Additively Regularized Multimodal Topic Hierarchies*

***N. A. Chirkova*[1]*, K. V. Vorontsov*[2]**
nadiinchi@gmail.com, vokov@forecsys.ru
[1]Lomonosov Moscow State University, Leninskie Gory, 1, Moscow, Russia; [2]ORGANIZATION

AAA “Machine Learning and Data Analysis”.
BBB

Background: One paragraph about the problem, existent approaches and its limitations.
Methods: One paragraph about proposed method and its novelty.
Results: One paragraph about major properties of the proposed method and experiment results if applicable.
Concluding Remarks: One paragraph about the place of the proposed method among existent approaches.

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

DOI: 10.21469/22233792

1 Introduction

Multiple inheritance:

It is common case in any field of knowledge when specific topic occurs on the edge of two or even more major topics. For example, bioinformatics combines applied mathematics and computer science to solve biology problems.

Feature: multiple inheritance & ability to make graph sparse

Feature: fast

Feature: combination with other regularizers

2 Related Work

Briefly: PLSA, LDA, ARTM. These are plate models/

hierarchical approaches

hPAM, hLDA: slow. hPAM allows multiple inheritance and Zavitsanos as well, but they don’t allow sparsing. And mainstream is tree hierarchy.

STROD: fast. CATHY too...

hvHDP: bottom up, child topic as documents

STROD: parent topic distribution is a mixture of child topic distributions

From Zavitsanod review: about evaluation. From STROD review: slow Gibbs sampling based methods.

3 Problem statement

In this paper we refer to the document collection as D consisting of documents of different subjects. Documents may contain not only words but other elements too, say tags, links, location marks etc. We refer to such types of elements as modalities. For example, scientific paper can be described at least by three modalities: text, keywords and references. M denotes 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. Thus collection can be represented as a counters matrix $\{n_{dw}\}_{D\times W}$, n_{dw} is a number of w occurencies in d.

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

Each topic in hierarchy is associated with distributions over each modality dictionary. This allows us 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.

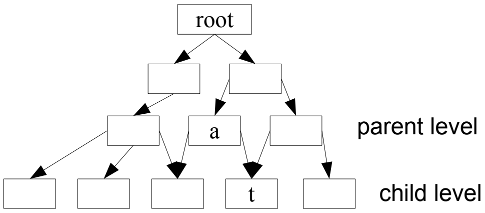

Figure 1 An example of topic hierarchy

To learn hierarchy we combine several plate topic models and tie them via regularization.

In the rest of the paper we will use operator $\operatorname*{norm}_{x\in X}[y_x] = \frac{(y_x)_+}{\sum_{x'\in X}(y_{x'})_+}$ transforming real vector to probability distribution, $(y_x)_+$ equals y_x if $y_x > 0$ and 0 otherwise.

4 hARTM framework

4.1 ARTM: plate topic models

Plate topic model describes collection D by finite topics set T. In ARTM [**?**] document distribution over each modality is modeled as a mixture of topics 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 topic model is a low-rank approximation

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

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

$$\sum_{w\in W^m} \varphi_{wt} = 1, \quad \sum_{t\in T} \theta_{td} = 1. \qquad (1)$$

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

We use regularized weighted maximum log-likelihood principle to learn Φ and Θ :

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

Weights $\varkappa_m$ are used to balance log-likehood of modalities. Regularizers R_i impose additional subject-specific criteria for model parameters. Regularizer coefficients τ_i balance optimization of regularizers and log-likehood. If 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 (Vorontsov, Potapenko, 2014). *If all regularizers are continuously differentiable on Φ and Θ, then the stationary point of problem* (2) *with constrains* (1) *satisfies the following system yielding EM-algorithm for model training:*

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

EM-algorithm is obtained by applying the fixed point iteration method to the system. An initial guess is random.

Sparsing regularizers. Frequently used sparsing regularizer [**?**] causes distributions $p(w|t)$ and $p(t|d)$ to be sparse meaning the majority of distribution's domain elements have zero probability. To achieve it 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, for uniform distribution $\alpha_t=\frac{1}{|T|}$. Similarly for Φ^m sparsing with uniform specified distribution:

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

Modified M-step formulas for parameters update:

$$\varphi_{wt}=\operatorname*{norm}_{w\in W^m}\left[n_{wt}-\frac{\tau_1}{|W^m|}\right], \quad \theta_{td}=\operatorname*{norm}_{t\in T}\left[n_{td}-\frac{\tau_2}{|T|}\right]. \qquad (4)$$

Hyperparameters of plate topic model are number of topics $|T|$, weights $\{\varkappa_m\}_{m\in M}$ and regularization coefficients $\{\tau_i\}_i$. While learning topic hierarchy, we will need to train plate topic model 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, we consider each level as a plate topic model. We 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 second level, we need not only to model topics, but also to establish parent-child topic relations. To do this, we introduce two additional matrix factorization problems and propose two new interchangeable regularizers based on them.

Assume we have already learned $\ell \geqslant 1$ hierarchy levels. Now we will learn $(\ell+1)$-th level that is child level for ℓ-th ancestor level. Not to confuse levels we denote parent level topics $a \in A$ and parameters Φ^ℓ, 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. We suppose that parent topic distribution over words and other modalities should be 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 in parent topic. If the measure of probability distributions dissimilarity is Kullback–Leibler divergence, we have the following regularizaion criteria:

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

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 topics proportionally to their size and to scale criteria up to log-likelihood scale, n_{wa} are parent topic counters from EM-algorithm. Regularizer criterias are weighted by modalities weights:

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

This regularizer is equivalent to adding $|A|$ pseudodocuments to collection represented by $\{n_{wa}\}_{W\times A}$ columns. Then Ψ forms additional columns to Θ corresponding to pseudodocuments. Note than child level couldn't 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 may be applied for regularizing Θ instead of Φ. Then for each document distribution over parent topics is a 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:

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

To train child model with regularizer we add new modality $\tilde{m}$ corresponding to parent topics and consider document counters for this modality are θ^{ℓ}_{ad}. Θ-regularizer coefficient will become modality weight and $\widetilde{\Psi}$ will correspond to $\Phi^{\tilde{m}}$.

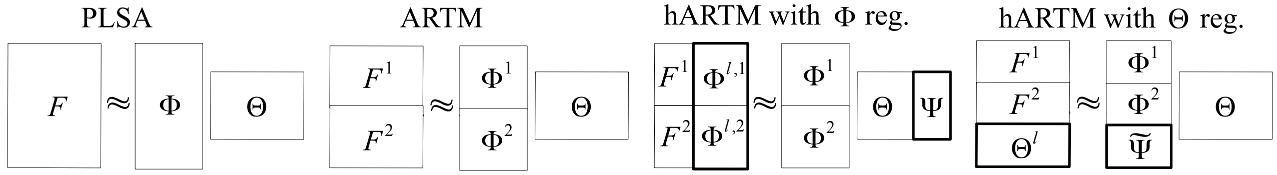

Figure 2 An illustration of child level regularization

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

Hierarchy sparsing regularizers. When we allow topics to inherit from a number of parents we assume that this number won't be large, i. e. 1–3, rarely greater parents. Such hierarchy is called *sparse* one. In other words, we want distributions $p(a|t)$ to be sparse. 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 $p(a|t)$ we can use Φ-sparsing regularizer described above to make the hierarchy sparse. We rewrite (4) replacing $\varphi \to \tilde{\psi}$, $w \to a$, $W^m \to A$ to show how $\widetilde{\Psi}$ updates on each iteration:

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

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 sparsing regularizers, we maximize KL-divergence between uniform distribution $\boldsymbol{\gamma} = \{\frac{1}{|A|}\}_{a\in A}$ and target one $\boldsymbol{p}_t = \{p(a|t)\}_{a\in A}$:

$$\sum_{t\in T} KL(\boldsymbol{\gamma}\|\boldsymbol{p}_t) \to \max_{\Psi}$$

or, equivalently,

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

To show how Ψ updates we rewrite M-step formula in (3) replacing $\theta \to \psi$ and $d \to a$ and taking derivatives of $R_5(\Psi)$ with respect to ψ_{ta}:

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

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 number of levels, 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 ot Θ interlevel rgularizer. If sparse hierarchy is desired, hierarchy sparsing regularizer should be also used. The process of training levels is stopped when 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 $(\ell+1)$-th level only ℓ-th level's topics are used for regularization, not all previous levels' topics.

When hierarchy is learned, topics on each level are represented by its distributions over words and other modalities. 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, open-source topic modeling library

6 Experiments

7 Discussion

References

[1] Goossens, M., F. Mittelbach, and A. Samarin. 1994. *The LaTeX companion*. 2nd ed. Reading, MA: Addison-Wesley. 528 p.

[2] Zagurenko, A. G., V. A. Korotovskikh, A. A. Kolesnikov, A. V. Timonov, and D. V. Kardymon. 2008. Tekhniko-ekonomicheskaya optimizatsiya dizayna gidrorazryva plasta [Technical and economic optimization of the design of hydraulic fracturing]. *Neftyanoe Khozyaystvo* [Oil Industry] 11(1):54–57. doi: `http://dx.doi.org/10.3114/S187007708007`. (In Russian)

[3] Blaga, P. A. 2007. Commutative Diagrams with XY-pic II. Frames and Matrices. *PracTEX J.* 4. Available at: `https://tug.org/pracjourn/2007-1/blaga/blaga.pdf` (accessed February 20, 2007).

[4] XYpic. Available at: `http://akagi.ms.u-tokyo.ac.jp/input9.pdf` (accessed April 09, 2015).

[5] Usmanov, T. S., A. A. Gusmanov, I. Z. Mullagalin, R. Yu. Mukhametshina, A. N. Chervyakova, and A. V. Sveshnikov. 2007. Osobennosti proektirovaniya razrabotki mestorozhdeniy s primeneniem gidrorazryva plasta [Features of the design of field development with the use of hydraulic fracturing]. *6th Symposium (International) "New Energy Saving Subsoil Technologies and the Increasing of the Oil and Gas Impact" Proceedings*. Moscow: Publisher. 267–272. (In Russian)

[6] Author, N. 2009. Paper title. *10th Conference (International) on Any Science Proceedings*. Place of publication: Publisher. 111–122.

[7] Lambert, P. 1993. *The title of the work*. Place of publication: The institution that published. Report 2.

Received May 12, 2016

Аддитивно регуляризованные многомодальные тематические иерархии[*]

Н. А. Чиркова[1]***, К. В. Воронцов***[2]

`nadiinchi@gmail.com, vokov@forecsys.ru`

[1]Московский государственный университет им. М. В. Ломоносова [2]Организация

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

DOI: 10.21469/22233792

Литература

[1] *Гуссенс М., Миттельбах Ф., Самарин А.* Путеводитель по пакету LaTeX и его расширению LaTeX 2$_\varepsilon$ / Пер. с англ. — М.: Мир, 1999. 606 с. (*Goossens M., Mittelbach F., Samarin A.* The LaTeX companion. — 2nd ed. — Reading, MA, USA: Addison-Wesley, 1994. 528 p.)

[2] *Загуренко А. Г., Коротовских В. А., Колесников А. А., Тимонов А. В., Кардымов Д. В.* Технико-экономическая оптимизация дизайна гидроразрыва пласта // Нефтяное хозяйство, 2008. Т. 11. № 1. С. 54–57. doi: `http://dx.doi.org/10.3114/S187007708007`.

[3] *Blaga P. A.* Commutative Diagrams with XY-pic II. Frames and Matrices // PracTEX J., 2007. Vol. 4. URL: `https://tug.org/pracjourn/2007-1/blaga/blaga.pdf`.

[4] XYpic. URL: `http://akagi.ms.u-tokyo.ac.jp/input9.pdf`.

[5] *Усманов Т. С., Гусманов А. А., Муллагалин И. З., Мухаметшина Р. Ю., Червякова А. Н., Свешников А. В.* Особенности проектирования разработки месторождений с применением гидроразрыва пласта // Труды 6-го Междунар. симп. «Новые ресурсосберегающие технологии недропользования и повышения нефтегазоотдачи». — М.: Издательство, 2007. С. 267–272.

[6] *Author N.* Paper title // 10th Conference (International) on Any Science Proceedings. — Place of publication: Publisher, 2009. P. 111–122.

[7] *Lambert P.* The title of the work. Place of publication: The institution that published, 1993. Report 2.

Поступила в редакцию 12.05.2016

[*]