Thematic classification for EURO/IFORS conference using expert model

Arsenty Kuzmin, Alexander Aduenko, and Vadim Strijov

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

IFORS 2014, Barcelona
17.06.2014

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

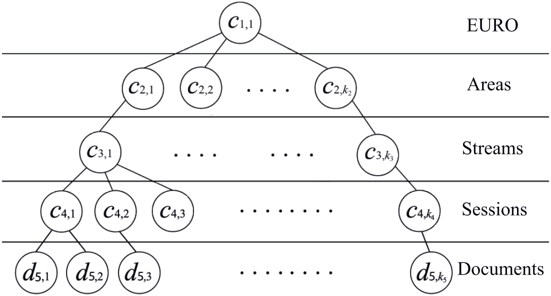
There given:

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

The main idea:

- to join all thematic models into one,
- to calculate the similarity of a new abstract and each Stream of the unified model
- to show the most similar Streams to the Experts

EURO/IFORS conference hierarchical model



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

Challenges

Causes of the problems

1. Great number of the experts (more than 200),
2. expert classification could be controversial,
3. there is no base thematic model.

The terms of the document determine its theme

$W = \{w_1, \dots, w_n\}$ is the terms dictionary of the conference

Let the document be the bag of words

The document d of the collection D is an unordered set of words of the dictionary W, $d = \{w_j\},\ j \in \{1, \dots, n\}$.

The more documents contain some term, the less information this term gives us about clustering.

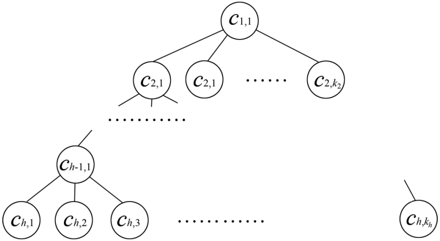
Terms significance matrix Λ:

$$\Lambda = \mathrm{diag}\{\lambda_{1,1},\ \dots,\ \lambda_{n,n}\},\ \text{normalization:}\quad \mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^{\mathsf{T}} \Lambda \mathbf{x}_s}}$$

Hierarchical representation of the thematic model

Each leaf (h, i) of the tree corresponds to the document d_i.

Each node (ℓ, i), $\ell \neq h$ corresponds to the cluster $c_{\ell,i}$, which consists of corresponding documents.



Here ℓ is a conference level, $h = 5$ is the number levels and i is the index of a node given level.

Similarity function

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

$$s(\mathbf{x}_i,\ \mathbf{x}_j) = \frac{\mathbf{x}_i^{\mathsf{T}}\Lambda\mathbf{x}_j}{\sqrt{\mathbf{x}_i^{\mathsf{T}}\Lambda\mathbf{x}_i}\sqrt{\mathbf{x}_j^{\mathsf{T}}\Lambda\mathbf{x}_j}} = \mathbf{x}_i^{\mathsf{T}}\Lambda\mathbf{x}_j.$$

Define the similarity function $S(\cdot,\cdot)$ between clusters $c_{\ell,i}$ and $c_{\ell,j}$ as the mean $s(\mathbf{x},\mathbf{y})$ between their documents $\mathbf{x}\in c_{\ell,i}, \mathbf{y}\in c_{\ell,j}$

$$S(c_{\ell,i},\ c_{\ell,j}) = \frac{1}{|A|}\sum_{(\mathbf{x},\ \mathbf{y})\in A} s(\mathbf{x},\ \mathbf{y}),$$

where A is the set of all document pairs from clusters $c_{\ell,i}$ and $c_{\ell,j}$, $\mathbf{x}\in c_{\ell,i}$, $\mathbf{y}\in c_{\ell,j}$, $\mathbf{x}\neq\mathbf{y}$.

Similarity function

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

$$s(\mathbf{x}_i,\ c_{\ell,j}) = \mathbf{x}^{\mathsf{T}}\Lambda\overline{\mathbf{x}}_{\ell,\,i},$$

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

Similarity between document and cluster of the h level

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

where θ_{h-j} is the sinificance of the level $h-j$ and B^j is the operator of the precedence that associate cluster $c_{h,\,i}$ with its predecessor on the level j.

The clustering quality function

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

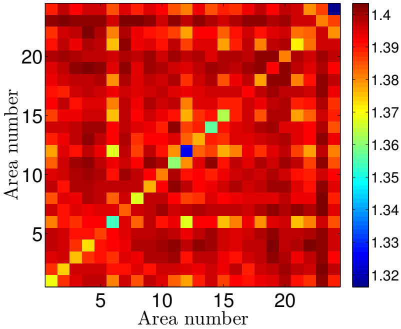
and F_1 is a mean inter-cluster similarity: $F_1 = \frac{2}{k_\ell(k_\ell - 1)} \sum_{i<j} S(c_{\ell,i},\ c_{\ell,j})$

Clustering quality criterion

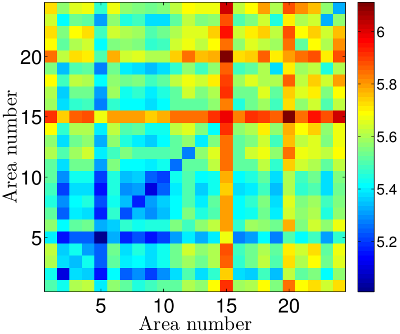
$$F = \frac{F_1}{F_0} \to \min$$

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

Distance and similarity functions comparison

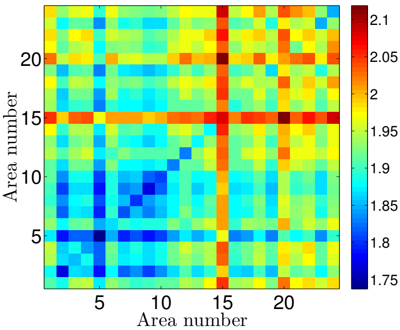


Euclidean distance

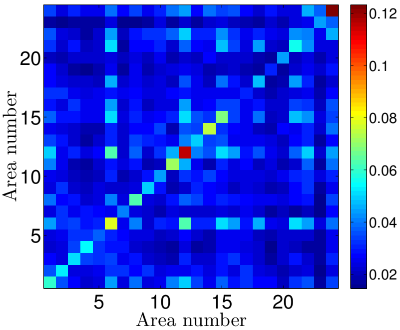


Hellinger distance

Distance and similarity functions comparison



Jenson-Shannon distance



Proposed similarity function

The relevance operator R

Let 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.
$S^{k_h} = \{3, 1, \dots, 6\}$

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 of the lowest hierarchy level clusters

Let $\mathrm{pos}(s, j) : S^q \times \{1, 2, \dots, q\} \to \{1, 2, \dots, q\}$ be the position function

It returns the position of the given number in the permutation.

The baseline relevance operator R_1

Sort all clusters of the h hierarchy level by their size.

Let $c_{h,\,i_1},\,\ldots,\,c_{h,\,i_{k_h}}$ be the corresponding order of level h clusters:

$$|c_{h,\,i_1}| \geq |c_{h,\,i_2}| \geq \ldots \geq |c_{h,\,i_{k_h}}|.$$

Clusters of the equal size have some fixed order.

Let $R_1(\cdot) = (i_1,\, i_2,\, \ldots,\, i_{k_3})$ be the baseline relevance operator

R_1 returns the permutation S^{k_h} of the ordered-by-size level h clusters for all documents.

Quality criterions $Q(R)$ and $AUC(R)$

$Q(R)$ quality criterion

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

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

AUC(R) quality criterion

AUC(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 $i \in [1,\ k_h]$.

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

Terms significance

Denote by $\mathbf{p}_\ell^j$ the vector of j-th components of mean vectors $\bar{\mathbf{x}}_{\ell,i}$

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

The word entropy

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

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

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 collections

The purpose of the experiment

Construct a thematic model of the conference EURO 2010

The collection D^1:

We matched the Areas and the Streams from collections:

- EURO 2012, $|D| = 1342$, 26 Areas, 141 Streams.
- EURO 2013, $|D| = 2313$, 24 Areas, 137 Streams.

The unified structure has 24 Areas, 178 Streams.

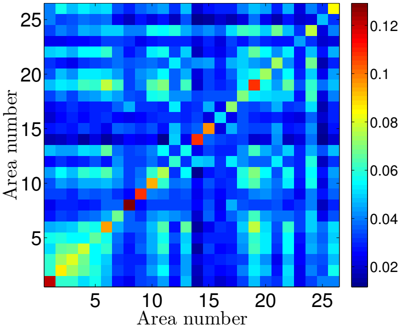
The collection D^2:

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

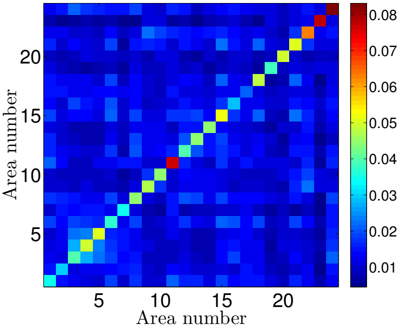
15 out of 178 streams are present only in the year 2010.

Size of the dictionary:

- $|W| = 1675$ terms.

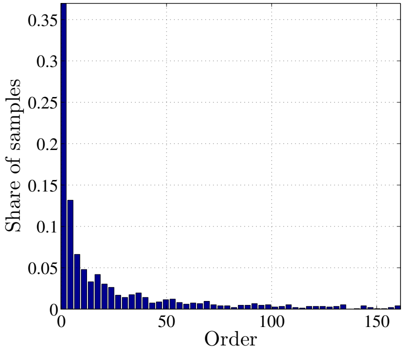


Areas similarity, $\lambda_i = 1$

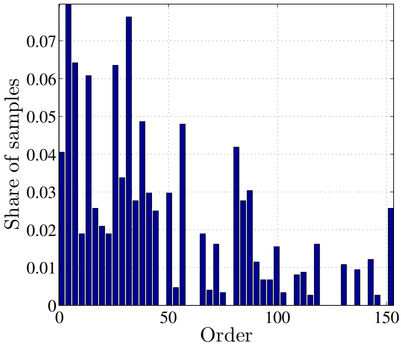


Areas similarity, optimized λ

Quality comparison Q(R)

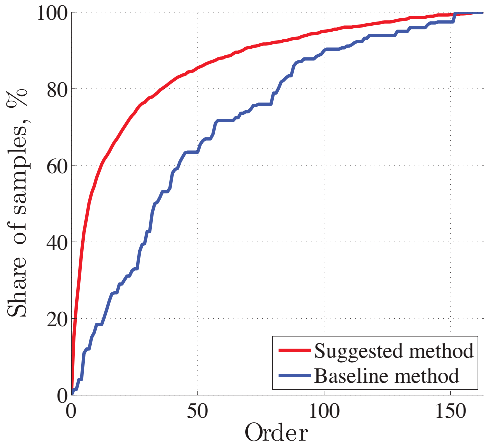


Proposed relevance operator, $Q = 22.54$



Baseline relevance operator $R_1(\cdot)$, $Q = 46.86$

Quality comparison AUC(R)



$$\mathrm{AUC}(R) = 0.868,\ \mathrm{AUC}(R_1) = 0.719$$

Implementation: http://europrogramadvisor.com

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

Conclusion

- The weighted cosine similarity function is proposed.
- The entropy-based method to calculate terms significance is proposed.
- The relevance operator is proposed.