Supervised topic classification for modeling a hierarchical conference structure

M. Kuznetsov*, V. Strijov,
M. Clausel, M.-R. Amini, E. Gaussier

Moscow Institute of Physics and Technology,
Grenoble Alpes University

ICONIP, 9-12 November 2015

Problem formulation

The goal

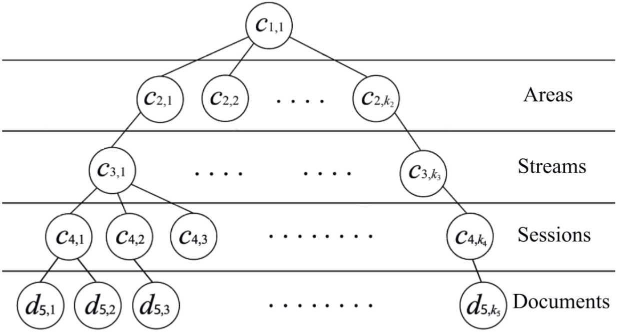
To construct a hierarchical topic model using supervised information about topics.

The approach

- To construct a model we use an ARTM (Additive Regularization of Topic Models) approach.
- The regularizer term penalizes the difference between estimated and expert-given topic models.
- The difference between models is a distance between corresponding hierarchy trees.

Hierarchical conference structure

A structure of the IFORS conference:



- At the upper level there are 26 main areas,
- each of areas contains about 5 streams,
- each stream then contains about 5 sessions,
- each session is formed by 4 abstracts,
- overall number of abstracts is 3000.

A document example

Area: Decision Analysis, Decision Support Systems

Stream: Intelligent Optimization in Machine Learning and Data Analysis

Session: Categorical Data Analysis and Preference Aggregation

Document:
We propose a new method for the ordinal-scaled object ranking problem. The method is based on the combining of partial orders corresponding to the ordinal features. Every partial order is described with a positive cone in the object space. We construct the solution of the object ranking problem as the projection to a superposition of the cones. To restrict model complexity and prevent overfitting we reduce dimension of the superposition and select most informative features. The proposed method is illustrated with the problem of the IUCN Red List monotonic categorization.

Topic modeling problem

Given:

- D is a set of documents (a collection), $d \in D$,
- W is a set of words (a vocabulary), $w \in W$,
- n_{dw} is a number of occurences of a term w in a document d,
- T is a set of latent topics.

Find:

- $\phi_{wt} = p(w|t)$, a distribution over terms for a topic,
- $\theta_{td} = p(t|d)$, a distribution over topics for a document,

Basic assumptions:

- We consider a *bag-of-words* model.
- Each observed word in a document has a latent topic.

Probabilistic topic model

Given:

- D is a set of documents (a collection), $d \in D$,
- W is a set of words (a vocabulary), $w \in W$,
- T is a set of latent topics.

A probabilistic topic model

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

Find:

parameters $\phi_{wt} = p(w|t)$ and $\theta_{td} = p(t|d)$ of a topic model

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

under constraints $\phi_{wt} \geqslant 0,\ \theta_{td} \geqslant 0, \quad \sum_{w \in W} \phi_{wt} = 1,\ \sum_{t \in T} \theta_{td} = 1.$

Basic methods

PLSA [Hofmann, 1999]

Constrained maximization of the log-likelihood:

$$L(\Phi, \Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \quad \to \quad \max(\Phi, \Theta),$$

n_{dw} is a number of appearances of a term w in d.

Additive regularization of topic models (ARTM)

Maximum log-likelihood with additive regularization criterion R:

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \lambda R(\Phi, \Theta) \quad \to \quad \max(\Phi, \Theta),$$

Supervised topic classification, flat case

Given:

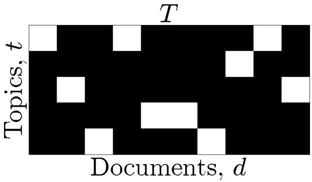
- D is a set of documents (a collection), $d \in D$,
- W is a set of words (a vocabulary), $w \in W$,
- $\mathbf{t}$ is a vector of the expert-given topics, $t_d \in T$.

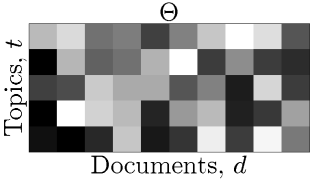
Distance between topics

The regularizer $R(\Phi, \Theta, \mathbf{t})$ computes the distance between topic profiles:

$$R(\Phi, \Theta, \mathbf{t}) = -\sum_{d \in D} r(t_d, \boldsymbol{\theta}_d) = -\sum_{d \in D} \sum_{t \in T} |z_{td} - \theta_{td}|,$$

where $z_{td} = [t_d = t]$.





Parameter optimization for ARTM

Optimization problem

Maximum log-likelihood with additive regularization criterion R:

$$\sum_{d\in D}\sum_{w\in d} n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td} + R(\Phi,\Theta,\mathbf{t}) \quad\to\quad \max(\Phi,\Theta),$$

EM approach

A EM extension for an additive part $R(\Phi,\Theta,\mathbf{t})$:

- E-step: $p_{tdw} = \mathrm{norm}_{t\in T}(\phi_{wt}\theta_{td})$,
- M-step: $\begin{cases} \phi_{wt} = \mathrm{norm}_{w\in W}\big(\sum\limits_{d\in D} n_{dw}p_{tdw} + \phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\big) \\ \theta_{td} = \mathrm{norm}_{t\in T}\big(\sum\limits_{w\in d} n_{dw}p_{tdw} + \theta_{td}\frac{\partial R}{\partial\theta_{td}}\big) \end{cases}$

M-step for supervised regularization

$$\theta_{td} = \frac{\eta_{td}}{\sum\limits_{t\in T}\eta_{td}}, \qquad \eta_{td} = \left[\sum_{w\in d} n_{dw}\frac{\phi_{wt}\theta_{td}}{\sum\limits_{t\in T}\phi_{wt}\theta_{td}} + \lambda\theta_{td}(2z_{td}-1)\right]_+ .$$

Topics hierarchy

Hierarchy levels

- Denote by $T = T = T_0 \sqcup \ldots \sqcup T_L$, where the sets $T_0, \ldots, .T_L$ denote disjoint sets of topics at different levels of hierarchy.
- Parent $p(t)$ and children $s(t)$ operators:

$$p(t) \in T_{l-1} \text{ for } t \in T_l, \quad s(t) \subset T_{l+1} \text{ for } t \in T_l.$$

Basic idea: expand Θ to hierarchy levels

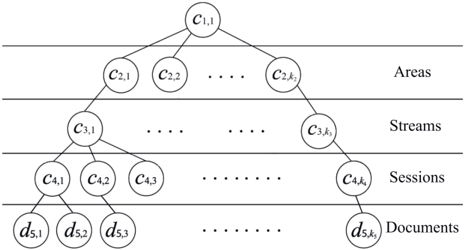
Consider the additional document-topic parameters $[\Theta, \Theta']$ corresponding to the different hierarchy levels:

$$\theta'_{td} = \begin{cases} \theta_{td}, & t \in T_l, \\ \frac{1}{\#s(t)} \sum\limits_{s \in s(t)} \theta'_{sd}, & \text{otherwise.} \end{cases}$$

Parameter optimization for a hierarchical model

Distance between hierarchical topics

$$R(\Phi, \Theta, \mathbf{t}) = -\sum_{d \in D} r(t_d, \boldsymbol{\theta}_d), \quad r(t_d, \boldsymbol{\theta}_d) = \sum_{l=0}^{L} \sum_{t \in T} |z_{td} - \theta'_{td}|,$$



M-step for supervised hierarchical regularization:

$$\eta_{td} = \left[\sum_{w \in d} n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum_{t \in T} \phi_{wt}\theta_{td}} + \lambda_1 \theta_{td} \left(2z_{td} - 1\right) + \lambda_2 \theta'_{p(t)d} \left(2z_{p(t)d} - 1\right) \right]_+$$

Conference topics visualization

Summary

- We constructed a hierarchical topic model using an additive regularization approach.
- To take into consideration supervised information we proposed a measure of distance between topic models.
- We extended the distance to a hierarchical case and modified an M-step of the EM algorithm to find the optimal model parameters.
- The proposed method was used to construct a hierarchical model of the conference.