Supervised topic classification for modeling a hierarchical conference structure

Mikhail Kuznetsov[1], Marianne Clausel[2],
Massih-Reza Amini[3], Eric Gaussier[3], and Vadim Strijov[1]

[1] Moscow Institute of Physics and Technology,
Institutskiy Lane 9, Dolgoprudny, Moscow, 141700, Russia,
mikhail.kuznecov@phystech.edu, strijov@phystech.edu
[2] Laboratoire Jean Kuntzmann, Université de Grenoble Alpes, CNRS F38041
Grenoble Cedex 9, France,
marianne.clausel@imag.fr
[3] Laboratoire d'Informatique de Grenoble, Université de Grenoble Alpes, CNRS
F38041 Grenoble Cedex 9, France,
Massih-Reza.Amini@imag.fr, eric.gaussier@imag.fr

Abstract. In this paper we investigate the problem of supervised latent modelling for extracting topic hierarchies from data. The supervised part is given in the form of expert information over document-topic correspondence. To exploit the expert information we use a regularization term that penalizes the difference between a predicted and an expert-given model. We hence add the regularization term to the log-likelihood function and use a stochastic EM based algorithm for parameter estimation. The proposed method is used to construct a topic hierarchy over the proceedings of the European Conference on Operational Research and helps to automatize the abstract submission system.

Keywords: hierarchical topic model, labeled classification, probabilistic latent semantic analysis, EM approach

1 Introduction

Probabilistic topic models are generally unsupervised generative models that describe document content in large document collections. These models assume that each document is associated with a set of hidden variables, called topics, that indicate how the words within the document are generated. Formally, a topic is a probability distribution over terms in a vocabulary. The two most popular topic models are the Probabilistic Latent Semantic Indexing (PLSI) [6] and the latent Dirichlet allocation (LDA) model [2] and their variants. The LDA model consists of two types of probability distributions: *a*) distributions of topics over documents and *b*) distributions of words over topics. After estimating the model parameters over a training corpus, the obtained distributions of words over topics can then be used to infer per-document topic distributions on unseen documents. LDA has found applications in many areas ranging from document clustering, text categorization, ad-hoc information retrieval, to signal analyzes. Several attempts have been made to extent PLSI and LDA to unsupervised hierarchical topic modelling. In [5], Dirichlet processes are used to model different levels of an hierarchy, while in [3] an extension of the PLSI model is proposed by introducing additional probabilities corresponding to different levels of the hierarchy.

In this paper, we address the problem of hierarchical topic modelling using an expert information over the document-topic correspondence in the form of a labeled document collection with a predefined hierarchical-structured topics. The problem is hence to predict a topic model for a new document collection using such past labeled information.

The application that we consider is the construction of an hierarchical topic model for the "European Conference on Operational Research" (EURO) containing over 3000 abstracts. The structure of the conference papers is shown in Figure 1. At the upper level there are 26 main areas, each of which contains about 10 streams. Each stream then contains about 10 sessions, and each session is formed by four abstracts. The main areas correspond to the broad topics of the operational research field like *Non-smooth optimization*, *timetabling*, *logistics*, etc. Every year the program committee, constituted by groups of experts, constructs by hand such an hierarchy for the submitted papers [7]. Each group is responsible for the organization of a stream or a set of streams. After the abstract submission deadline each group of experts starts to fill a stream with unassigned abstracts and to form sessions within a stream. The practical goal of our research is to construct an efficient structure from the supervised expert information of the previous years using the topic modeling methodology. For that, we consider the additive regularization of topic models (ARTM) [9]. In general, this method finds topic-document and word-topic probabilities by optimizing a log-likelihood quality measure with an additional regularization term. Here we propose, a regularizer term that penalizes the difference between the predicted and the expert-given topic models.

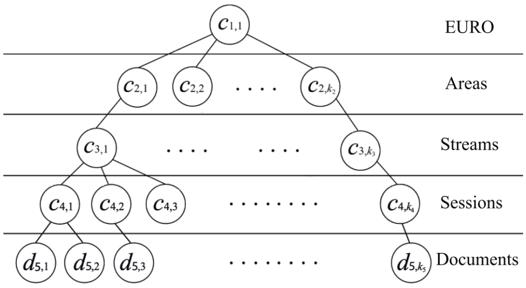


Fig. 1. Hierarchical structure of the conference

Compared to [8], where the prior probabilities are modified with respect to the projections of document-topic vectors on the set of identified topics, here we propose a unified formalism to measure the distance between the hierarchy trees by introducing a set of hyperparameters that describes the hierarchy and summarizes penalizations on different hierarchy levels. The optimization of the regularized likelihood is then carried out using a stochastic version of the

Expectation-Maximization algorithm [6, 10]. The algorithm has a modified M-step that takes into account the regularization term and samples a current topic from the conditional distribution on a given word-document pair.

The structure of the paper is as follows. In Section 2, we present our framework and a Bayesian interpretation of the proposed supervised topic model. Section 3 presents its hierarchy extension and empirical results are shown in Section 4. Finally, in Section 5 we discuss the outcomes of this study and give some pointers to further research.

2 Supervised classification, flat case

Let D denote a collection of documents, $d_i \in D$, and W denote a vocabulary, a set of terms describing the documents. Let T denote a set of topics such that each document d_i may refer to a topic $t(d_i) \in T$. Let $t_1, ..., t_n$ denote an initial expert topic classification of the documents $d_1, ..., d_n$. The given sample consists of the document-topic pairs, $\{d_i, t_i\}_{i=1}^n$.

To construct a probabilistic model we use conditional independence assumption. The collection D is generated from the distributions $\theta_{td} = p(t|d)$ and $\phi_{wt} = p(w|t)$ in the following way:

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

To estimate the probabilities $(\theta_{td})_{t \in T, d \in D}$ and $(\phi_{wt})_{w \in W, t \in T}$, we consider the PLSI approach [6], where the optimization problem consists in maximizing the log-likelihood $L(\boldsymbol{\Phi}, \boldsymbol{\Theta})$ under non-negativity and normalization conditions:

$$\begin{gathered} \boldsymbol{\Phi}^*, \boldsymbol{\Theta}^* = \operatorname*{argmax}_{\boldsymbol{\Phi}, \boldsymbol{\Theta}} L(\boldsymbol{\Phi}, \boldsymbol{\Theta}) = \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td}, \\ \text{u.c.} \quad \phi_{wt} \geqslant 0, \quad \theta_{td} \geqslant 0, \text{ and } \sum_{w \in W} \phi_{wt} = 1, \quad \sum_{t \in T} \theta_{td} = 1. \end{gathered} \tag{1}$$

The PLSI model (1) does not take into account the initial topic classification $t_1, ..., t_n$, we tackle the problem by introducing the expert-given topic labels using a regularization term $R(\mathbf{t}, \hat{\mathbf{t}})$ that measures the similarity between the predicted and the expert-given topic vectors, $\mathbf{t}$ and $\hat{\mathbf{t}}$:

$$\begin{gathered} \boldsymbol{\Phi}^*, \boldsymbol{\Theta}^* = \operatorname*{argmax}_{\boldsymbol{\Phi}, \boldsymbol{\Theta}} L(\boldsymbol{\Phi}, \boldsymbol{\Theta}) + \lambda R(\mathbf{t}, \hat{\mathbf{t}}), \\ \text{u.c.} \quad \phi_{wt} \geqslant 0, \quad \theta_{td} \geqslant 0, \text{ and } \sum_{w \in W} \phi_{wt} = 1, \quad \sum_{t \in T} \theta_{td} = 1. \end{gathered} \tag{2}$$

Where is λ the regularization parameter.

Bayesian interpretation of the regularized PLSI. As stated in [4], a penalized approach can be interpreted within the Bayesian framework. According to such an interpretation, the penalized likelihood function corresponds to the *a posteriori* density whereas the penalty is the density of the prior. The solution of the maximization of the penalized likelihood of the model is then a maximum *a posteriori* estimate of the parameters of interest. In our setting, adding a regularization to the PLSI model means that we are setting the following prior for the latent variables (θ, ϕ)

$$\pi(\theta, \phi) = C\, exp\, (\lambda R(\phi, \theta)) \tag{3}$$

where $C > 0$ is a normalizing constant.

Our corpus can then be assumed to be generated as follows :

- Step 1: Generate the whole set of the topic and of the matrix word–topic $(\theta, \phi) \sim \pi$ where π is the distribution defined in 3.
- Step 2: for each document d and each word of the document
  - Draw the n^{th} topic $t_n^w \sim mult(\theta_{td})$
  - Draw the n^{th} word w_n with probability ϕ_{w_n, t_n^w}.

Labeled classification. Let $\mathbf{Z} = \|z_{td}\|$ be a document-topic correspondence matrix of size $D \times T$ such that

$$z_{td} = \mathbb{1}_{\hat{t}_d = t}.$$

Where $\mathbb{1}_\pi$ is the indicator function, equal to 1 if the predicate π holds and 0 otherwise. We define similarity $R(\mathbf{t}, \hat{\mathbf{t}})$ as a matrix norm of difference between matrices $\boldsymbol{\Theta}$ and $\mathbf{Z}$:

$$R(\mathbf{t}, \hat{\mathbf{t}}) = -\|\boldsymbol{\Theta} - \mathbf{Z}\|_1.$$

This form of regularization leads us to the following optimization problem:

$$\begin{gathered}\boldsymbol{\Phi}^*, \boldsymbol{\Theta}^* = \underset{\boldsymbol{\Phi}, \boldsymbol{\Theta}}{\operatorname{argmax}} \sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} + \lambda \left(\sum_{d \in D} \sum_{t \in T} \theta_{td} (2 z_{td} - 1) \right), \\ \phi_{wt} \geqslant 0, \quad \theta_{td} \geqslant 0, \qquad \sum_{w \in W} \phi_{wt} = 1, \quad \sum_{t \in T} \theta_{td} = 1. \end{gathered} \tag{4}$$

Parameter optimization: EM approach. To solve the optimization problem (4) we use the Expectation-Maximization algorithm. To derive the explicit expectation and maximization formulas we use theorem 1 from [9] that gives properties of the local optimum of the general expression of (Eq. 2). Following this result if $R(t, \hat{t})$ is continuously differentiable then at the local maximum of R we have:

$$\phi_{wt} \propto \left(n_{wt} + \phi_{wt} \frac{\lambda \partial R}{\partial \phi_{wt}} \right)_+, \quad \theta_{td} \propto \left(n_{td} + \theta_{td} \frac{\lambda \partial R}{\partial \theta_{td}} \right)_+. \tag{5}$$

Note that in our problem the function R depends only of θ_{td} variables, therefore we will use only a second equation for θ.

For the problem (4) we hence obtain the following formula for the M-step:

$$\theta_{td} = \frac{\eta_{td}}{\sum\limits_{t \in T} \eta_{td}}, \qquad \eta_{td} = \left[\sum_{w \in d} n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum\limits_{t \in T} \phi_{wt}\theta_{td}} + \lambda \theta_{td} \left(2 z_{td} - 1\right) \right]_+ . \tag{6}$$

Stochastic EM. To speed up the proposed EM algorithm we rather use its stochastic version that is similar to the Gibbs sampling method for LDA [2]. The approach consists in sampling a topic t from the estimated distribution $p(t|d,w)$, where the distribution of a topic t given w, d is given by a formula

$$p(t|d,w) \propto \left(\frac{\hat{n}_{wt}}{\hat{n}_t} \frac{\hat{n}_{dt} + \lambda \hat{n}_{dt}(2z_{td} - 1)}{n_d + \lambda \sum\limits_{t \in T} \hat{n}_{dt}(2z_{td} - 1)} \right)_+ ,$$

where

$$\hat{n}_{dt} = \sum_{w \in d} n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum\limits_{t \in T} \phi_{wt}\theta_{td}}, \quad \hat{n}_{wt} = \sum_{d \in D} n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum\limits_{t \in T} \phi_{wt}\theta_{td}}, \quad \hat{n}_t = \sum_{w \in d} \hat{n}_{wt}.$$

3 Topics hierarchy

We extend the model by taking into account the expert-given hierarchy defined on the set of topics. To model the hierarchical structure we introduce the following notations. Let us denote by $T = T = T_0 \sqcup ... \sqcup T_L$ a set of topics, or a set of vertices of a hierarchical tree, where the sets $T_0, ..., .T_L$ denote disjoint sets of topics at different levels of hierarchy. For further reading we consider a two-level hierarchical structure. However, the proposed method can be used for any number of levels.

For further convenience we introduce parent $p(t)$ and children $s(t)$ operators defined as follows:

$$p(t) \in T_{l-1} \quad \text{for} \quad t \in T_l, \quad l = 1, ..., L,$$

$$s(t) \subset T_{l+1} \quad \text{for} \quad t \in T_l, \quad l = 0, ..., L-1.$$

To define the loss function $R(t, \hat{t})$ between topics we propose to measure a summary loss over the hierarchy levels:

$$R(t, \hat{t}) = \sum_{l=0}^{L-1} r(p^l(t), p^l(\hat{t})).$$

Here the vertex t belongs to the lowest level of hierarchy, $t \in T_L$, and $p^l(t)$ is the l-th predecessor of the vertex t.

To measure the value of single loss $r(t_i, \hat{t_i})$ on a document d_i we expand (Eq. 4) to the different hierarchy levels:

$$r(t, \hat{t}) = |\mathbb{1}_{\hat{t}=t} - \theta'_{td}|,$$

where θ'_{td} is defined for an arbitrary hierarchy level as follows:

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

According to the introduced hierarchy addition we obtain the following modification of the M-step formula (6):

$$\eta_{td} = \left[\sum_{w \in d} n_{dw} \frac{\phi_{wt}\theta_{td}}{\sum\limits_{t \in T} \phi_{wt}\theta_{td}} + \lambda_1 \theta_{td} \left(2z_{td} - 1\right) + \lambda_2 \theta'_{p(t)d} \left(2z_{p(t)d} - 1\right) \right]_+ . \quad (7)$$

4 Empirical results

We use the proposed method to construct a topic model for the European Conference on Operational Research. We use the collection of abstracts for the 2012 year. Each abstract contains less than 600 symbols, the collection contains 1342 abstracts, and vocabulary contains 1675 words after preprocessing. The preprocessing stage includes removing stop words and lemmatization. Together with the collection we used an initial expert-given conference structure as described in introduction [1].

To show the hierarchical results we first need to choose the hyperparameters λ_1 and λ_2 (Eq. 7). To do this we perform the following steps.

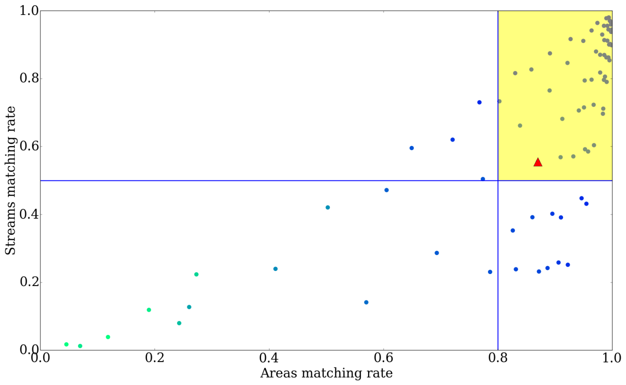
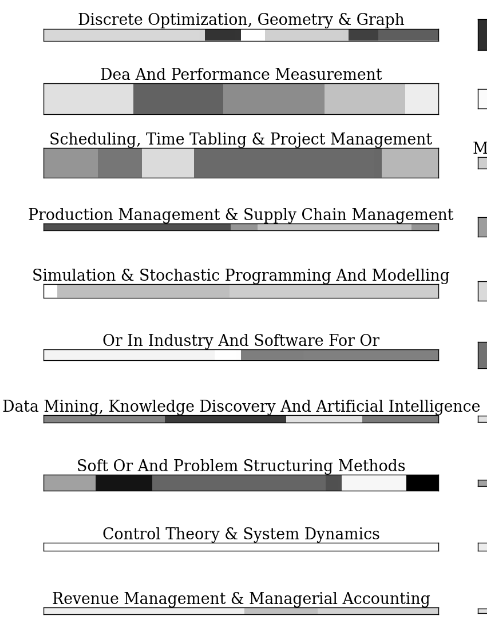


Fig. 2. Hierarchical model matching rates for different regularization values

1. Estimate model parameters for different sets of parameters λ_1, λ_2. We took about 100 different parameter sets from the range $\lambda_1, \lambda_2 \in [0, 1]$.

2. For each set of parameters we obtain three values measuring the quality of a hierarchical model: 1) the normalized number of documents matched with the expert model within the areas $na \in [0, 1]$, 2) the same for the streams, $ns \in [0, 1]$, 3) the value of perplexity.
3. We chose those regularization values λ_1, λ_2 that minimize the perplexity for the values $na > 0.8$, $ns > 0.5$.

Figure 2 illustrates the mentioned steps. x- and y-axis correspond to the values na and ns, respectively. Each point corresponds to the different set of parameters λ_1, λ_2. The color of each point indicates the values of perplexity: the darker the color, the higher the perplexity. The optimal point (of minimum perplexity with $na > 0.8$ and $ns > 0.5$) indicated by the triangle. The regularization values for this point are $\lambda_1 = 0.15$, $\lambda_2 = 0.2$.



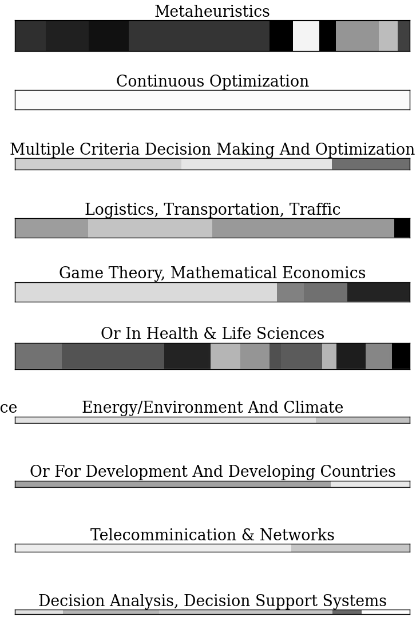


Fig. 3. Conference hierarchy matching

Figure 3 shows matching of hierarchical model for the EURO conference. Each block corresponds to the main area such that the height of each block indicates total number of documents belonging to the corresponding area due to the expert-given model. Each block consists of the subblocks corresponding to the streams; the length of the subblock indicates the size of the stream. The color of each subblock indicates rate of documents ns matched with the expert-given model: the more white is subblock, the better is the matching (ns is closer to

1). According to our method we can specify stable and non-stable areas. We see that good matched areas (mostly white-coloured) are ”Continiuos optimization”, ”Control theory” and ”Revenue management”, whereas bad-matched are, e.g., ”Metaheuristics” and ”OR in health”.

5 Conclusion

We proposed a supervised hierarchical topic model, where the expert knowledge is encompassed into a regularization term measuring the distance between the predicted and the expert-given topic models. The optimization of the regularized likelihood is then carried out using a stochastic version of the Expectation-Maximization algorithm where the modified M-step takes into account the regularization term and samples a current topic from the conditional distribution on a given word-document pair. Our experiments on the EURO proceedings showed that the proposed topic model is able to find expert-given topics, with high perplexity.

References

1. EURO conference abstracts and data. http://sourceforge.net/p/mlalgorithms/code/HEAD/tree/EURO_data/. [Online; accessed 14-May-2015].
2. David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3:993–1022, 2003.
3. Eric Gaussier, Cyril Goutte, Kris Popat, and Francine Chen. A hierarchical model for clustering and categorising documents. In *Advances in Information Retrieval*, pages 229–247. Springer Berlin Heidelberg, 2002.
4. I.J. Good and R.A. Gaskins. Nonparametric roughness penalties for probability densities. *Biometrika*, 58(2):255–277, 1971.
5. Thomas L. Griffiths, Michael I. Jordan, Joshua B. Tenenbaum, and David M. Blei. Hierarchical topic models and the nested chinese restaurant process. In S. Thrun, L.K. Saul, and B. Schölkopf, editors, *Advances in Neural Information Processing Systems 16*, pages 17–24. 2004.
6. Thomas Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57. ACM, 1999.
7. A. A. Kuzmin and V. V Strijov. Validation of the thematic models for document collections. *Informacionnie technologii*, 4:16–20, 2013.
8. Daniel Ramage, David Hall, Ramesh Nallapati, and Christopher D Manning. Labeled lda: A supervised topic model for credit attribution in multi-labeled corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*, pages 248–256. Association for Computational Linguistics, 2009.
9. K.V. Vorontsov and A.A. Potapenko. Additive regularization of topic models. *Machine Learning Journal*, Special Issue ”Data Analysis and Intelligent Optimization”, 2014.
10. Hanna M Wallach. Topic modeling: beyond bag-of-words. In *Proceedings of the 23rd international conference on Machine learning*, pages 977–984. ACM, 2006.