Thematic classification for EURO/IFORS conference using expert model

Arsentiy Kuzmin, Alexander Aduenko, and Vadim Strijov

Moscow Institute of Physics and Technology
Department of Control and Applied Mathematics

EURO 2016, Poznan
05.06.2016

Construct a decision support system to assist the program committee and stream organizers make the forthcoming conference program

The goal:

- to construct a thematic model of the conference

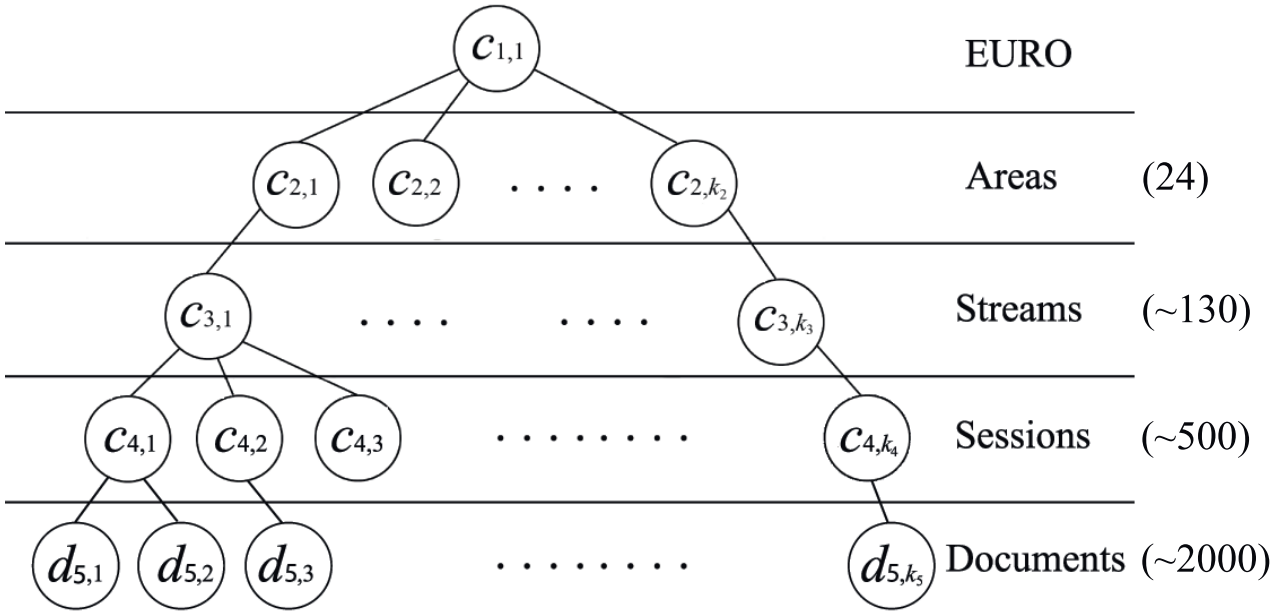
There given:

- historical expert thematic models of the previous conferences
- submitted abstracts for the forthcoming conference

The main idea:

- to calculate the similarity of a new abstract and each Stream,
- to show the most similar Streams to the Experts

EURO/IFORS conference hierarchical model



1. A group of experts is responsible for an Area,
2. participants submit their Abstracts to the collection,
3. the experts distribute the Abstracts over the Streams,
4. the Abstracts are organized into the Sessions.

Document-vector representation

Let $W = \{w_1, \ldots, w_n\}$ be the terms dictionary of the collection.

There exists two different approaches to represent a word:

Vocabulary vector

$C(w_j) \in \mathbb{R}^{|W|}, \quad |W| \approx 10^4$

$C(w_j) = e(j) = [\ldots 0\ 1\ 0 \ldots]^{\mathsf{T}}$

$\|C(w_j) - C(w_i)\| = 2 \cdot \mathrm{I}(i \neq j)$

Distance between words "logistic" and "transport" equals distance between "logistic" and "finance".

Document-vector representation:
$\mathbf{x}_d = \sum_{w_j \in d} C(w_j)$

word2vec, gloVec

$C(w_j) \in \mathbb{R}^K, \quad |W| \approx 10^2$

$\frac{1}{T} \sum_{t=1}^{T} \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j} | w_t)$

$\|C(w_j) - C(w_i)\| \in \mathbb{R}$

Words with similar contexts are closer, so
"logistic" $\approx$ "transport", but
"logistic" $\neq$ "finance".

Document-vector representation: doc2vec, convolutional NN, recursive NN.

The clustering quality function

Suppose F_0 is a mean intra-cluster similarity: $F_0 = \frac{1}{k_\ell} \sum_{i=1}^{k_\ell} S(c_{\ell,i},\ c_{\ell,i})$,

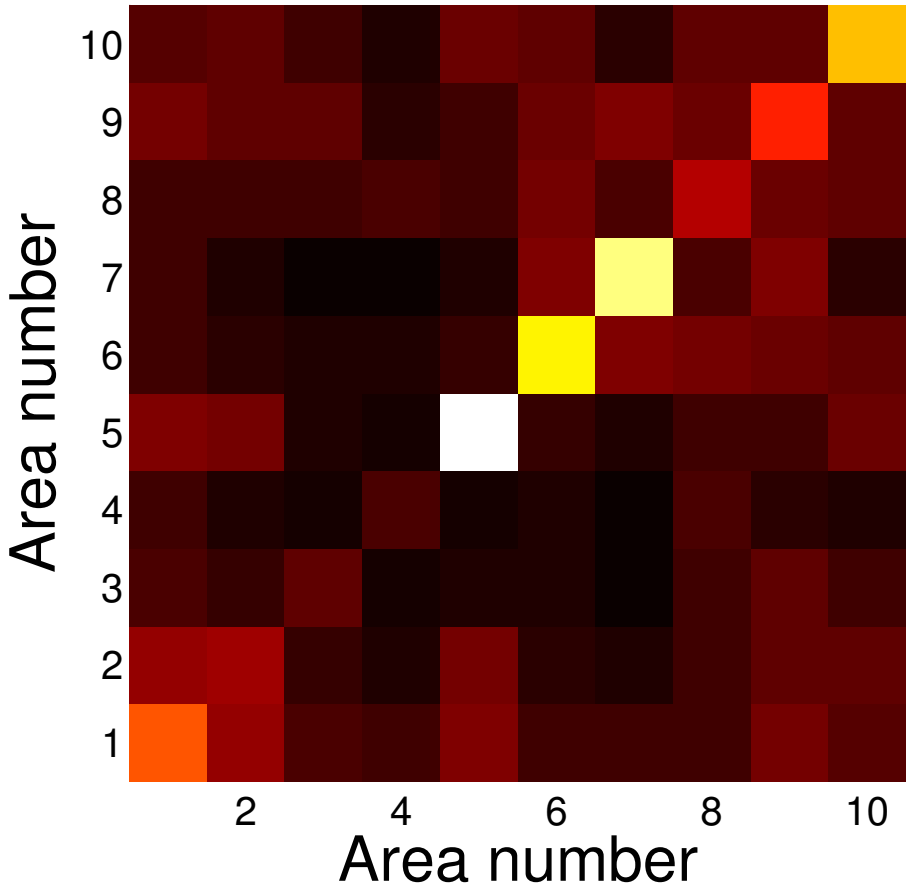
and F_1 is a mean inter-cluster similarity: $F_1 = \frac{2}{k_\ell(k_\ell - 1)} \sum_{i<j} S(c_{\ell,i},\ c_{\ell,j})$

Clustering quality criterion

$$F = \frac{F_0}{F_1} \to \max$$

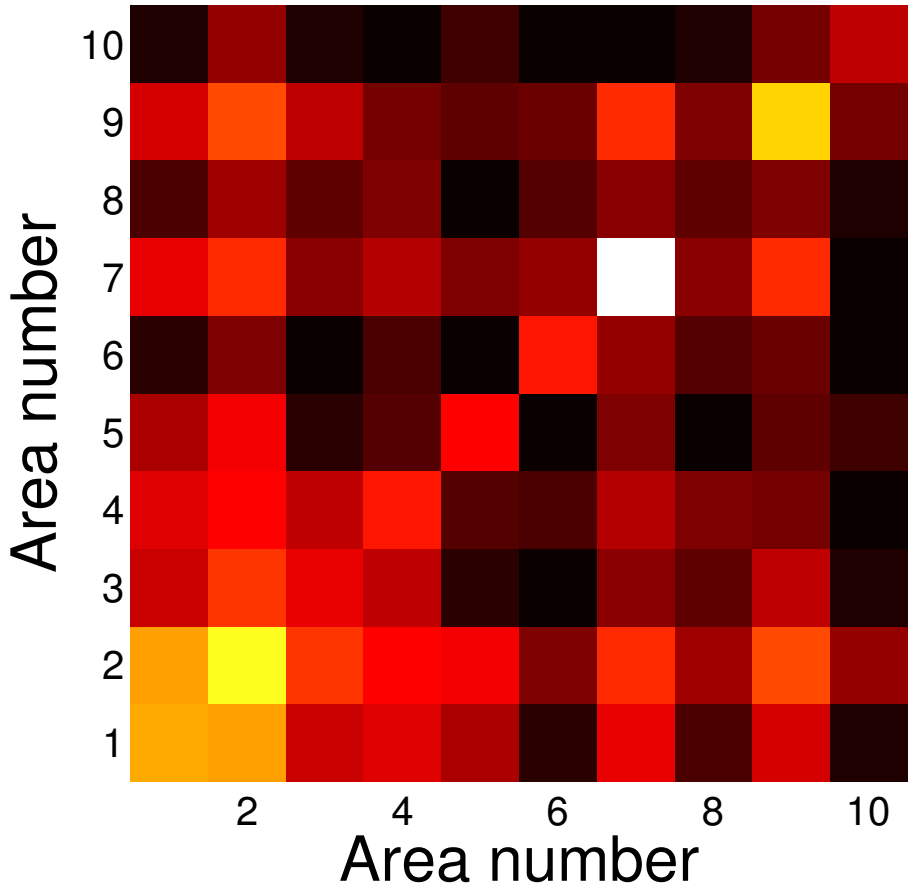
The expert model is the origin for the algorithmic thematic model.

Document representation comparison



Document vector - sum of it's word vectors $e(w_i)$.

F0/F1 = 1.98



Document vector - calculated with doc2vec approach.

$F0/F1 = 1.31$

Similarity function

Define the similarity function $s(\cdot, \cdot)$ between documents $\mathbf{x}_i$ and $\mathbf{x}_j$ as:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^\top \mathbf{\Lambda} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^\top \mathbf{\Lambda} \mathbf{x}_i} \sqrt{\mathbf{x}_j^\top \mathbf{\Lambda} \mathbf{x}_j}} = \mathbf{x}_i^\top \mathbf{\Lambda} \mathbf{x}_j, \text{ normalization: } \mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^\top \mathbf{\Lambda} \mathbf{x}_s}},$$

where $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_{1,1}, \ \ldots, \ \lambda_{n,n}\}$ is a term-importance matrix.

Define the similarity function $s(\cdot, \cdot)$ between the document $\mathbf{x}_i$ and the cluster $c_{\ell,j}$ on the ℓ hierarchy level as:

$$s(\mathbf{x}_i, \ c_{\ell,j}) = \mathbf{x}^\top \Lambda \bar{\mathbf{x}}_{\ell,\,i},$$

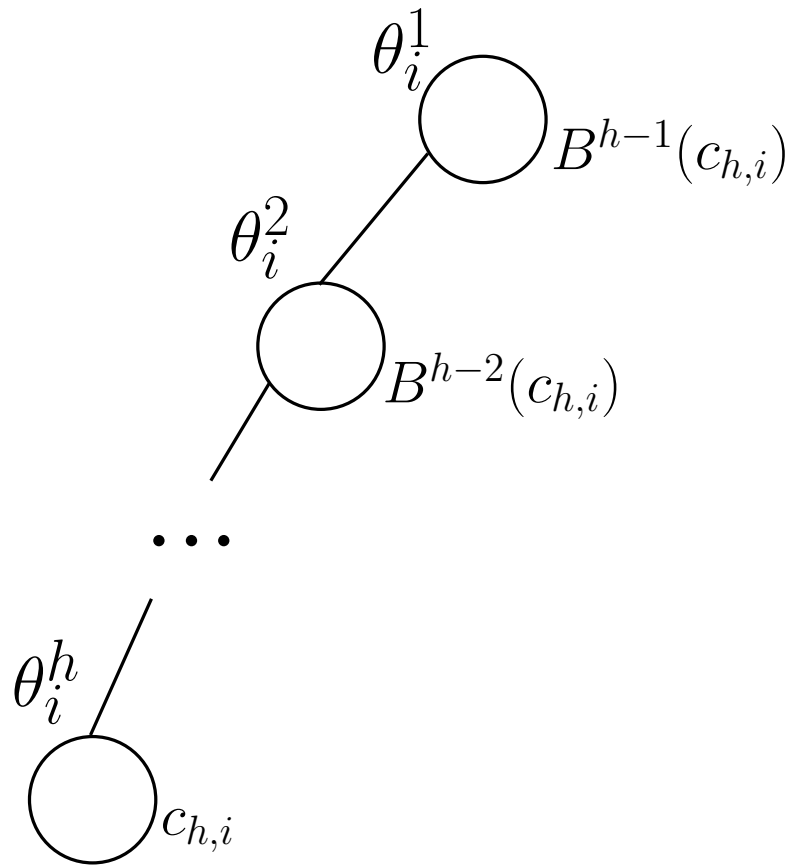
where $\bar{\mathbf{x}}_{\ell,\,i}$ is the mean vector of the cluster $c_{\ell,\,i}$.

Hierarchical similarity function

Similarity between document and cluster of the h level

$$s(\mathbf{x}, c_{h,i}) = \sum_{j=0}^{h-1} \theta_i^{h-j} s\left(\mathbf{x}, B^j(c_{h,i})\right),$$

where θ_i^{h-j} is the weights parameter of the level $h-j$ for the cluster $c_{h,i}$, and B^j is the operator of the precedence that associate cluster $c_{h,i}$ with its predecessor on the level j.



Optimize weight parameters θ

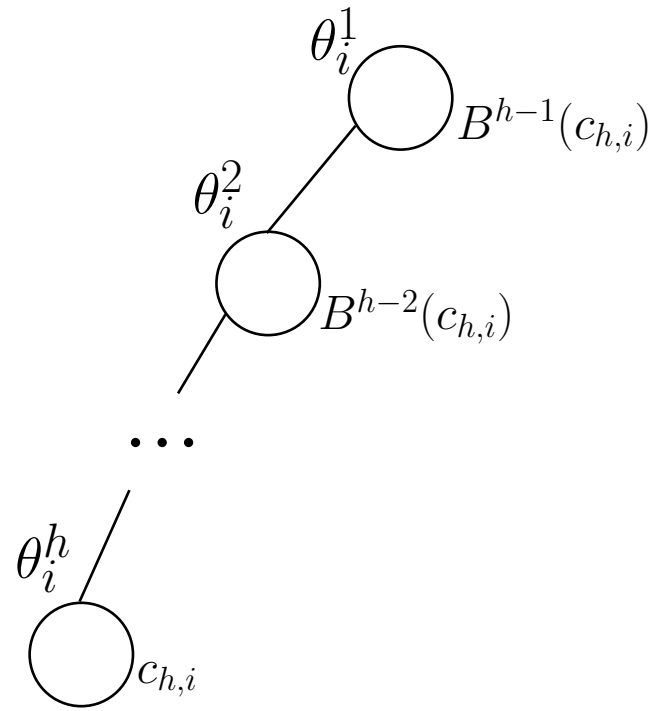
Find optimal parameters $\boldsymbol{\theta}_i$ given expert hierarchy:

$$\boldsymbol{\theta}_i^* = \arg\max_{\boldsymbol{\theta}_i} \sum_{\mathbf{x} \in c_{h,i}} \sum_{j=0}^{h-1} \theta_i^{h-j} s\big(\mathbf{x},\ B^j(c_{h,i})\big) + \mu \sum_{j=1}^{h} \left(\theta_i^j - \frac{1}{h}\right)^2,$$

$$\sum_{j=1}^{h} \theta_i^j = 1, \qquad \theta_i^j \geq 0, \quad j \in \{1 \ldots h\}.$$

Iterative procedure:

- find optimal $\boldsymbol{\Lambda}$ with fixed $\boldsymbol{\theta}_i$ to maximize similarity $s\big(\mathbf{x},\ B^j(c_{h,i})\big)$ (stay tuned);
- find optimal $\boldsymbol{\theta}_i$ using fixed $\boldsymbol{\Lambda}$ and similarities;
- stop when $\Delta\boldsymbol{\theta}_i$ and $\Delta\boldsymbol{\Lambda}$ for consecutive iterations are small.



Return for a new document all clusters sorted by similarity in descending order instead of the most similar one

Definition

Let $q \in S^{k_h}$ be the permutation of the level h clusters. The clusters in this permutation are sorted by the similarity to an object $\mathbf{x}$ in the descending order, k_h is the clusters quantity.

Example: $q = \{3, 1, \ldots, 6\}$.

Definition

Let $R : \mathbb{R}^n \to S^{k_h}$ be the relevance operator. It maps the document $\mathbf{x} \in \mathbb{R}^n$ to the permutation q.

Definition

Let $\mathrm{pos}(q, j) : S^{k_h} \times \{1, 2, \ldots, k_h\} \to \{1, 2, \ldots, k_h\}$ be the position function. It returns the position of the given number in the permutation.

Example: $\mathrm{pos}(q, 1) = 2$.

Quality criteria $Q(R)$ and $\mathrm{AUCH}(R)$

$Q(R)$ quality criterion

Denote $Q(R)$ by the average position of the expert cluster $z_{j,h}$ in the permutation $R(\mathbf{x}_j)$:

$$Q(R) = \frac{1}{|D|} \sum_{j=1}^{|D|} \mathrm{pos}\big(R(\mathbf{x}_j),\, z_{j,h}\big).$$

$\mathrm{AUCH}(R)$ quality criteria

$\mathrm{AUCH}(R) \in [0,\, 1]$ is the area under the top curve for a histogram $\#\{\mathrm{pos}(R(\mathbf{x}_j),\, z_{j,h}) \le i\}$, where $i \in [1,\, k_h]$.

$$\mathrm{AUCH}(R) = \sum_{i=1}^{k_L} \frac{\#\{\mathrm{pos}\big(R(\mathbf{x}_j),\, z_{j,h}\big) \le i\}}{k_h |D|}.$$

Optimize $\boldsymbol{\Lambda}$ using the expert thematic model

$$\boldsymbol{\Lambda}^* = \arg\min_{\boldsymbol{\Lambda}} Q(R_{SIM}, \mathbf{X}).$$

Terms significance

Denote by $\mathbf{p}_{\ell,j}$ the vector of j-th components of cluster vectors $\bar{\mathbf{x}}_{\ell,i}$ of the level ℓ.

$$\mathbf{p}_{\ell,j} = [\bar{\mathbf{x}}_{\ell,1,j}, \ldots, \bar{\mathbf{x}}_{\ell,k_\ell,j}]^\mathsf{T}, \quad \mathbf{p}_{\ell,j} \mapsto \frac{\mathbf{p}_{\ell,j}}{\sum_{i=1}^{k_\ell} p_{\ell,i,j}}$$

The word entropy

Define the entropy $I_\ell(w_j)$ of the word w_j for hierarchy level ℓ as:

$$I_\ell(w_j) = \sum_{i=1}^{k_\ell} -p_{\ell,i,j} \log(p_{\ell,i,j}).$$

Term w_j significance according to its entropy

$$\lambda_j = 1 + \alpha_\ell \log\bigl(1 + I_\ell(w_j)\bigr).$$

Optimization using the collection with the expert model

$$\alpha_\ell^* = \arg\min_{\alpha_\ell} Q(R).$$

The documents collection: EURO abstracts

The purpose of the experiment

Construct a thematic model of the conference EURO 2013

The collection D^1:

- EURO 2010, $|D| = 1663$, 26 Areas, 113 Streams.
- EURO 2012, $|D| = 1342$, 26 Areas, 141 Streams.

The collection D^2:

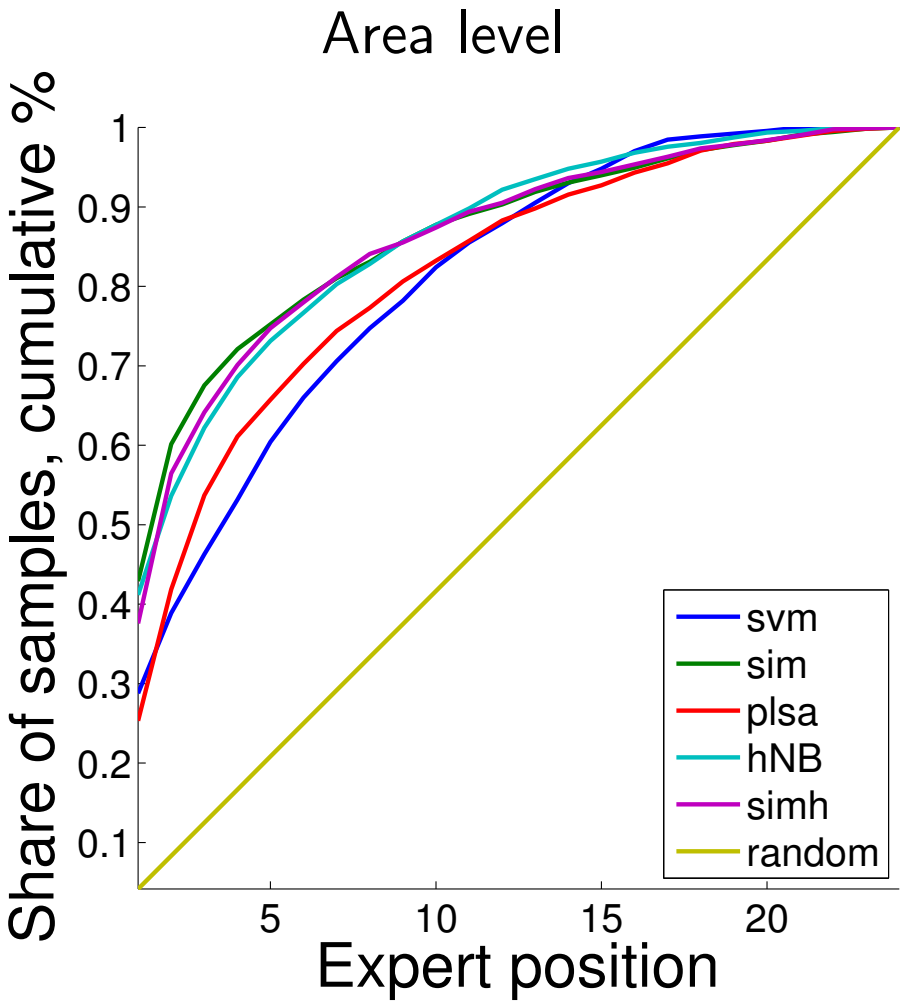
- EURO 2013, $|D| = 2313$, 24 Areas, 137 Streams.

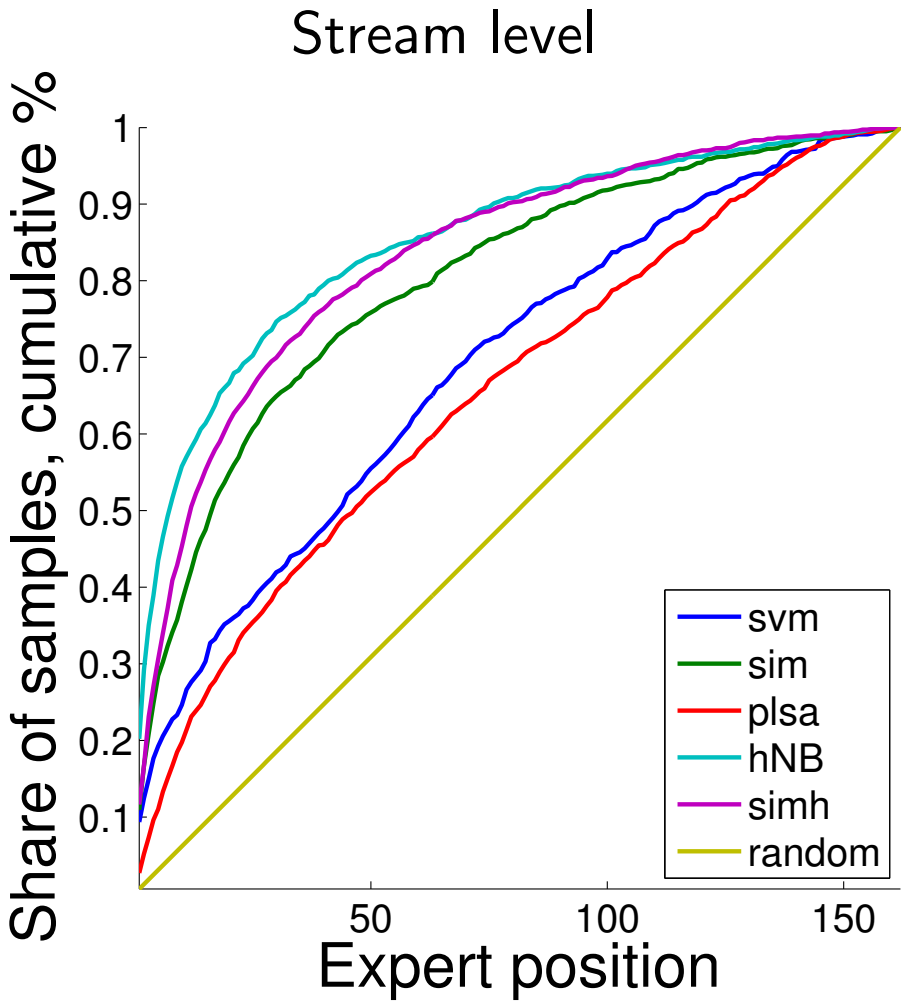
We matched the Areas and the Streams from all collections:
The unified structure has 24 Areas, 178 Streams.
15 out of 178 streams are present only in the EURO 2013.

Size of the expert dictionary:

- $|W| = 1675$ terms.

Quality comparison, EURO abstracts





	svm	sim	plsa	hNB	simh
AUCH(R), Area level	0.80	**0.85**	0.81	**0.84**	**0.85**
AUCH(R), Stream level	0.69	0.78	0.65	**0.84**	**0.83**

Realization

Conference program validation for EURO/INFORMS abstract collection

Paste title and abstract here

Title:

Hierarchical thematic model visualizing algorithm

Abstract:

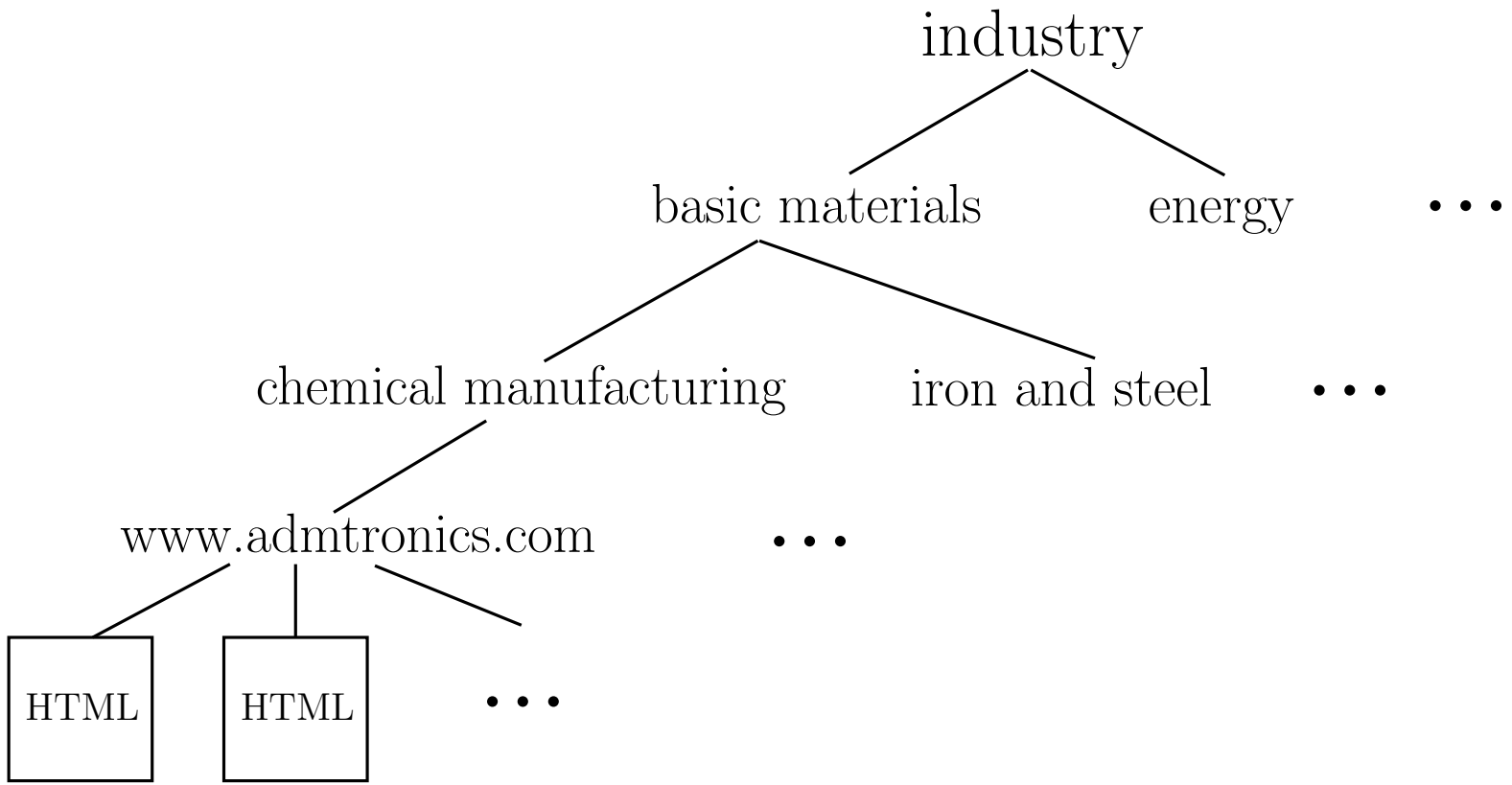
The talk is devoted to the problem of the thematic hierarchical model construction. One must to construct a hierarchcal model of a scientific conference abstracts using machine learning clustering approach, to check the adequacy of the expert models and to visualize hierarchical differences between the algorithmic and expert models. An algorithms of hierarchical thematic model constructing is developed. It uses the notion of terminology similarity to construct the model. The obtained model is visualized as the plane graph.

Clear | Search

Search results (page 1 of 18)

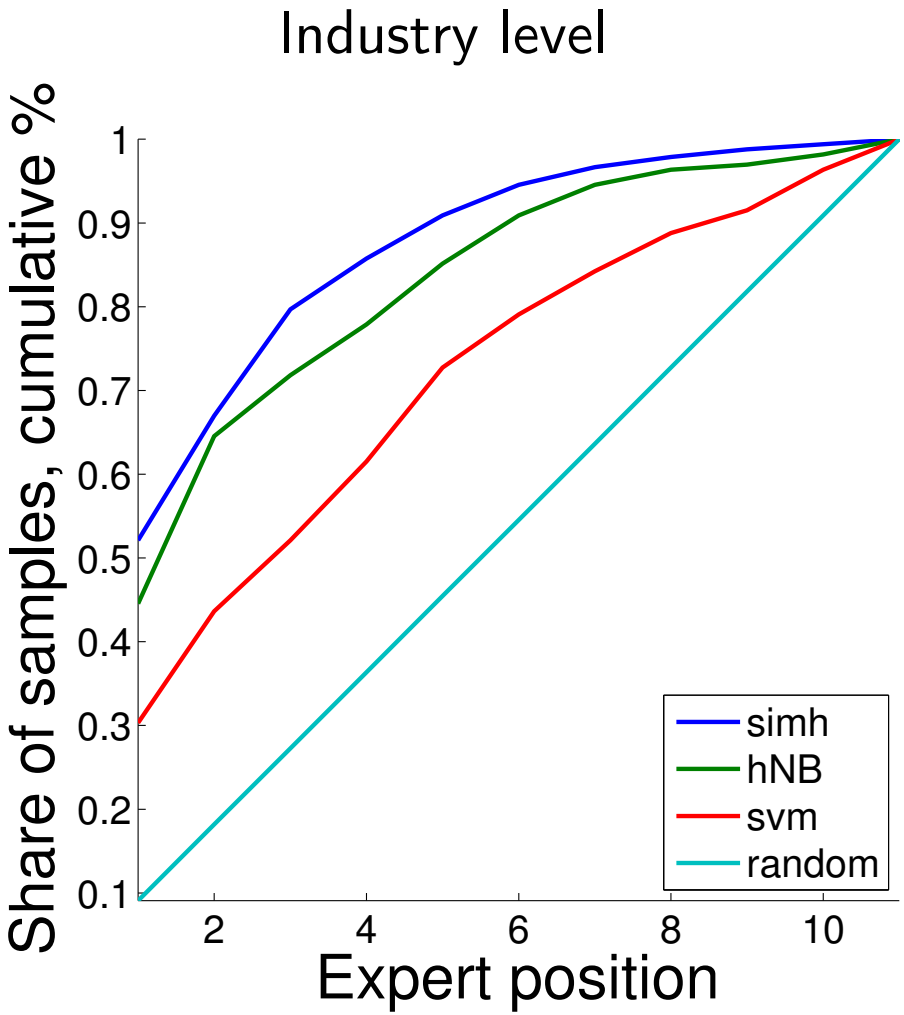
Area: Emerging Applications of OR **Stream:** Models of Embodied Cognition	Select
Area: OR in Health, Life Sciences & Sports **Stream:** Medical Decision Making	Select
Area: Discrete Optimization, Geometry & Graphs **Stream:** Graphs and Networks	Select
Area: Data Science, Business Analytics, Data Mining **Stream:** Machine Learning and its Applications	Select
Area: Discrete Optimization, Geometry & Graphs **Stream:** Boolean and Pseudo-Boolean Optimization	Select
Area: Discrete Optimization, Geometry & Graphs **Stream:** Geometric Clustering	Select
Area: Multiple Criteria Decision Making and Optimization **Stream:** Preference Learning	Select
Area: Multiple Criteria Decision Making and Optimization **Stream:** Innovative Software Tools for MCDA	Select

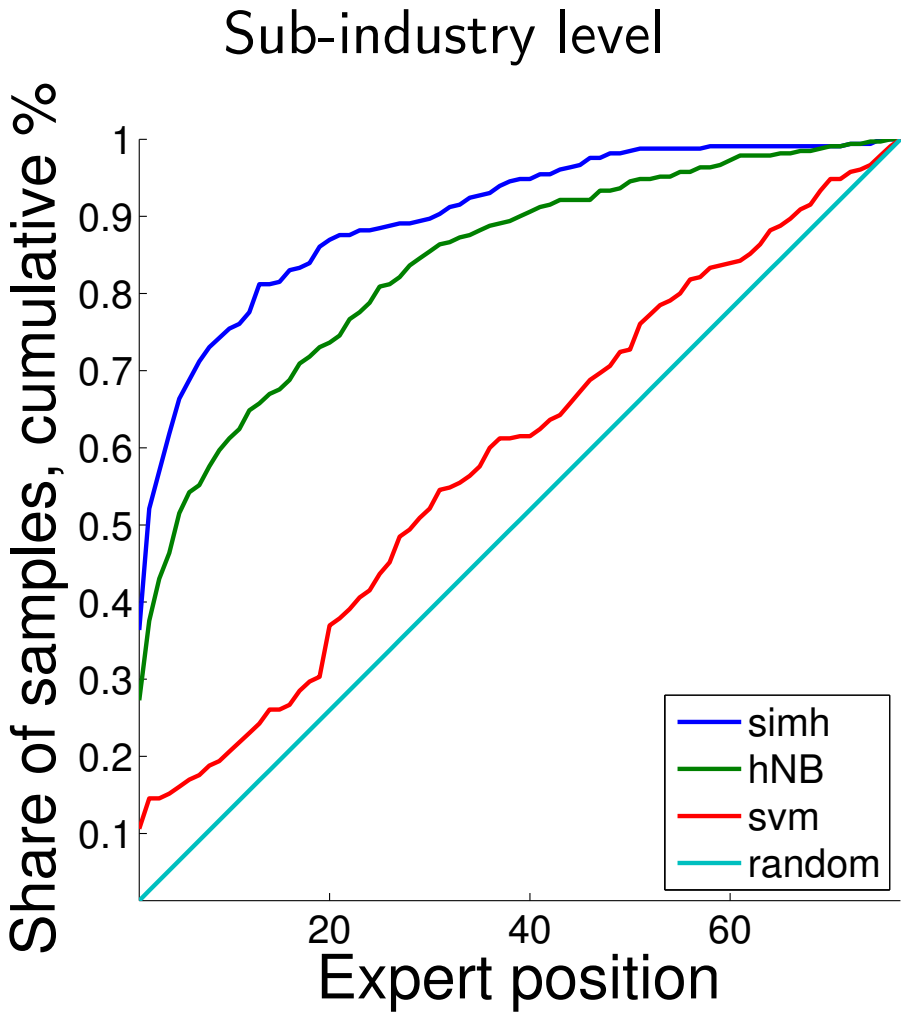
The documents collection: industry sector web sites



- Collection: $|D| = 1036$ web sites classified by experts into 11 industries and 78 sub-industries.
- Dictionary: $|W| = 18000$ terms.
- Preprocessing: combine all web pages of a site, remove all html-tags

Quality comparison, Industry sector web sites





	svm	hNB	simh
AUCH(R), Industry level	0.70	0.80	**0.83**
AUCH(R), Sub-industry level	0.59	0.83	**0.89**

Conclusion

- The weighted hierarchical similarity function is proposed.
- The entropy-based method to calculate terms significance matrix $\mathbf{\Lambda}$ and optimize level weights $\boldsymbol{\theta}$ is proposed.
- The relevance operator is proposed.
- The hierarchical similarity approach shows the same quality as the strongest base line with the expert dictionary.
- The hierarchical similarity approach shows significantly better quality with the automatically created dictionary.