Hierarchical Thematic Classification of Major Conference Proceedings

Arsentii Kuzmin, Alexander Aduenko, and Vadim Strijov
Moscow Institute of Physics and Technology — Laboratory of Machine Intelligence

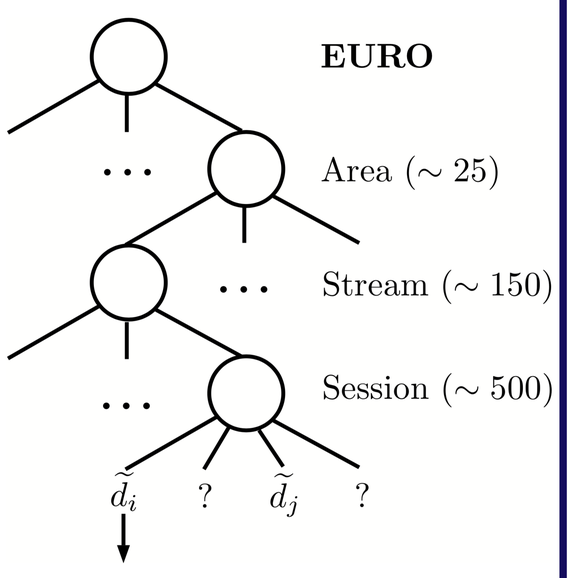
strijov@phystech.edu

Hierarchical text classification

The thematic model of a text collection is a map, which determines a set of topics from some given hierarchical structure of topics for each document from the collection. The text collections are abstracts of conference proceedings, or text messages from social networks, or research papers. The thematic model assists in searching through collections efficiently. However the model construction is often labour-intensive. Some collections already have their structure and subset of documents, which have been partially classified by experts. To simplify the classification procedure the authors propose an algorithm to rank topics for a given document in a collection [1, **?**].

EURO conference thematic modelling

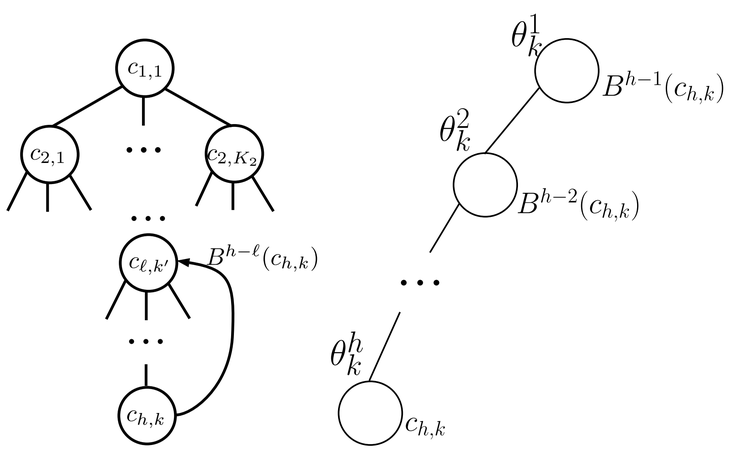
Every year the program committee the European Conference on Operational Research (EURO) builds a schedule (thematic model) for this conference using a set of submitted abstracts. The structure of this model consists of 26 major areas. Each Area consists of $10-15$ streams. Each stream consists of $5-10$ Sessions. Each session consists of four talks in a row. The participants submit short abstracts to the program committee to proceed. There are two types of participants: invited participants and new, contributed, participants. The invited participants already have a determined session, so the collection of abstracts is partly labelled. For each contributed participant, the program committee should select the most relevant session according to the content of the abstract and according to the conference structure. The program committee invites up to 200 experts from different research areas to construct the thematic model.



Hierarchical similarity function

We propose a weighted hierarchical similarity function to calculate the topic relevance. The function calculates the similarity of a document and a tree branch. The weights in this function determine word importance. We use the entropy of words to estimate the weights [2, 3].

The weighted similarity of the document $\mathbf{x}$ and the cluster $c_{\ell,k}$ $s(\mathbf{x}, c_{\ell,k}) = \mathbf{x}^\mathsf{T}\mathbf{\Lambda}\boldsymbol{\mu}(c_{\ell,k})$, where $\mathbf{\Lambda} = \operatorname{diag}(\lambda_1, \cdots, \lambda_{|W|})$ is the word importance matrix, and λ_i is the importance of the word w_i.



Cluster hierarchy Cluster branch $c_{h,k}$, weights

The hierarchical similarity function of the document $\mathbf{x}$ and the cluster $c_{h,k}$ is $s_\mathsf{h}(\mathbf{x}_n,\ c_{h,\,k}) =$

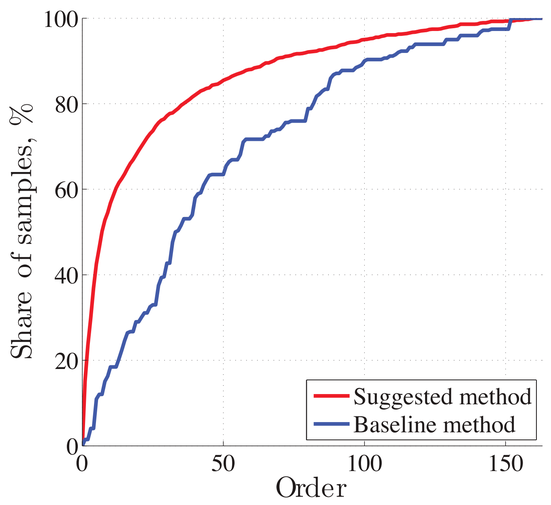
$$= \sum_{\ell=1}^{h} \theta_k^\ell s\big(\mathbf{x}_n,\ B^{h-\ell}(c_{h,\,k})\big) = \mathbf{x}_n^\mathsf{T}\mathbf{\Lambda}\mathbf{M}_k\boldsymbol{\theta}_k,$$

where θ_k^ℓ weight of the cluster of level ℓ on the branch k, and $\mathbf{M}_k$ – matrix of centers of parent clusters $c_{h,k}$:

$$\boldsymbol{\mu}_{\ell,k} = \boldsymbol{\mu}\big(B^{h-\ell}(c_{h,k})\big), \quad \mathbf{M}_k = [\boldsymbol{\mu}_{1,k}, \ldots, \boldsymbol{\mu}_{h,k}].$$

Hierarchical ranking quality

We introduce the ranking operator $R: \mathbb{R}^{|W|} \to S^{K_h}$. It maps the document $\mathbf{x} \in \mathbb{R}^{|W|}$ to the permutation $q(\mathbf{x}) \in S^{K_h}$ of the clusters $\{h\}$.



$$\operatorname{Hist}(k) = \#\{n:\ \operatorname{pos}\big(R(\mathbf{x}_n),\, c(\mathbf{x}_n)\big) \le k\},$$

$$\operatorname{AUCH}(R) = \frac{1}{k_h|D|}\sum_{k=1}^{k_h}\operatorname{Hist}(k).$$

Joint model of document clusters

The probability of some document $\mathbf{x}_n$ belongs some cluster $c_{h,k}$ is

$$p(\mathbf{x}_n \in c_{h,k}|\boldsymbol{\theta}, \boldsymbol{\alpha}) = \operatorname{softmax}(\mathbf{s}_\mathsf{h}(\mathbf{x}_n|\boldsymbol{\theta}_k, \boldsymbol{\alpha}))_k.$$

The probabilistic model $p(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})$ of the document classes $\mathbf{Z}$, $z_{nk} = 1$ for $\mathbf{x}_n$ belong $c_{h,k}$ is

$$p(\mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha}) = L(\mathbf{Z}|\boldsymbol{\theta}, \boldsymbol{\alpha})\mathcal{N}(\boldsymbol{\alpha}|\mathbf{0}, a^{-1}\mathbf{I})\times$$

$$\times \prod_k \mathcal{N}(\boldsymbol{\theta}_k|\mathbf{m}_k, \mathbf{V}_k^{-1})\mathcal{NW}(\mathbf{m}_k, \mathbf{V}_k|\mathbf{m}_0, b, \mathbf{W}, \nu).$$

The probability estimate of some *unclassified* document $\tilde{\mathbf{x}}_t$ belongs some cluster $c_{h,k}$:
the MAP estimates

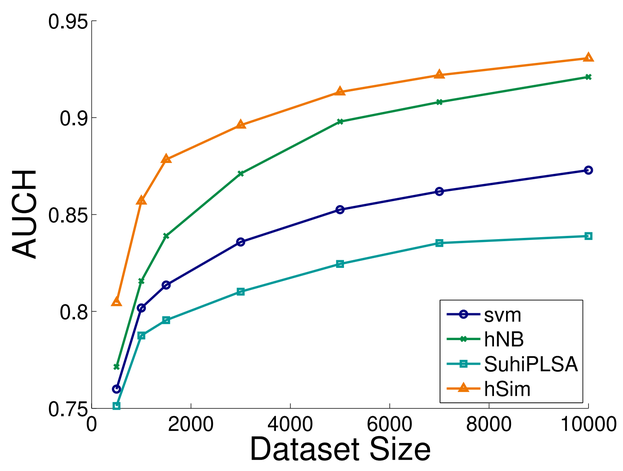
$$p(\tilde{z}_{tk}|\tilde{\mathbf{x}}_t) = p(\tilde{z}_{tk}|\tilde{\mathbf{x}}_t, \boldsymbol{\theta}_k^{\text{MAP}}, \boldsymbol{\alpha}^{\text{MAP}}),$$

the evidence estimates

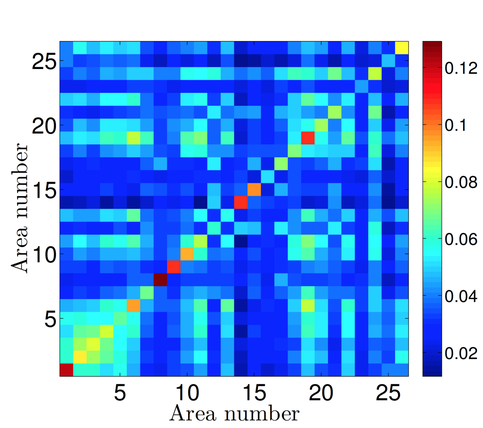
$$p(\tilde{z}_{tk}|\tilde{\mathbf{x}}_t) = \int p(\tilde{z}_{tk}|\tilde{\mathbf{x}}_t, \boldsymbol{\theta}, \boldsymbol{\alpha})p(\boldsymbol{\theta}, \boldsymbol{\alpha}|\mathbf{Z})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\alpha}.$$

Ranking results of hierarchical similarity function

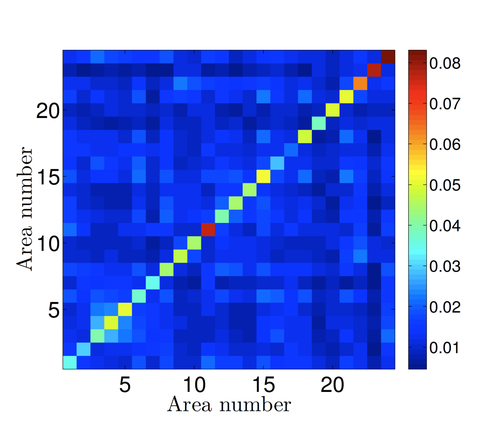
For the ranking experiment, we considered the area and the stream levels of the conferece hierarchy. We constructed the ranking operator R using the proposed weighted hierarchical similarity function hSim and used the EM algorithm to optimize its parameters on the training subset. The results of this function were compared with results several algorithms of hierarchical ranking: 1) the hierarchical naive Bayes hNB, 2) the probabilistic regularized model SuhiPLSA and 3) the hierarchical multiclass SVM.



Test AUCH quality dependence on the training sample size



Same importance for all words, $\mathbf{\Lambda} = \mathbf{I}$



Optimized importance of words, $\mathbf{\Lambda} = \mathbf{\Lambda}^*$.

References

[1] M. Alexandrov, A. Gelbukh, and P. Rosso. An approach to clustering abstracts. *Natural Language Processing and Information Systems*, 20:275–285, 2005.

[2] M. P. Kuznetsov, A. A. Tokmakova, and V. V. Strijov. Analytic and stochastic methods of structure parameter estimation. *Informatica*, 37(4):607–624, 2016.

[3] M. P. Kuznetsov, M. Clausel, M.-R. Amini, E. Gaussier, and V. V. Strijov. Supervised topic classification for modeling a hierarchical conference structure. *International conference on neural information processing*, 9489:90—-97, 2015.

[4] T. Joachims. A probabilistic analysis of the rocchio algorithm with tfidf for text categorization. *Machine Learning: ECML*, pages 143–151, 1997.