Построение иерархических тематических моделей крупных конференций

Кузьмин Арсентий

Московский физико-технический институт
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Научный руководитель д.ф.-м.н. В. В. Стрижов

Москва,
2015 г.

Цель работы

При поиске экспертом наиболее релевантной темы для доклада на крупной конференции возникают следующие проблемы:

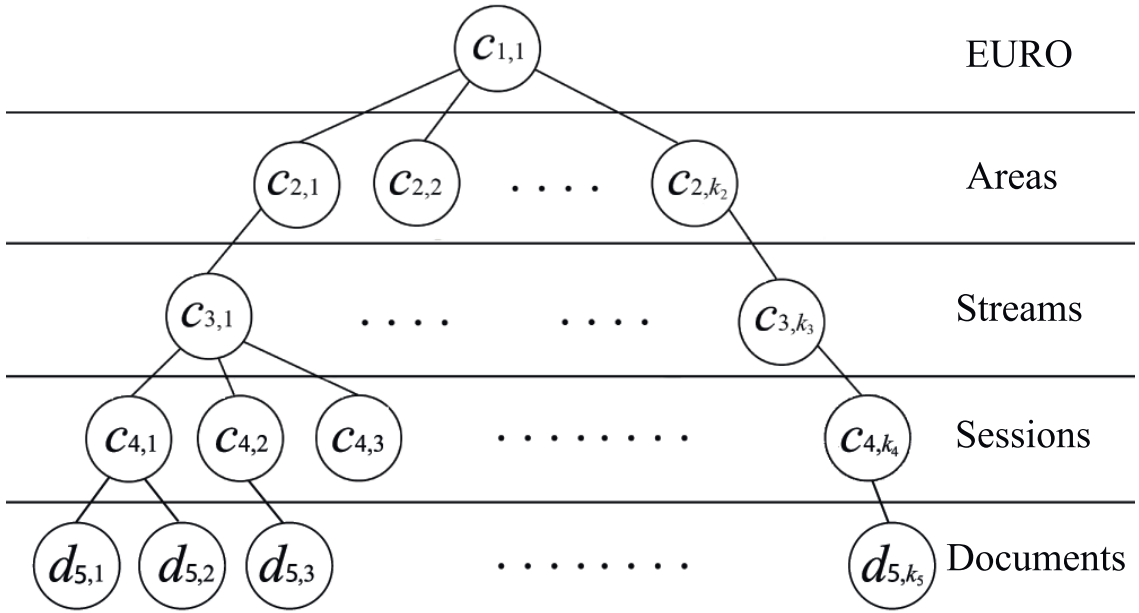
- Большое число тем
- Отсутствие эталонной модели
- Субъективность экспертного решения

Задача

Предложить способ поиска наиболее релевантных тем для нового документа, используя экспертные модели прошлых конференций.

Для решения задачи предлагается использовать методы иерархической жесткой кластеризации.

Иерархическая модель конференции EURO/IFORS



1. Участники подают документы в общую коллекцию.
2. За каждую область отвечает группа экспертов.
3. Эксперты распределяют документы в свои направления.
4. Внутри каждого направления формируются сессии.

Тема документа определяется его терминами

$W = \{w_1, \ldots, w_n\}$ терминологический словарь конференции

Документ — мешок слов

Документ d из коллекции D – неупорядоченный набор слов из словаря W, $d = \{w_j\}$, $j \in \{1, \ldots, n\}$.

Матрица важности терминов $\mathbf{\Lambda}$:

$$\mathbf{\Lambda} = \operatorname{diag}\{\lambda_{1,1}, \ \ldots, \ \lambda_{n,n}\}, \text{ нормализация: } \mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^{\mathsf{T}} \mathbf{\Lambda} \mathbf{x}_s}}$$

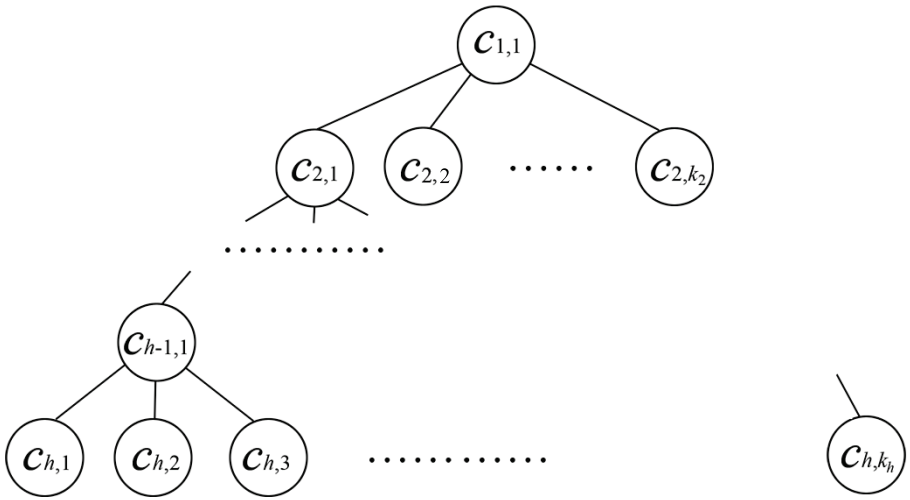
Оптимизация $\mathbf{\Lambda}$ по экспертной модели

$$\mathbf{\Lambda}^* = \arg\min_{\mathbf{\Lambda}} Q(R)$$

Иерархическое представление тематической модели

Каждому листу дерева (h, i) соответствует документ d_i.

Каждому узлу (ℓ, i), $\ell \neq h$ соответствует кластер $c_{\ell,i}$, содержащий в себе документы, путь до которых от вершины дерева $c_{1,1}$ проходит через данный узел (ℓ, i).



h — число уровней конференции, ℓ — уровень конференции, i — порядковый номер узла на уровне.

Функция сходства двух документов и двух кластеров

Сходством $s(\cdot,\cdot)$ документов $\mathbf{x}_i$ и $\mathbf{x}_j$ называется:

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

Сходством $S(\cdot,\cdot)$ кластеров $c_{\ell,i}$ и $c_{\ell,j}$ называется сходство $s(\mathbf{x},\mathbf{y})$ между их документами $\mathbf{x}\in c_{\ell,i}, \mathbf{y}\in c_{\ell,j}$

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

где A – множество всех пар документов из кластеров $c_{\ell,i}$ и $c_{\ell,j}$, $\mathbf{x}\in c_{\ell,i}$, $\mathbf{y}\in c_{\ell,j}$, $\mathbf{x}\neq\mathbf{y}$.

Функция сходства документа и кластера

Сходством $s(\cdot,\cdot)$ документа $\mathbf{x}_i$ и кластера $c_{\ell,j}$ на уровне иерархии ℓ называется:

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

где $\overline{\mathbf{x}}_{\ell,\,i}$ – средний вектор кластера $c_{\ell,\,i}$.

Сходство документа и кластера на уровне h

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

где θ_{h-j} значимость уровня $h-j$, а B^j – оператор, возвращающий для каждого кластера $c_{h,\,i}$ его родительский кластер уровня j.

Оператор релевантности R

Определение

Пусть $s \in S^{k_h}$ – перестановка, соответствующая сортировке кластеров уровня h по сходству с документом $\mathbf{x}$ в порядке убывания, где k_h – количество кластеров.

Определение

Пусть $R : \mathbb{R}^n \to S^{k_h}$ – оператор релевантности, ставящий в соответствие каждому документу $\mathbf{x} \in \mathbb{R}^n$ перестановку $s \in S^{k_h}$.

Определение

Пусть $\mathrm{pos}(s, j) : S^q \times \{1, 2, \ldots, q\} \to \{1, 2, \ldots, q\}$ – функция позиции, возвращающая индекс числа j в перестановке s.

Базовый оператор релевантности R_1

Для проверки качества предложенного оператора R сравним его с оператором R_1, возвращающим псевдослучайную перестановку.

Пусть $c_{h,i_1}, \ldots, c_{h,i_{k_h}}$ – порядок кластеров уровня h, такой, что:

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

Определение

Пусть $R_1(\cdot) = (i_1, i_2, \ldots, i_{k_3})$ – базовый оператор релевантности R_1 возвращает перестановку S^{k_h} отсортированных кластеров уровня h для всех документов.

Критерии качества $Q(R)$ и $\mathrm{AUC\ CH}(R)$

Критерий качества $Q(R)$

Пусть $Q(R)$ – усредненная позиция экспертного кластера $z_{j,h}$ в перестановке $R(\mathbf{x}_j)$:

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

Критерий качества $\mathrm{AUC\ CH}(R)$

$\mathrm{AUC\ CH}(R) \in [0,\, 1]$ – площадь под кривой гистограммы $\#\{\mathrm{pos}(R(\mathbf{x}_j),\, z_{j,h}) \leq i\}$, где $z_{j,h}$ – номер экспертного кластера документа $\mathbf{x}_j$, а $i \in [1,\, k_h]$.

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

Важность терминов

Пусть $\mathbf{p}_{\ell,j}$ – вектор из j-ых компонент средних векторов $\bar{\mathbf{x}}_{\ell,i}$ кластеров $c_{\ell,i}$ уровня ℓ.

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

Энтропия слов

Определим энтропию $I_\ell(w_j)_{k_\ell}$ слова w_j для уровня иерархии ℓ как

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

Важность термина w_j через энтропию

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

Оптимизация по экспертной тематической модели

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

Коллекция документов

Цель эксперимента

Построить тематическую модель конференции EURO 2010

Коллекция D^1:

Области и направления объединены для коллекций:

- EURO 2012, $|D| = 1342$, 26 Областей, 141 Направление.
- EURO 2013, $|D| = 2313$, 24 Области, 137 Направлений.

Объединенная модель содержит 24 Области, 178 Направлений.

Коллекция D^2:

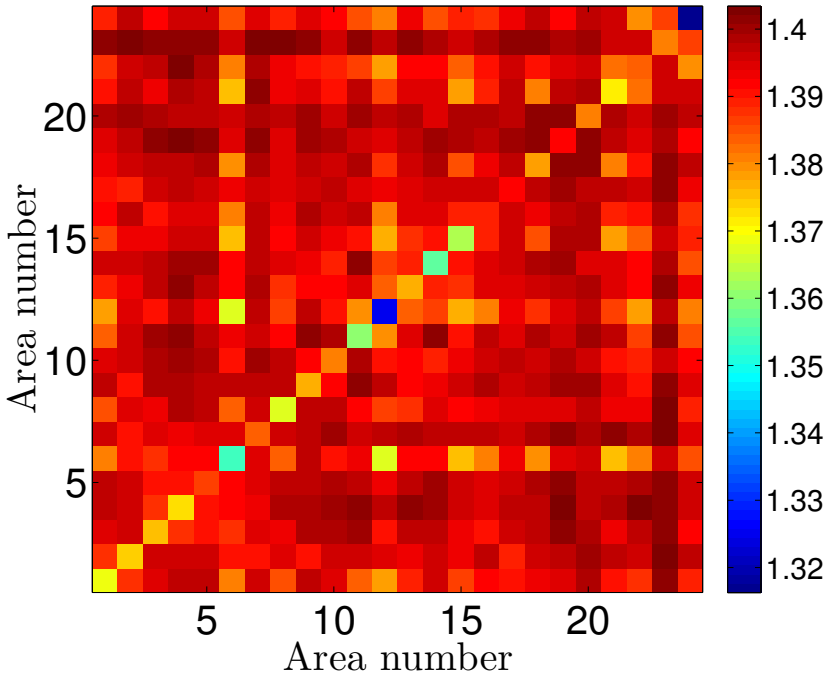
- EURO 2010, $|D| = 1663$, 26 Областей, 113 Направлений.

15 из 178 Направлений представлены только в коллекции 2010 года.

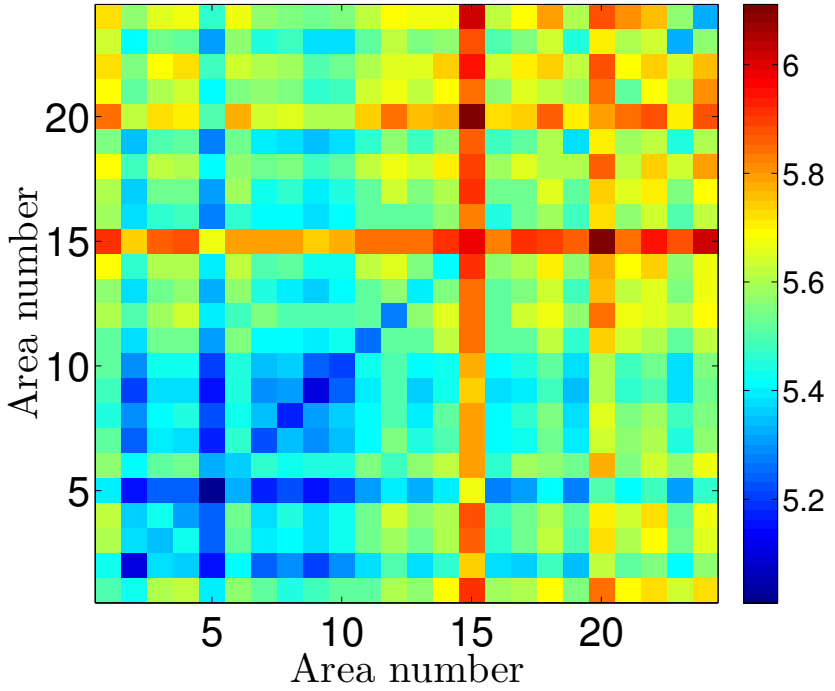
Размер словаря:

- $|W| = 1675$ терминов.

Сравнение функции расстояния и сходства

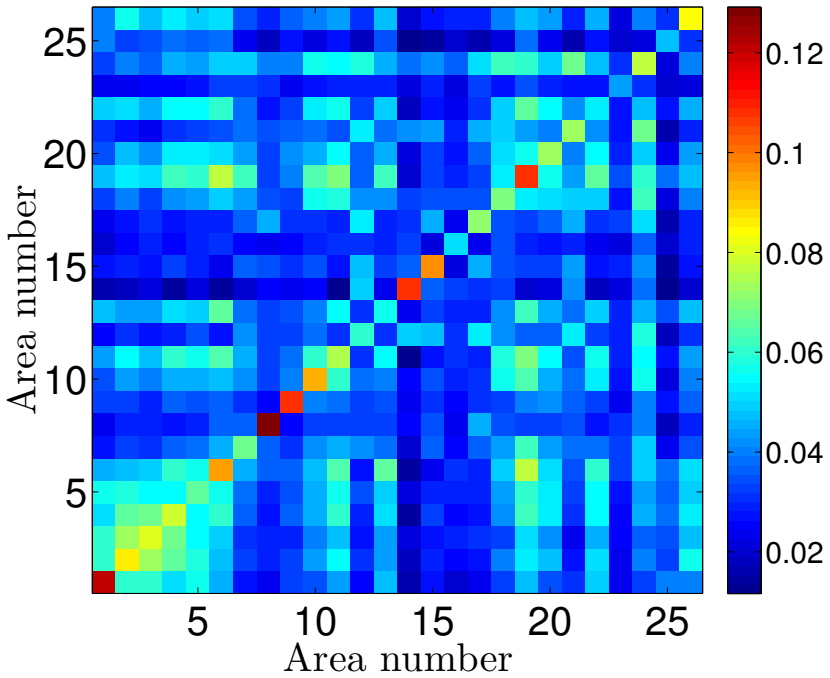


Расстояние Евклида

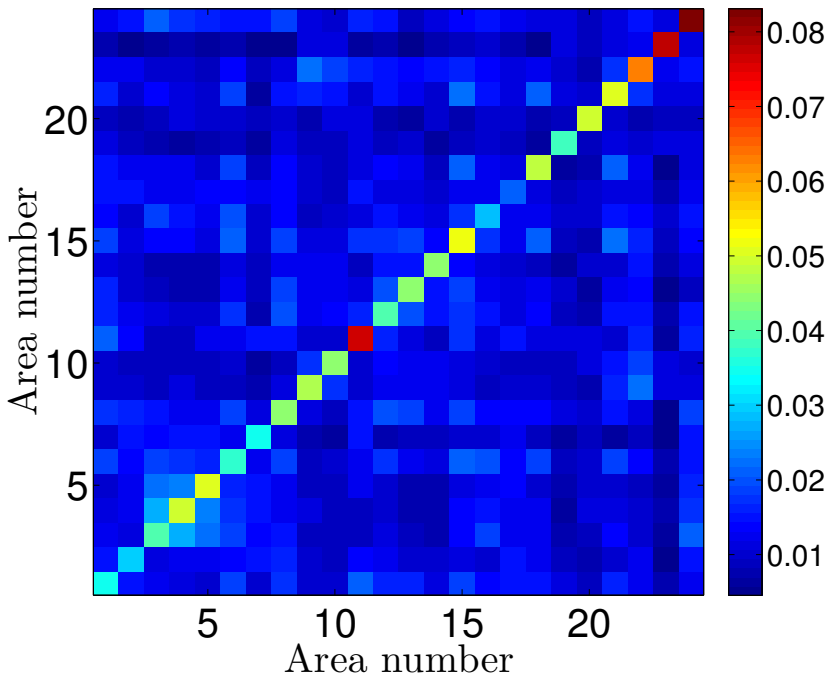


Расстояние Хеллингера

Сравнение функции расстояния и сходства

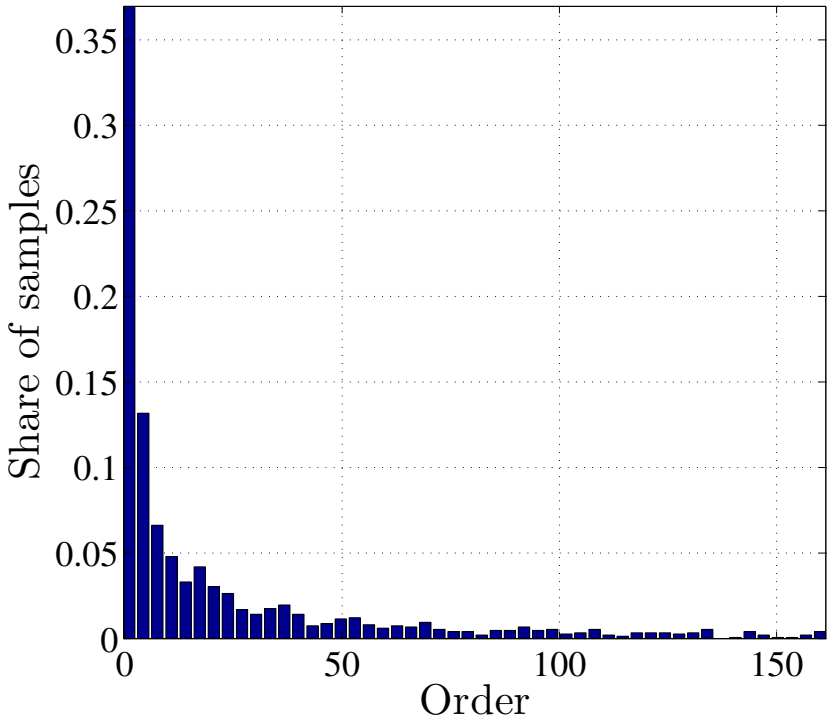


Сходство Областей, $\lambda_i = 1$

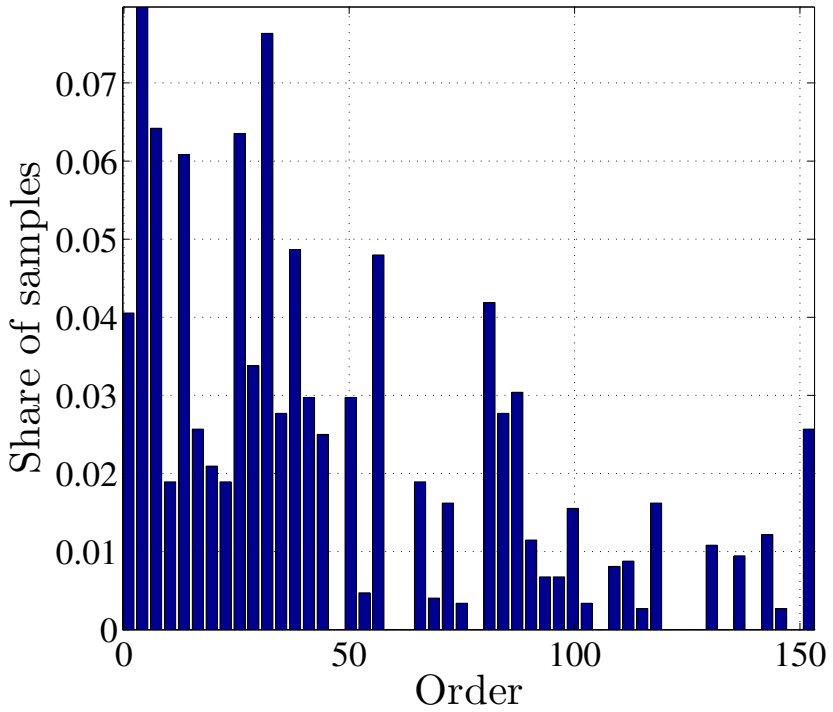


Сходство Областей, оптимизированные λ

Сравнение качества Q(R)

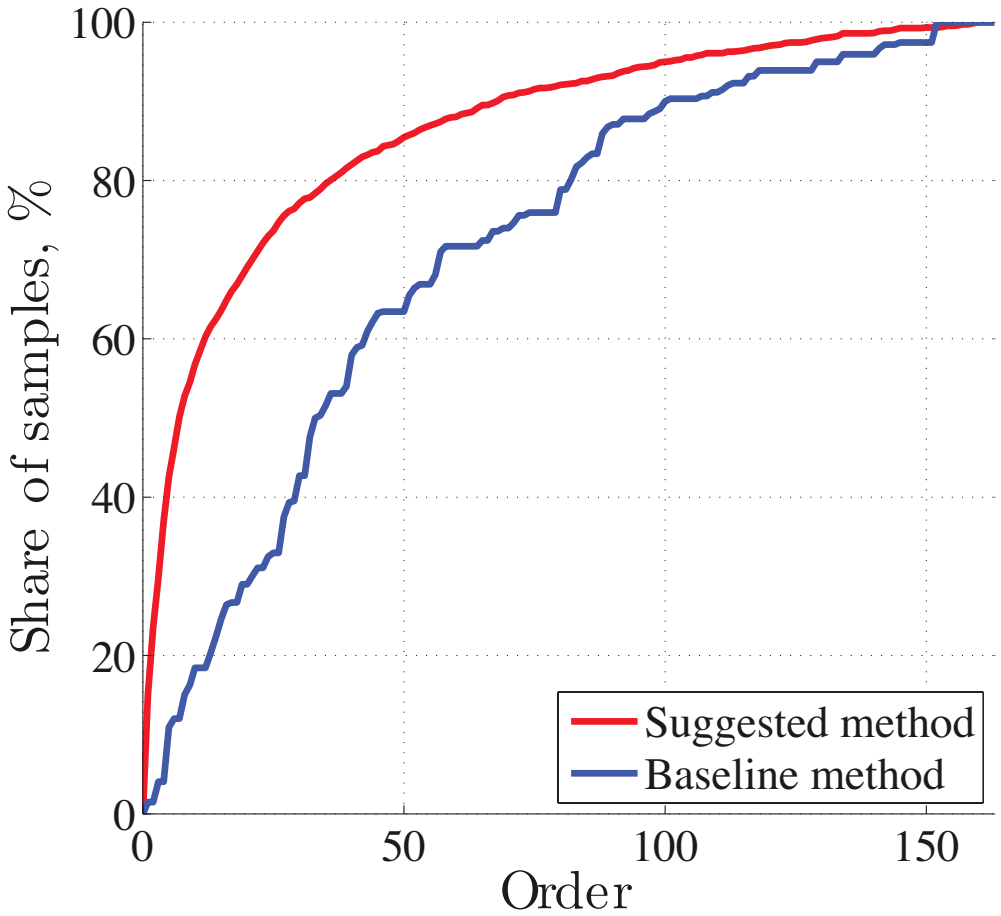


Предложенный оператор релевантности, $Q = 22.54$



Базовый оператор релевантности $R_1(\cdot)$, $Q = 46.86$

Сравнение качества AUC CH(R)



$$\text{AUC CH}(R) = 0.868, \quad \text{AUC CH}(R_1) = 0.719$$

Реализация: 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

Заключение

- Предложена взвешенная косинусная функция сходства двух документов, двух кластеров, документа и кластера;
- Предложен оператор релевантности и его базовая версия;
- Предложено два способа оценки качества оператора релевантности;
- Предложен энтропийный метод оценки важности терминов и метод оптимизации весов $\boldsymbol{\Lambda}$;
- Разработана экспертная система, ранжирующая области и направления по убыванию сходства с заданным документом.

Публикации по теме

- Кузьмин А. А. Многоуровневая классификация при обнаружении движения цен // Машинное обучение и анализ данных. 2012. № 3. С. 318-327.
- А. А. Кузьмин, А. А. Адуенко, В. В. Стрижов Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ. 2012. № 3. С. 119-131.
- А. А. Кузьмин, В. В. Стрижов Проверка адекватности тематических моделей коллекции документов. // Программная инженерия. 2013. № 4. С. 16-20
- Кузьмин А. А., Адуенко А. А., Стрижов В. В. Тематическая классификация тезисов крупной конференции с использованием экспертной модели // Информационные технологии, 2014, № 6. С. 22-26.