in subsumption criteria controls the hierarchy sparsity. The disadvantage

is that specific topics are modeled independently from their parent topics. Also, the authors propose a simple agglomerative clustering based method for determining the number of topics in levels.

In [1], the hierarchy is constructed by bottom-up strategy. The last level of topics is learned first and then, these topics are treated as pseudodocuments and the next level model is learned from them. In this case, subtopic-pseudodocument proportions specify the topic graph structure and there is no ability to control the graph sparsity.

Almost all hierarchical topic models are based on Bayesian inference, it makes difficult to combine other topic model modifications with hierarchy. The present authors propose a top-down hierarchy learning framework based on ARTM that incorporates few reasonable ideas from other approaches.

3 Problem Statement

Let D denote the text collection. Documents $d \in D$ may contain not only words but also other elements such as tags, links, location marks, etc. Let us refer to such types of elements as modalities. For example, a scientific paper usually contains three modalities: text, keywords, and references. Let M denote a set of all modalities in the collection. Modalities $m \in M$ are defined by disjoint dictionaries $W = \bigsqcup_{m \in M} W^m$.

A document $d \in D$ is a sequence of n_d elements: $(w_1, w_2, w_3, \dots)$, $w_i \in W$. In this paper, an order of elements is not important. Thu, collection can be represented as a counters matrix $\{n_{dw}\}_{D \times W}$ where n_{dw} is the number of w occurencies in d.

Given the text collection, the goal is to organize its documents into comprehensive hierarchical structure. Let us define a *topic hierarchy* as an oriented multiparticle (multilevel) acyclic graph of topics so that the edges connect the topics from the neighboring levels. If there is an edge $a \to t$ in the hierarchy, then the topic a is called *parent*, or *ancestor*, topic and t is called *child topic*, or *subtopic*. The parent topic is divided into several more specific child topics. The number of topics in each following (child) level is usually greater than in the previous (parent) level. Zero level consists of only one topic called *root*. An example of topic hierarchy is given in Fig. 1.

Each topic in the hierarchy is associated with distributions over each modality dictionary. This allows one to represent a topic by a top of most probable words saying what this topic is about. The same can be done with other modalities.

To construct the hierarchy, let us learn several flat topic models and tie them via regularization.

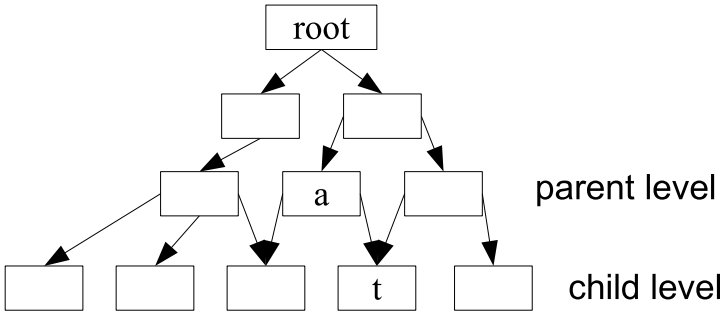


Figure 1 An example of topic hierarchy

In the rest of the paper, an operator $\underset{i\in I}{\text{norm}}[y_i] = \max\{y_i, 0\}/\sum_{i'\in I}\max\{y_{i'}, 0\}$ transforming real vector $(y_i)_{i\in I}$ to a probability distribution is used.

4 hARTM framework

4.1 ARTM: flat topic models

The flat topic model describes collection D by finite topics set T. In ARTM [10], document distribution over each modality is modeled as a mixture of topic distributions:

$$p(w|d) \approx \sum_{t\in T} p(w|t)p(t|d) \quad d \in D,\, w \in W^m.$$

In other words, for each modality m, the topic model is a low-rank approximation

$$F^m \approx \Phi^m\Theta$$

of the frequency matrix $F^m = \{f_{wd}\}_{W^m\times D}$ where $f_{wd} = \underset{w\in W^m}{\text{norm}}[n_{dw}]$ is the frequency $p(w|d)$. The model parameters are the matrices $\Phi^m = \{\varphi_{wt}\}_{W^m\times T}$ with $\varphi_{wt} = p(w|t)$ and $\Theta = \{\theta_{td}\}_{T\times D}$ with $\theta_{td} = p(t|d)$, Φ and Θ being the stochastic matrices:

$$\sum_{w\in W^m} \varphi_{wt} =; \quad \sum_{t\in T} \theta_{td} = 1. \tag{1}$$

For brevity, let is denote vertically stacked Φ^m and F^m, $m \in M$, by Φ and F, respectively. Then, the topic model in an approximate matrix factorization $F \approx \Phi\Theta$.

Let us maximize the weighted sum of modality log-likelihoods and regularizers R_i to learn Φ and Θ :

$$\sum_{m\in M} \varkappa_m \sum_{d\in D}\sum_{w\in W^m} n_{dw} \ln \sum_{t\in T} \varphi_{wt}\theta_{td} + \sum_i \tau_i R_i(\Phi,\Theta) \to \max_{\Phi,\Theta}. \tag{2}$$

Weights $\varkappa_m$ are used to balance log-likehood of modalities. Regularizers R_i impose additional problem-specific criteria on the model parameters. Regularizer coefficients τ_i balance the importance of regularizers and log-likehoods. If the regularizer term $R = \sum_i \tau_i R_i(\Phi,\Theta)$ equals zero and there is only text modality, then described model simplifies to PLSA.

Theorem 1 (see [10]). *If all regularizers are continuously differentiable on Φ and Θ, then the stationary point of the problem* (2) *with constrains* (1) *satisfies the following system yielding expectation-maximization (EM) algorithm for model training:*

$$\left.\begin{aligned}
&\textit{E-step}: \quad p(t|d,w) = \underset{t\in T}{\text{norm}}[\varphi_{wt}\theta_{td}], \quad w \in W,\, d \in D;\\
&\textit{M-step}: \quad \varphi_{wt} = \underset{w\in W^m}{\text{norm}}\left[n_{wt} + \frac{\partial R}{\partial \varphi_{wt}}\varphi_{wt}\right], \quad n_{wt} = \sum_{d\in D} n_{dw}p(t|d,w),\\
&\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad w \in W^m, t \in T, m \in M;\\
&\qquad\quad \theta_{td} = \underset{t\in T}{\text{norm}}\left[n_{td} + \frac{\partial R}{\partial \theta_{td}}\theta_{td}\right], \quad n_{td} = \sum_{w\in W} n_{dw}p(t|d,w), \quad t \in T, d \in D.
\end{aligned}\right\} \tag{3}$$

The EM-algorithm is obtained by applying the fixed point iteration method to the system. Matrices Φ and Θ are initialized randomly.

Sparsing regularizers. Frequently used sparsing regularizer [10] causes distributions $p(w|t)$ and $p(t|d)$ to be sparse meaning the majority of distribution domain elements have zero probability. To do this, Kullback–Leibler divergence between specified distribution $\boldsymbol{\alpha}$, usually uniform, and target distribution is maximized. For instance, Θ-sparsing regularizer:

$$\sum_{d\in D} KL(\boldsymbol{\alpha}\|\boldsymbol{\theta}_d) \to \max_{\Theta} \quad \Leftrightarrow \quad R_1(\Theta) = -\sum_{d\in D}\sum_{t\in T} \alpha_t \ln \theta_{td} \to \max_{\Theta},$$

$\boldsymbol{\theta}_d$ denotes Θ column and for uniform distribution, $\alpha_t = 1/|T|$. Similarly, for Φ^m sparsing with uniform specified distribution,

$$R_2(\Phi^m) = -\sum_{t\in T}\sum_{w\in W^m} \frac{1}{|W^m|} \ln \varphi^m_{wt} \to \max_{\Phi^m}.$$

Modified M-step formulas for parameters update:

$$\varphi_{wt} = \underset{w\in W^m}{\operatorname{norm}}\left[n_{wt} - \frac{\tau_1}{|W^m|}\right]; \quad \theta_{td} = \underset{t\in T}{\operatorname{norm}}\left[n_{td} - \frac{\tau_2}{|T|}\right]. \tag{4}$$

Hyperparameters of flat topic model are number of topics $|T|$, weights $\{\varkappa_m\}_{m\in M}$, and regularization coefficients $\{\tau_i\}_i$. While learning topic hierarchy, flat topic model is trained for each level of hierarchy, every time with new hyperparameters settings.

4.2 hARTM: Top-down hierarchy learning

Since topic hierarchy is a multilevel graph, let us consider each level as a flat topic model. The authors propose top-down, level by level hierarchy learning algorithm. Zero level is associated with the whole collection. The first level contains small number of major topics. Starting from the second level, it is necessary not only to model the topics, but also to establish parent–child topic relations. To do this, the authors introduce two additional matrix factorization problems and propose two new interchangeable regularizers based on them.

Assume one has already learned $\ell \geqslant 1$ hierarchy levels. Now, let us learn $(\ell+1)$th level that is a child level for the ℓth ancestor level. Not to confuse levels, let us denote parent level topics $a \in A$ and parameters Φ^ℓ and Θ^ℓ instead of $t \in T$, Φ, and Θ used for child level. Note that Φ^ℓ and Θ^ℓ are already modeled.

Φ interlevel regularizer. Let us model parent topic distribution over words and other modalities as a mixture of child topics distributions:

$$p(w|a) = \sum_{t\in T} p(w|t)p(t|a), \quad w \in W^m,\ a \in A.$$

This means an approximation

$$\Phi^\ell \approx \Phi\Psi \tag{5}$$

with new parameters matrix $\Psi = \{\psi_{ta}\}_{T\times A}$, $\psi_{ta} = p(t|a)$ containing *interlevel distributions* of children topics t in parent topics a. This gives the following regularizaion criteria:

$$\sum_{a\in A} n_a\, KL(\boldsymbol{\varphi}^{\ell,m}_a \| \Phi^m\, \boldsymbol{\psi}_a) \to \min_{\Phi^m,\Psi}$$

or, equivalently,

$$R_3(\Phi^m, \Psi) = \sum_{a\in A}\sum_{w\in W^m} n_{wa} \ln \sum_{t\in T} \varphi_{wt}\psi_{ta} \to \max_{\Phi^m,\Psi},$$

$\boldsymbol{\varphi}_a^{\ell,m}$ and $\boldsymbol{\psi}_a$ denote columns of $\Phi^{\ell,m}$ and Ψ, respectively. Weights $n_a = \sum_{w \in W^m} n_{wa}$ are imposed to balance parent opics proportionally to their size and to scale regularization criteria up to the log-likelihood, n_{wa} being the parent topic counters from the EM-algorithm. Regularizer criterias are weighted by the modality weights:

$$R_3(\Phi, \Psi) = \sum_{m \in M} \varkappa_m R_3(\Phi^m, \Psi).$$

This regularizer is equivalent to adding $|A|$ pseudodocuments represented by $\{n_{wa}\}_{W \times A}$ columns. Then, the matrix Ψ forms $|A|$ additional columns to the matrix Θ corresponding to pseudodocuments. Note than child level could not be trained only on pseudodocuments because internal dimension in approximation (5) is higher than the minimum dimension of Φ^ℓ and Φ will just copy columns of Φ^ℓ.

Θ interlevel regularizer. The same idea can be applied for regularizing Θ instead of Φ. Then, for each document, distribution over parent topics is modeled by the mixture of topic distributions:

$$p(a|d) = \sum_{t \in T} p(a|t) p(t|d).$$

Additional matrix approximation looks like

$$\Theta^\ell \approx \widetilde{\Psi} \Theta$$

with interlevel distributions $\widetilde{\Psi} = \{\tilde{\psi}_{at}\}_{A \times T}$, $\tilde{\psi}_{at} = p(a|t)$. This means that parent topic's documents set is a union of children's documents sets. Regularizer criteria is

$$R_4(\Theta, \widetilde{\Psi}) = \sum_{a \in A} \sum_{d \in D} \theta_{ad}^\ell \ln \sum_{t \in T} \tilde{\psi}_{at} \theta_{td} \to \max_{\widetilde{\Psi}, \Theta}.$$

To train child model with the regularizer, let us add a new modality $\tilde{m}$ corresponding to parent topics and consider document counters for this modality θ_{ad}^ℓ. The Θ-regularizer coefficient becomes the modality weight and $\widetilde{\Psi}$ corresponds to the matrix $\Phi^{\tilde{m}}$.

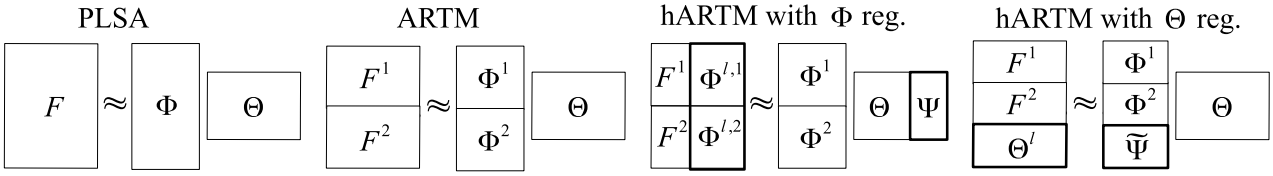


Figure 2 An illustration of child level regularization

An illustration of manipulating with pseudodocuments and new modality while the regularization of child level is given in Fig. 2.

Hierarchy sparsing regularizers. When the topics are allowed to inherit from a number of parents, it is assumed that this number will not be large, i. e., 1–3, rarely greater parents. Such hierarchy is called the *sparse* one. In other words, we want distributions $p(a|t)$ to be sparse. The regularization allows us to achieve this requirement.

Since in Θ interlevel regularization approach $\widetilde{\Psi}$ is a child $\Phi^{\tilde{m}}$ and its columns represent distributions $p(a|t)$, one can use Φ-sparsing regularizer described above to make the hierarchy

sparse. Let us rewrite (4) replacing $\varphi \to \tilde{\psi}$, $w \to a$, and $W^m \to A$ to show how $\widetilde{\Psi}$ updates on each iteration:

$$\tilde{\psi}_{at} = \underset{a \in A}{\text{norm}}\left[n_{at} - \frac{\tau_1}{|A|}\right].$$

In case of Φ interlevel regularization, Ψ columns represent $p(t|a)$ distributions that can be converted to $p(a|t)$ using Bayes formula. Following the idea of other parsing regularizers, let us maximize KL-divergence between uniform distribution $\boldsymbol{\gamma} = \{1/|A|\}_{a \in A}$ and the arget one $\tilde{\boldsymbol{\psi}}_t = \{p(a|t)\}_{a \in A}$:

$$\sum_{t \in T} \mathrm{KL}\,(\boldsymbol{\gamma} \| \tilde{\boldsymbol{\psi}}_t) \to \max_{\Psi}$$

or, equivalently,

$$R_5(\Psi) = \sum_{t \in T} \sum_{a \in A} \frac{1}{|A|} \ln p(a|t) = \frac{1}{|A|} \sum_a \sum_t \ln \frac{\psi_{ta}\, p(a)}{\sum_{a'} \psi_{ta'}\, p(a')} \to \min_{\Psi}.$$

Probabilities $p(a)$ are counted from Θ^ℓ.

To show how Ψ updates, let us rewrite M-step formula in (**??**) replacing $\theta \to \psi$ and $d \to a$ and taking derivatives of $R_5(\Psi)$ with respect to ψ_{ta}:

$$\psi_{ta} = \underset{t \in T}{\text{norm}}\left[n_{ta} - \tau_5\left(\frac{1}{|A|} - p(a|t)\right)\right].$$

For each topic t, parent topics a with high $p(a|t)$ get higher and parents with low $p(a|t)$ get lower. Note that R_5 cannot zeroize all components of Ψ column whereas R_1 can do this with $\widetilde{\Psi}$ column.

Hierarchy learning scenario. Thus, hyperparameters of topic hierarchy are the number of levels, the number of topics on each level, modalities weights, and regularization coefficients. One can learn hierarchy level by level, on each level finding parents for topics from previous level using Φ or Θ interlevel regularizer. If sparse hierarchy is desired, hierarchy sparsing regularizer should also be used. The process of training levels is stopped when the topics on the last level are highly specialized.

Regularization coefficients may be tuned for each level individually or used the same for all levels. Note that when learning the $(\ell + 1)$th level, only ℓth level's topics are used for regularization, not all previous levels' topics.

When hierarchy is learned, the topics on each level are represented by their distributions over words and other modalities. The documents on each level are assigned to several topics with proportions specified in this level's Θ matrix. The hierarchy structure is defined by interlevel distributions. To draw the topic graph, one may impose a threshold on $p(a|t)$ or $p(t|a)$.

5 Implementation in BigARTM

BigARTM is an open-source topic modeling library with C++ kernel [20]. BigARTM provides command line, C++ and python interfaces, and rich built-in library with regularizers and scores. BigARTM takes multimodal input data in a range of formats and transforms it into a series of *batches*, internal format. All batches store about the same number of documents, each batch is assigned a float weight (default 1.0).

BigARTM provides offline and online multithread learning algorithms. Offline algorithm performs a number of scans over the entire collection. During one scan, each thread processes one batch at a time, calculating n_{td} and θ_{td} (applying Θ-regularizers) and contributing local, batch-specific n_{wt} multiplied by batch weight to global n_{wt} counters. After that, the scan algorithm applies Φ-regularizers to global n_{wt} and normalizes them to calculate Φ. Online algorithm improves the convergence rate by recalculating Φ after every portion of batches.

The hierarchy learning is implemented as a wrapper over library interface without changing the kernel. To use Φ interlevel regularizer, an additional batch is created from parent Φ matrix, the weight of this batch equals to regularization coefficient. This parent batch is appended to the collection batches during the learning of child level, it does not affect algorithm efficiency.

To use Θ interlevel regularizer during child level learning, each batch should be appended the new modality corresponding to current batch parent Θ. This is time consuming operation. In experiments, it will be shown that two proposed interlevel regularizers are interchangeable; so, there is no need to use ineffective algorithm.

The Ψ sparsing regularizer is implemented as usual Θ regularizer since Ψ is the parent batch Θ.

6 Experiments

In this section, two proposed interlevel regularizers will be compared and the properties of the present hierarchy construction method will be studied.

Datasets. Let us run the experiments on two text datasets:

1) English Wikipedia dump (08.12.2014): $|D| = 3665223$ and $|W| = 100000$ after lemmatization and filtering words by frequency; and
2) dump of http://postnauka.ru site (scientific lectures in Russian): $|D| = 1728$ and $|W| = 38467$ after lemmatization.

Let us use only the text modality.

Regularizers comparison. Since both proposed regularizers impose additional matrix factorization task, compare the quality of this approximation varying $|W|/|D|$ proportion. Let us measure Hellinger distance between two stochastic matrices $A_{n\times m}$ and $B_{n\times m}$:

$$H(A,B) = \frac{1}{m}\sum_{j=1}^{m}\sqrt{\frac{1}{2}\sum_{i=1}^{n}\left(\sqrt{a_{ij}} - \sqrt{b_{ij}}\right)^2}.$$

Two-level hierarchy was learned with 50 and 250 topics in each level on Wikipedia subset $D' \subset D$ several times. For $|D'| = 1$, 10, 50, 100, and 200, second level was learned twice: with Φ and Θ interlevel regularizer, respectively. For each run, $H(\Phi^\ell, \Phi\Psi)$ and $H(\Theta^\ell, \widetilde{\Psi}\Theta)$, $\ell = 1$, were measured. Coefficients of interlevel regularizers are set so that Θ^ℓ is approximated at the same rate with both regularizers, there is no hierarchy sparsing. The results are given in Fig. 3.

The graphic shows that with described coefficients, setup strategy Φ^ℓ is also approximated at the same rate for any $|W|/|D|$ proportion. In other words, both regularizers approximate both matrices Φ^ℓ and Θ^ℓ. Moreover, Φ-regularizer approximates both matrices a bit better than its counterpart. Also, remember Φ-regularizer allows more efficient realization. Hence, the authors recommend to use it instead of Θ-regularizer. Let us run the following experiments with Φ interlevel regularizer.

Children number study. One can trust children topics number estimated by the proposed method if these estimates are robust. In this experiment, one can see how the number of children topics and its deviation depend on hierarchy sparsing regularizer coefficient τ_5.

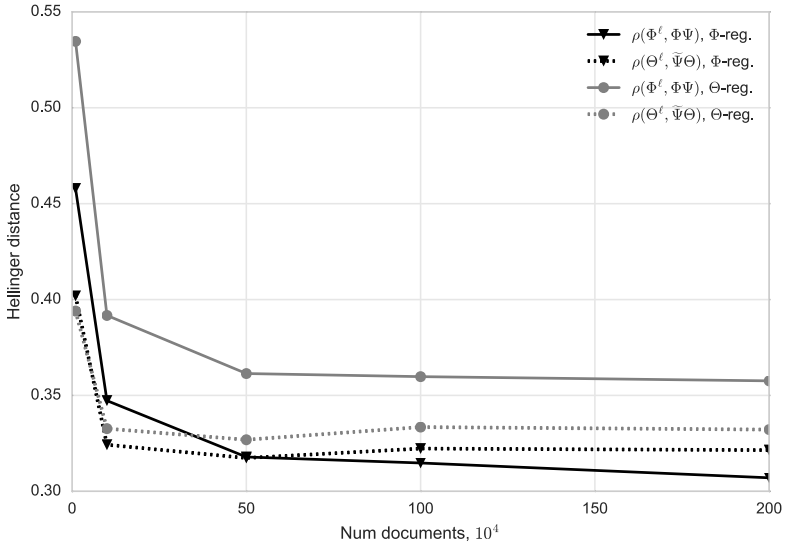


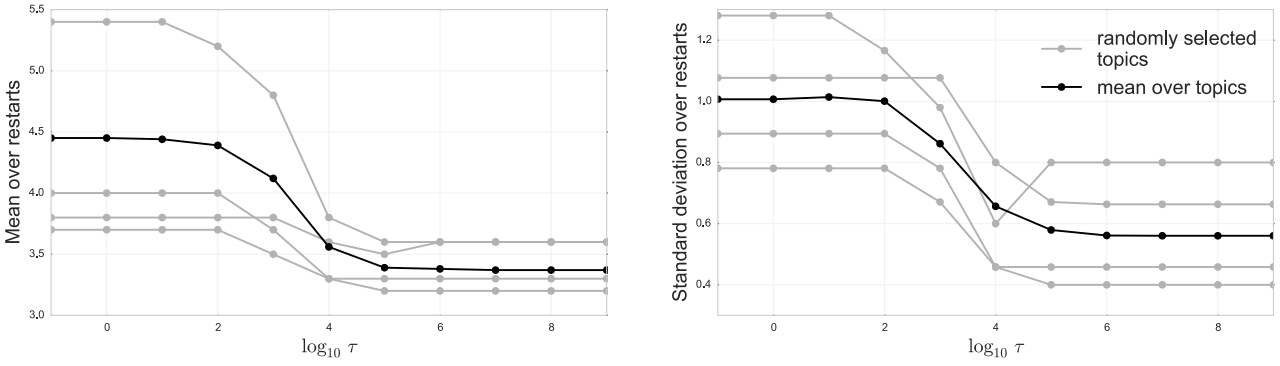
Figure 3 Interlevel regularizers comparison

Postnauka hierarchy was learned with two levels T_1 and T_2, $|T_1| = 10$ and $|T_2| = 30$. The first level was modeled once and fixed. Then, for each $\tau = \tau_5 = 0.1, 1, \ldots, 10^9$, second level with 10 restarts was learned from different random initializations. The mean m_t^τ and the standard deviation v_t^τ of children number were counted for each topic $t \in T_1$ and they were averaged over children topics:

$$m^\tau = \frac{1}{|T^1|} \sum_{t \in T^1} m_t^\tau; \quad v^\tau = \frac{1}{|T^1|} \sum_{t \in T^1} v_t^\tau.$$

We set a threshold on ψ_{ta} as maximum so that the hierarchy is still a connected graph.

Figure 4 shows that the number of children topics and its variance falls with certain hierarchy sparsing regularizer coefficient τ. For some topics, there is global minimum of m_t^τ and v_t^τ in $\tau_* = 10^4$ or 10^5. For $\tau \ll \tau_*$, regularizer affects Ψ weakly; for $\tau \gg \tau_*$, it zeroizes some ψ_{ta}

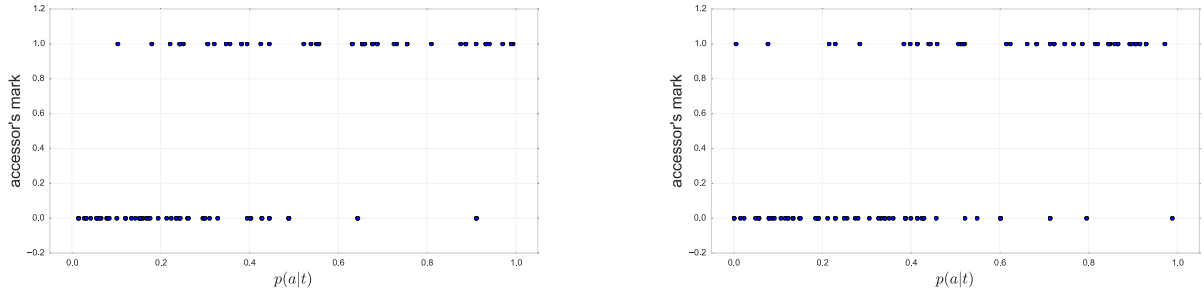


(a) (b)

Figure 4 Children number study. Dependences m_t^τ vs. τ (a) and v_t^τ vs. τ (b) for 4 randomly selected topics are in gray, black line displays m^τ vs. τ and v^τ vs. τ

immediately after initialization and hides real parent–child relations. The variance of children number is small (0.55–1), it proves that this estimate is robust. Mean parents number that equals $(|T_1|/|T_2|)m^\tau = 1.16$ for $\tau = 10^4$ is also small showing that the topic graph is close to a tree.

Parent–child relations study. In this experiment, see how well probabilities $p(a|t)$ reflect the existence of parent–child relations. We constructed three-level Postnauka hierarchy with 10, 30, and 90 topics in each level and chose 100 random pairs a–t. We asked an expert to mark each pair as interpretable (relation exists) or uninterpretable. Figure 5 shows the scatter of expert mark vs. $p(a|t)$.



(a) (b)

Figure 5 The difference in $p(a|t)$ between pairs a–t with 0 and 1 expert marks: (*a*) no sparsing and (*b*) with hierarchy sparsing

The bias between 0 and 1-marked pairs is greater for the sparse hierarchy. Although there is no explicit gap, one can impose a threshold on $p(a|t)$ so that the majority of 0- and 1-marked pairs are determined right.

Few randomly selected braches of a topic hierarchy learned from Wikipedia are shown in Fig 6.

7 Discussion

In this paper, a method of constructing topic hierarchy via regularization of the flat topic model is proposed. An experiment showed that both described regularizers do the same work; so, the more efficient one has been chosen. An idea of this regularizer is based on the assumption that a parent topic is a mixture of children topics. Some other works [16] make this assumption as well.

The authors suggest to learn hierarchy top down, level by level, not the whole hierarchy at once. Thus, the quality of topics is controled on the higher levels before splitting them into subtopics and the hierarchy is preserved from having uninterpretable branches. While other top-down approaches are recursive and split each topic node into subtopics, the suggested algorithm determines parent–child relations during child level learning and allows topics to have more than one parents. At the same time, it determines children number of each parent topic. An experiment shows that these estimates are robust. To the best of our knowledge, it is the first top-down approach with multiple inheritance support.

An open question is how to specify the number of topics on each level. To do this, one can apply the clustering technique proposed in [19]. Another way is to use a topic selection regularizer [21] that chooses as possible linearly independent topics from certainly excess topics

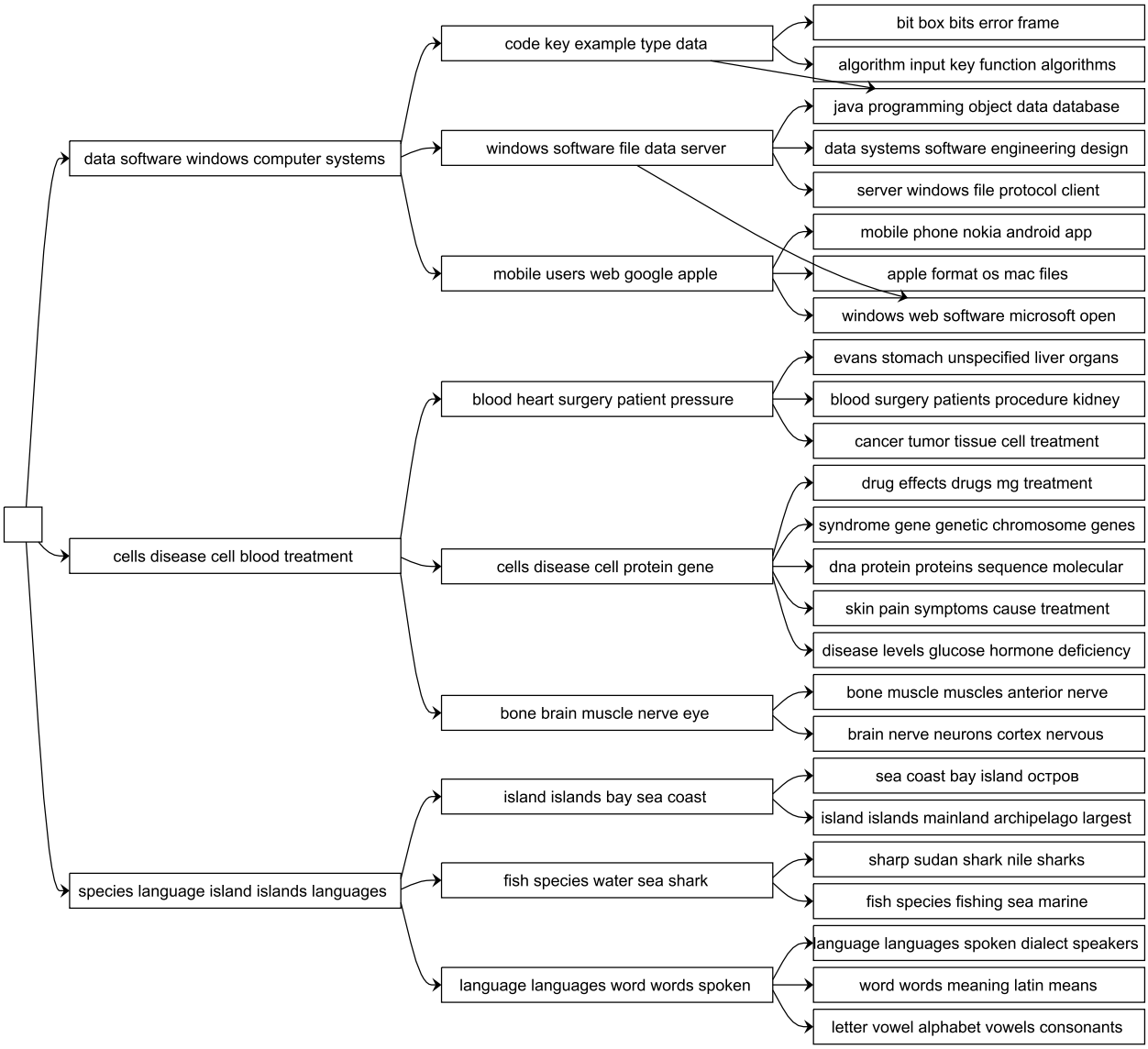


Figure 6 A part of Wikipedia hierarchy

set. The regularizer coefficients and modalities weights are usually tuned to maximize particular criteria or visual hierarchy interpretability.

References

[1] Zavitsanos, E., G. Paliouras, and G. A. Vouros. 2011. Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes. *J. Mach. Learn. Res.* 12:2749–2775.

[2] Blei, D. M., and J. D. Lafferty. 2006. Dynamic topic models. *23rd Conference (International) on Machine Learning Proceedings*. New York, NY: ACM. 113–120.

[3] Yan, X., J. Guo, Y. Lan, and X. Cheng. 2013. A biterm topic model for short texts. *22nd Conference (International) on World Wide Web Proceedings*. New York, NY: ACM. 1445–1456.

[4] Mimno, D., H. M. Wallach, J. Naradowsky, D. A. Smith, and A. McCallum. 2009. Polylingual topic models. *Conference on Empirical Methods in Natural Language Processing Proceedings*.

Stroudsburg, PA: Association for Computational Linguistics. 2:880–889. Available at: http://dl.acm.org/citation.cfm?id=1699571.1699627 (accessed December 22, 2016).

[5] Virtanen, S., Y. Jia, A. Klami, and T. Darrell. 2012. Factorized multi-modal topic model. *28th Conference on Uncertainty in Artificial Intelligence Proceedings*. Eds. N. de Freitas and K. Murphy. Corvallis, OR: AUAI Press. 843–851.

[6] Wang, D., M. Thint, and A. Al-Rubaie. 2012. Semi-supervised latent Dirichlet allocation and its application for document classification. *IEEE/WIC/ACM Conferences (International) on Web Intelligence and Intelligent Agent Technology*. 3:306–310.

[7] Blei, D. M., and J. D. Lafferty. 2006. Correlated topic models. *23rd Conference (International) on Machine Learning Proceedings*. MIT Press. 113–120.

[8] Than, K., and T. B. Ho. 2012. Fully sparse topic models. *European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings*. Berlin–Heidelberg: Springer-Verlag. I:490–505.

[9] Vorontsov, K. V. 2014. Additive regularization for topic models of text collections. *Dokl. Math.* 89(3):301–304.

[10] Vorontsov, K., O. Frei, M. Apishev, P. Romov, M. Suvorova, A. Yanina. 2015. Non-Bayesian additive regularization for multimodal topic modeling of large collections. *Workshop on Topic Models: Post-Processing and Applications Proceedings*. New York, NY: ACM. 29–37.

[11] Hofmann, T. 1999. Probabilistic latent semantic indexing. *22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings*. New York, NY: ACM. 50–57.

[12] Blei, D. M., A. Y. Ng, and M. I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3:993–1022.

[13] Blei, D. M., and J. D. Lafferty. 2009. Topic models. *Text mining: Classification, clustering, and applications*. Eds. A. N. Srivastava and M. Sahami. Chapman & Hall/CRC data mining and knowledge ser. CRC Press. 71–94.

[14] Blei, D. M., T. Griffiths, M. I. Jordan, and J. B. Tenenbaum. 2003. Hierarchical topic models and the nested Chinese restaurant process. *Advances in neural information processing systems*. Eds. S. Thrun, L. K. Saul, and P. B. Schölkopf. NIPS. 8 p.

[15] Mimno, D., W. Li, and A. McCallum. 2007. Mixtures of hierarchical topics with Pachinko allocation. *24th Conference (International) on Machine Learning Proceedings*. ACM. 633–640.

[16] Wang, C., X. Liu, Y. Song, and J. Han. 2014. Scalable and robust construction of topical hierarchies. arXiv:1403.3460.

[17] Pujara, J., and P. Skomoroch. 2012. Large-scale hierarchical topic models. *NIPS Workshop on Big Learning*.

[18] Wang, C., M. Danilevsky, N. Desai, *et al.* 2013. A phrase mining framework for recursive construction of a topical hierarchy. *19th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings*. New York, NY: ACM. 437–445.

[19] Srivastava, N. 2010. Learning size and structure of document ontologies using generative topic models. Available at: http://www.cs.toronto.edu/∼nitish/iitk_page/updated_cs397.pdf (accessed December 22, 2016).

[20] BigARTM. Available at: http://bigartm.org (accessed December 22, 2016).

[21] Vorontsov, K., and A. Potapenko. 2015. Additive regularization of topic models. *Machine Learn.* 101(1):303–323. doi: http://dx.doi.org/10.1007/s10994-014-5476-6.

Received September 3, 2016

Аддитивная регуляризация мультимодальных иерархических тематических моделей*

Н. А. Чиркова[1,2]***, К. В. Воронцов***[3]
nadiinchi@gmail.com; vokov@forecsys.ru
[1]ЗАО Антиплагиат, Россия, г. Москва, ул. Нагатинская, 33
[2]МГУ им. М. В. Ломоносова, Россия, г. Москва, Ленинские горы, 1
[3]ФИЦ «Информатика и управление» РАН, Россия, г. Москва, ул. Вавилова, 44/2

Вероятностные тематические модели выявляют семантику текстовых коллекций, описывая каждый документ дискретным распределением вероятностей на множестве тем. Иерархические модели рекурсивно делят темы на подтемы, что упрощает информационный поиск и навигацию по большим мультидисциплинарным коллекциям. В большинстве работ по иерархическому тематическому моделированию применяется байесовский вывод, что затрудняет введение тематических иерархий в тематические модели других видов. Не-байесовская аддитивная регуляризация тематических моделей, наоборот, позволяет комбинировать любые тематические модели, если их специфические особенности формализуемы в виде критериев-регуляризаторов. Однако до сих пор иерархические модели не имели такой формализации. Предлагаются регуляризаторы тематических иерархий, адаптируемые для широкого класса задач, в частности для тематизации мультимодальных и мультиязычных данных научных электронных библиотек и социальных сетей. Рассматриваются иерархии, в которых каждая подтема может иметь несколько родительских, что особенно актуально для междисциплинарных коллекций научных статей. Предлагаемый подход позволяет контролировать разреженность отношения тема–подтема и автоматически определять число подтем каждой темы. При построении модели задается только число тем на каждом уровне иерархии. Аддитивная регуляризация не усложняет процесс обучения тематической модели, что делает данный подход масштабируемым на большие текстовые коллекции.

Ключевые слова: *тематическое моделирование; APTM; тематические иерархии; регуляризация*

DOI: 10.21469/22233792.2.2.05

Литература

[1] *Zavitsanos E., Paliouras G., Vouros G. A.* Non-parametric estimation of topic hierarchies from texts with hierarchical Dirichlet processes // J. Mach. Learn. Res., 2011. Vol. 12. P. 2749–2775.

[2] *Blei D. M., Lafferty J. D.* Dynamic topic models // 23rd Conference (International) on Machine Learning Proceedings. — New York, NY, USA: ACM, 2006. P. 113–120.

[3] *Yan X., Guo J., Lan Y., Cheng X.* A biterm topic model for short texts // 22nd Conference (International) on World Wide Web Proceedings. — New York, NY, USA: ACM, 2013. P. 1445–1456.

[4] *Mimno D., Wallach H. M., Naradowsky J., Smith D. A., McCallum A.* Polylingual topic models // Conference on Empirical Methods in Natural Language Processing Proceedings. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. Vol. 2. P. 880–889. URL: http://dl.acm.org/citation.cfm?id=1699571.1699627.

*Работа выполнена при финансовой поддержке РФФИ, проекты №№ 16-37-00498, 14-07-00847 и 14-07-00908.

[5] *Virtanen S., Jia Y., Klami A., Darrell T.* Factorized multi-modal topic model // 28th Conference on Uncertainty in Artificial Intelligence Proceedings / Eds. N. de Freitas, K. Murphy. — Corvallis, OR, USA: AUAI Press, 2012. P. 843–851.

[6] *Wang D., Thint M., Al-Rubaie A.* Semi-supervised latent Dirichlet allocation and its application for document classification // IEEE/WIC/ACM Conferences (International) on Web Intelligence and Intelligent Agent Technology, 2012. Vol. 3. P. 306–310.

[7] *Blei D. M., Lafferty J. D.* Correlated topic models // 23rd Conference (International) on Machine Learning Proceedings, 2006. P. 113–120.

[8] *Than K., Ho T. B.* Fully sparse topic models // European Conference on Machine Learning and Knowledge Discovery in Databases Proceedings. — Berlin–Heidelberg: Springer-Verlag, 2012. Part I. P. 490–505.

[9] *Воронцов К. В.* Аддитивная регуляризация тематических моделей коллекций текстовых документов // Докл. Акад. наук, 2014. Т. 455. №3. С. 268–271.

[10] *Vorontsov K., Frei O., Apishev M., Romov P., Suvorova M., Yanina A.* Non-Bayesian additive regularization for multimodal topic modeling of large collections // Workshop on Topic Models: Post-Processing and Applications Proceedings. — New York, NY, USA: ACM, 2015. P. 29–37.

[11] *Hofmann T.* Probabilistic latent semantic indexing // 22nd Annual ACM SIGIR Conference (International) on Research and Development in Information Retrieval Proceedings. — New York, NY, USA: ACM, 1999. P. 50–57.

[12] *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet allocation // J. Mach. Learn. Res., 2003. Vol. 3. P. 993–1022.

[13] *Blei D. M., Lafferty J. D.* Topic models // Text mining: classification, clustering, and applications / Eds. A. N. Srivastava, M. Sahami. — Chapman & Hall/CRC data mining and knowledge ser. — CRC Press, 2009. P. 71–94.

[14] *Blei D. M., Griffiths T., Jordan M. I., Tenenbaum J. B.* Hierarchical topic models and the nested chinese restaurant process // Advances in neural information processing systems / Eds. S. Thrun, L. K. Saul, P. B. Schölkopf. — NIPS, 2003. 8 p.

[15] *Mimno D., Li W., McCallum A.* Mixtures of hierarchical topics with Pachinko allocation // 24th Conference (International) on Machine Learning Proceedings. — ACM, 2007. P. 633–640.

[16] *Wang C., Liu X., Song Y., Han J.* Scalable and robust construction of topical hierarchies // CoRR, 2014. arXiv:1403.3460.

[17] *Pujara J., Skomoroch P.* Large-scale hierarchical topic models // NIPS Workshop on Big Learning, 2012.

[18] *Wang C., Danilevsky M., Desai N., et al.* A phrase mining framework for recursive construction of a topical hierarchy // 19th ACM SIGKDD Conference (International) on Knowledge Discovery and Data Mining Proceedings. — New York, NY, USA: ACM, 2013. P. 437–445.

[19] *Srivastava N.* Learning size and structure of document ontologies using generative topic models, 2010. http://www.cs.toronto.edu/~nitish/iitk_page/updated_cs397.pdf.

[20] BigARTM. `http://bigartm.org`.

[21] *Vorontsov K., Potapenko A.* Additive regularization of topic models // Machine Learning, 2015. Vol. 101. No. 1. P. 303–323. doi: `http://dx.doi.org/10.1007/s10994-014-5476-6`.

Поступила в редакцию 03.09.2016