Построение иерархических тематических моделей крупных конференций

А. А. Кузьмин, А. А. Адуенко, В. В. Стрижов

Московский физико-технический институт

ММРО, Светлогорск
23 сентября 2015 г.

Предложить экспертную систему, упрощающую процесс построения тематической модели предстоящей конференции

Задача

- Построить тематическую модель крупной конференции

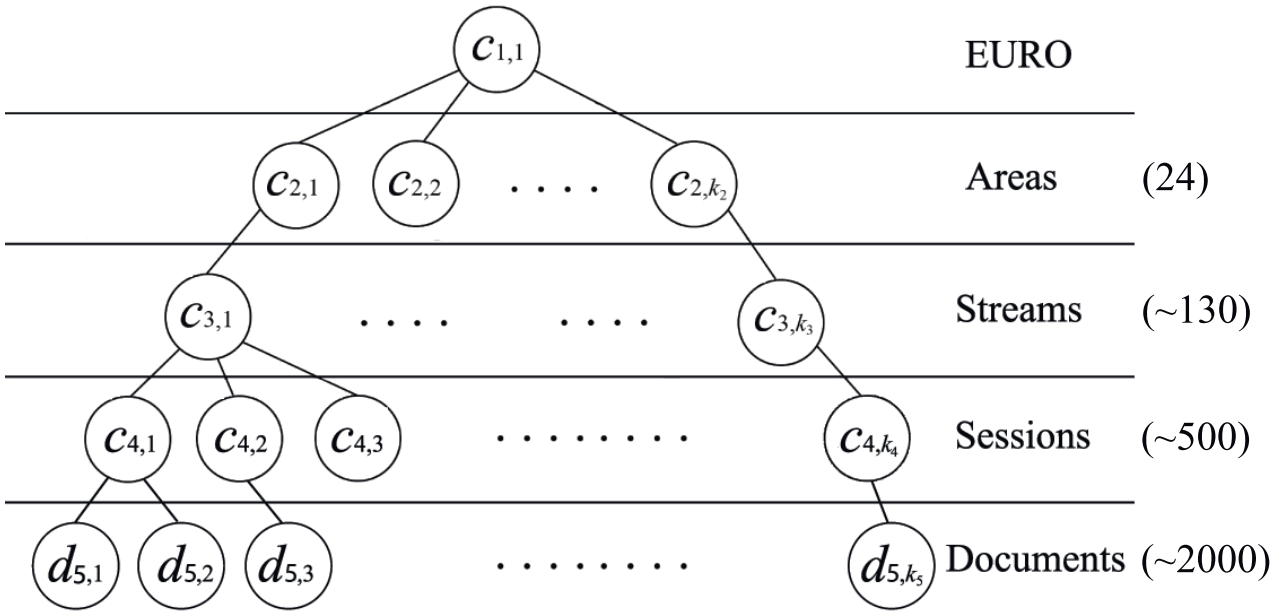
Данные

- Экспертные тематические модели конференций прошлых лет
- Аннотации к докладам участников предстоящей конференции

Требуется

- Предложить способ поиска наиболее релевантных тем для нового документа

Иерархическая модель конференции EURO/IFORS



1. Участники подают документы в общую коллекцию.
2. Часть участников являются приглашенными, их документы уже имеют фиксированный кластер.
3. За каждую область отвечает группа экспертов.
4. Эксперты распределяют оставшиеся документы в свои направления.

Тема документа определяется его терминами

$W = \{w_1, \ldots, w_n\}$ – терминологический словарь конференции.
Для построения W производится:

- Удаление стоп-слов
- Нормализация слов в документах
- Объединение синонимов (с помощью экспертов)
- Удаление часто (редко) встречающихся слов

Документ — мешок слов

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

Пусть $\mathbf{x}(d) \in \mathbb{R}^{|W|}$ – векторное представление документа d, где $x_i = \#\{w_i \in d\}$

Построение списка индексов кластеров, ранжированного по релевантности документу

Определение

Пусть $q(\mathbf{x}) \in S^{k_h}$ – перестановка, соответствующая сортировке кластеров нижнего уровня h в порядке убывания релевантности документу $\mathbf{x}$, где k_h – количество кластеров.

Пример: $q(\mathbf{x}) = \{3, 1, \ldots, 6\}$.

Определение

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

Определение

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

Пример: $\mathrm{pos}(q, 1) = 2$.

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 $i \in [1,\, k_h]$:

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

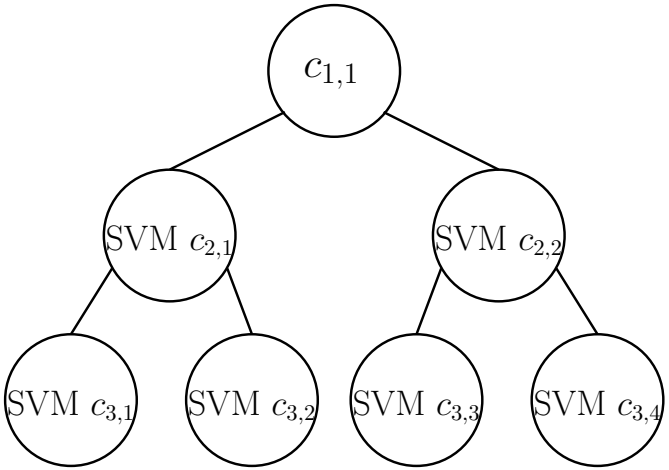
Построение оператора R с помощью SVM

Для каждого кластера $c_{\ell,i}$ уровня ℓ обучается двухклассовый SVM по принципу "один против всех".

Вероятность оценивается с помощью метода Платта (Platt, 2000):

$$P(c_{l,i}|\mathbf{x}) = \frac{1}{1+\exp(A\cdot\hat{m}(\mathbf{x})+B)},$$

где $\hat{m}(\mathbf{x})$ – отступ объекта $\mathbf{x}$,
A, B – числовые параметры.



Пусть для некоторого объекта $\mathbf{x}$ вероятности $p(c_{2,1}) > p(c_{2,1})$, $p(c_{3,1}) > p(c_{3,2})$, $p(c_{3,4}) > p(c_{3,3})$.

Оператор релевантности $R_{SVM}(\mathbf{x})$ вернет следующую перестановку:

$$R_{SVM}(\mathbf{x}) = (\overbrace{1,2}^{c_{2,1}}, \overbrace{4,3}^{c_{2,2}})$$

Функция сходства двух документов и двух кластеров

Матрица важности терминов $\mathbf{\Lambda} = \operatorname{diag}\{\lambda_{1,1}, \ \ldots, \ \lambda_{n,n}\}$.

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

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\mathsf{T}} \mathbf{\Lambda} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^{\mathsf{T}} \mathbf{\Lambda} \mathbf{x}_i}\sqrt{\mathbf{x}_j^{\mathsf{T}} \mathbf{\Lambda} \mathbf{x}_j}} = \mathbf{x}_i^{\mathsf{T}} \mathbf{\Lambda} \mathbf{x}_j, \ \text{нормализация:} \quad \mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^{\mathsf{T}} \mathbf{\Lambda} \mathbf{x}_s}}$$

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 $c_{\ell,j}$, $\mathbf{x} \in c_{\ell,i}$, $\mathbf{y} \in c_{\ell,j}$, $\mathbf{x} \neq \mathbf{y}$.

Построение оператора R с помощью SuHiPLSA (ARTM)

Строится регуляризированная вероятностная тематическая модель SuHiPLSA:

$$\boldsymbol{\Phi}^*, \boldsymbol{\Theta}^* = \arg\max_{\boldsymbol{\Phi},\boldsymbol{\Theta}} L(\boldsymbol{\Phi}, \boldsymbol{\Theta}) + \lambda F(\mathbf{t}, \hat{\mathbf{t}}),$$

где $\lambda F(\mathbf{t}, \hat{\mathbf{t}})$ – штраф за несоответствие прогнозных тем $\hat{\mathbf{t}}$ экспертным $\mathbf{t}$.

Матрица "слово-тема" $\boldsymbol{\Phi}^*$ – определяет вектора кластеров (тем). Используя меру сходства $s(\mathbf{x}, \phi)$ с $(\boldsymbol{\Lambda} = E)$ построим ранжированный список кластеров нижнего уровня по убыванию сходства с новым документом $\mathbf{x}$:

$$s(\mathbf{x}, \phi_{i_1}) \geq s(\mathbf{x}, \phi_{i_2}) \geq \cdots \geq s(\mathbf{x}, \phi_{i_{k_h}})$$

Оператор релевантности $R_{PLSA}(\mathbf{x}) = (i_1, i_2, \ldots, i_{k_h})$.

Функция сходства документа и кластера

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

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

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

Иерархическое сходство документа $\mathbf{x}_i$ и кластера $c_{\ell,j}$ на уровне ℓ называется:

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

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

Построение оператора R с помощью иерархической функции сходства

Поставим в соответствие новому документу $\mathbf{x}$ перестановку $q = (i_1, i_2, \ldots, i_{k_h})$ такую, что:

$$s(\mathbf{x}, c_{h,i_1}) \geq s(\mathbf{x}, c_{h,i_2}) \geq \cdots \geq s(\mathbf{x}, c_{h,i_{k_h}}).$$

Предложенный оператор релевантности R_{SIM} будет иметь вид:

$$R_{SIM}(\mathbf{x}) = (i_1, i_2, \ldots, i_{k_h}).$$

Оптимизация по экспертной тематической модели

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

Модель оценки важности 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 $I_\ell(w_j)$ слова w_j для уровня иерархии ℓ как

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

Важность термина w_j через энтропию

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

Оптимизация по экспертной тематической модели

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

Вычислительный эксперимент

Построить тематическую модель конференции EURO 2010.

Обучающая выборка D^1:

- EURO 2012, $|D| = 1342$, 26 областей, 141 направление.
- EURO 2013, $|D| = 2313$, 24 области, 137 направлений.

Объединенная модель содержит 24 области, 178 направлений.

Тестовая выборка D^2:

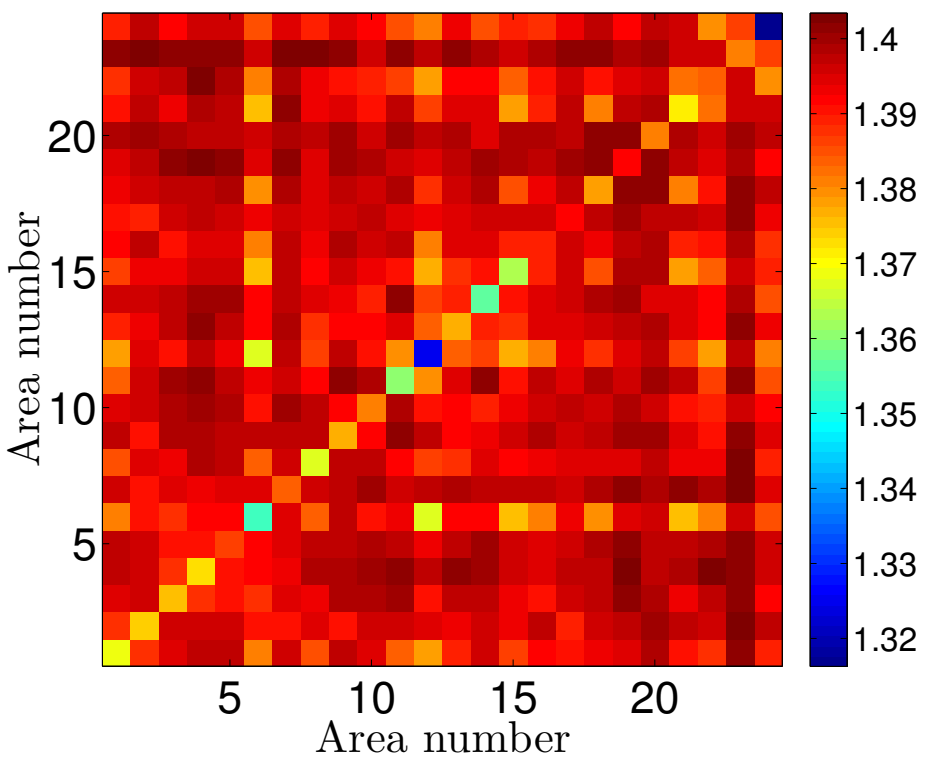
- EURO 2010, $|D| = 1663$, 26 областей, 113 направлений.

15 из 178 направлений представлены только в коллекции 2010 года.

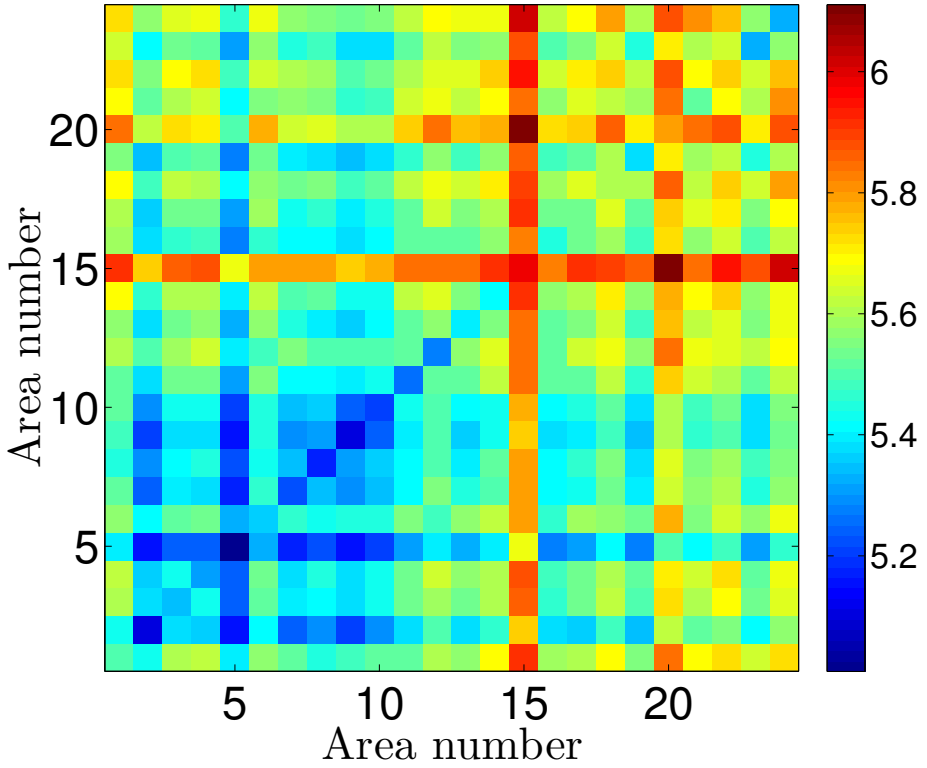
Размер словаря:

- $|W| = 1675$ терминов.

Сравнение функции расстояния и сходства



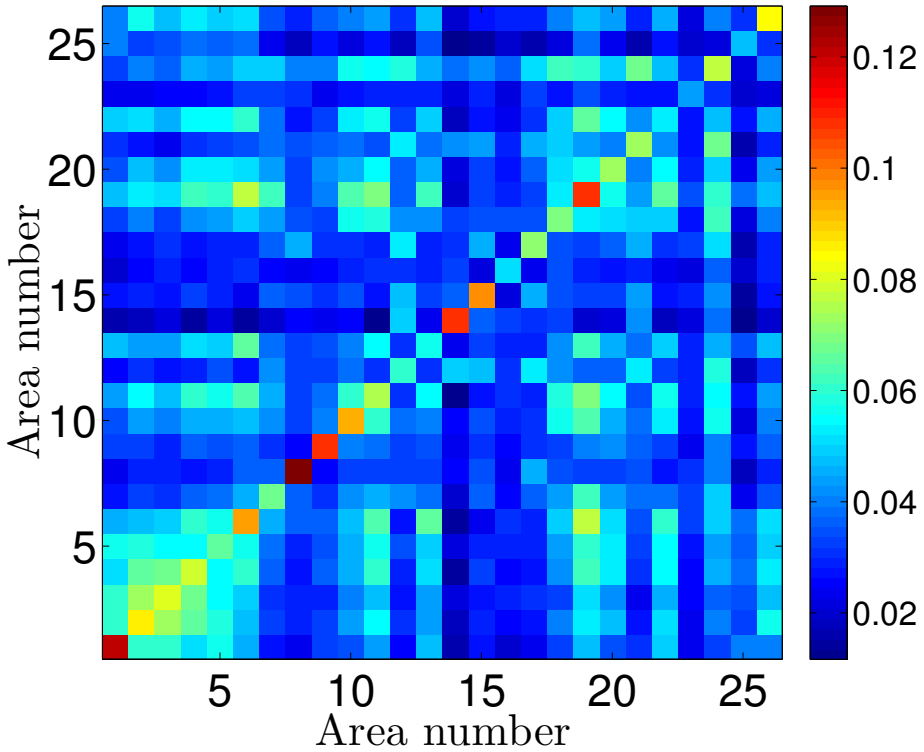
Расстояние Евклида



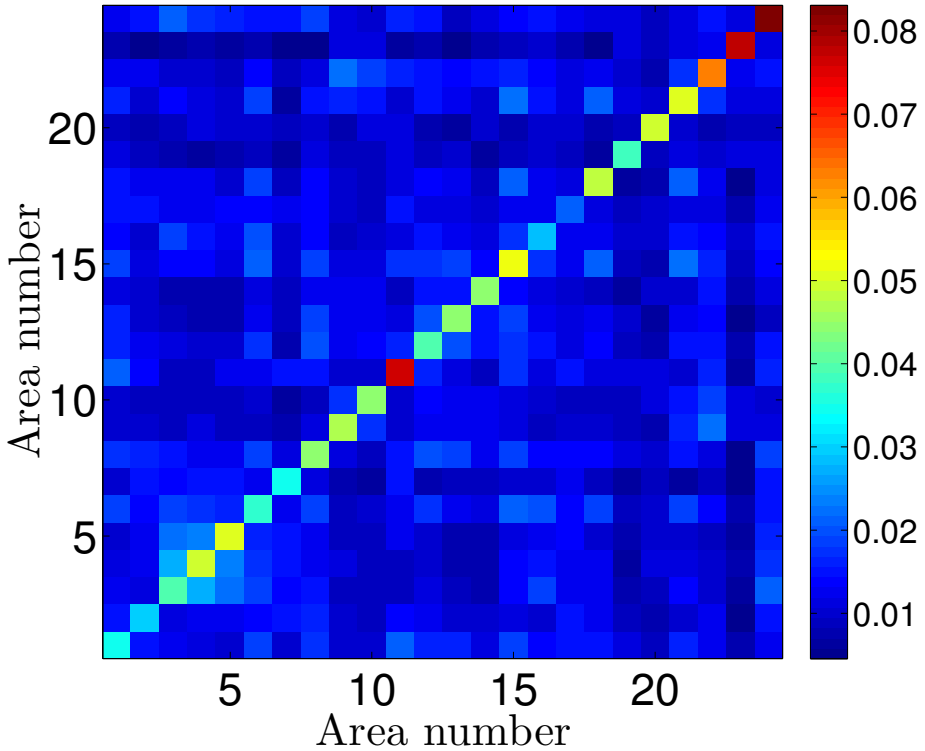
Расстояние Хеллингера

Цвет точки (x, y) на графике соответствует значению расстояния между кластерами уровня area с номерами x и y.

Сравнение функции расстояния и сходства



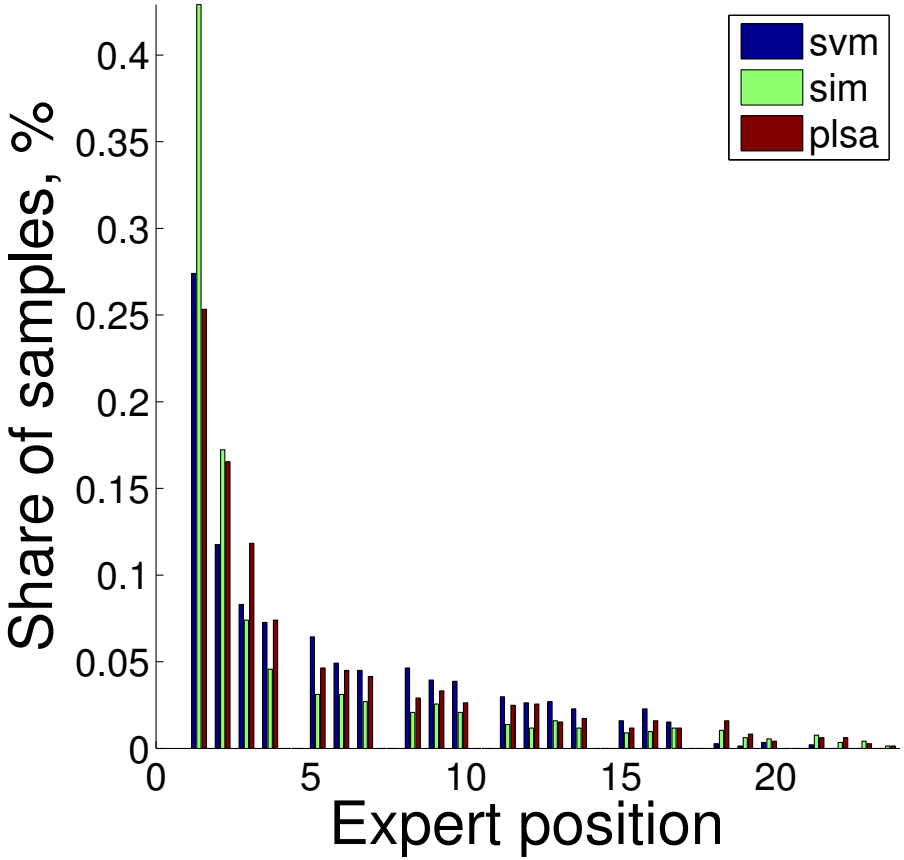
Сходство Областей, $\lambda_i = 1$



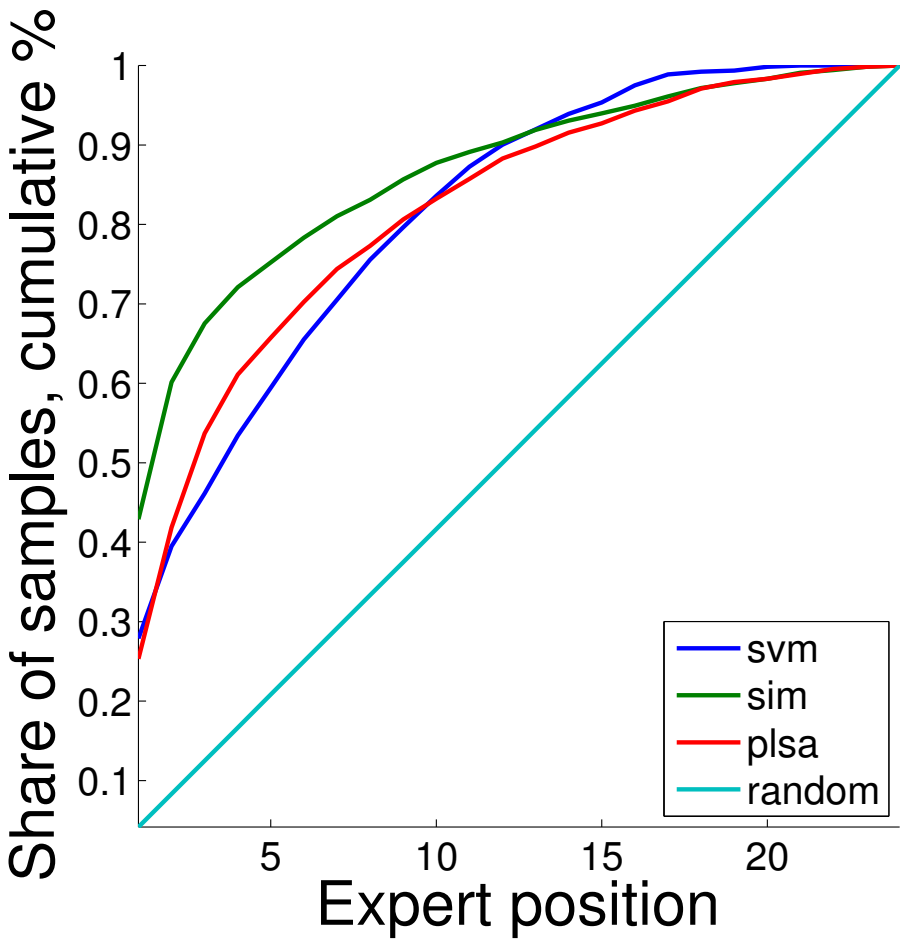
Сходство Областей, оптимизированные λ

Цвет точки (x, y) на графике соответствует значению сходства между кластерами уровня area с номерами x и y.

Сравнение качества, уровень Area

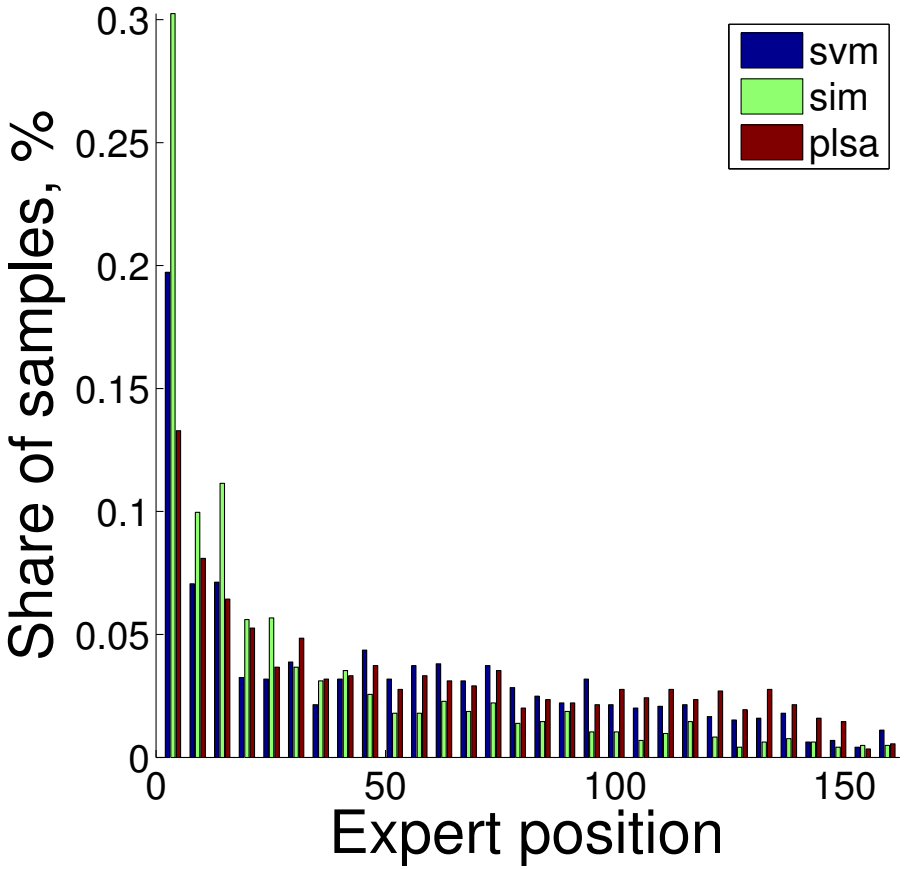


Средняя позиция экспертного кластера Q(R): SVM - 5.5, SIM - **4.3**, PLSA - 5.4

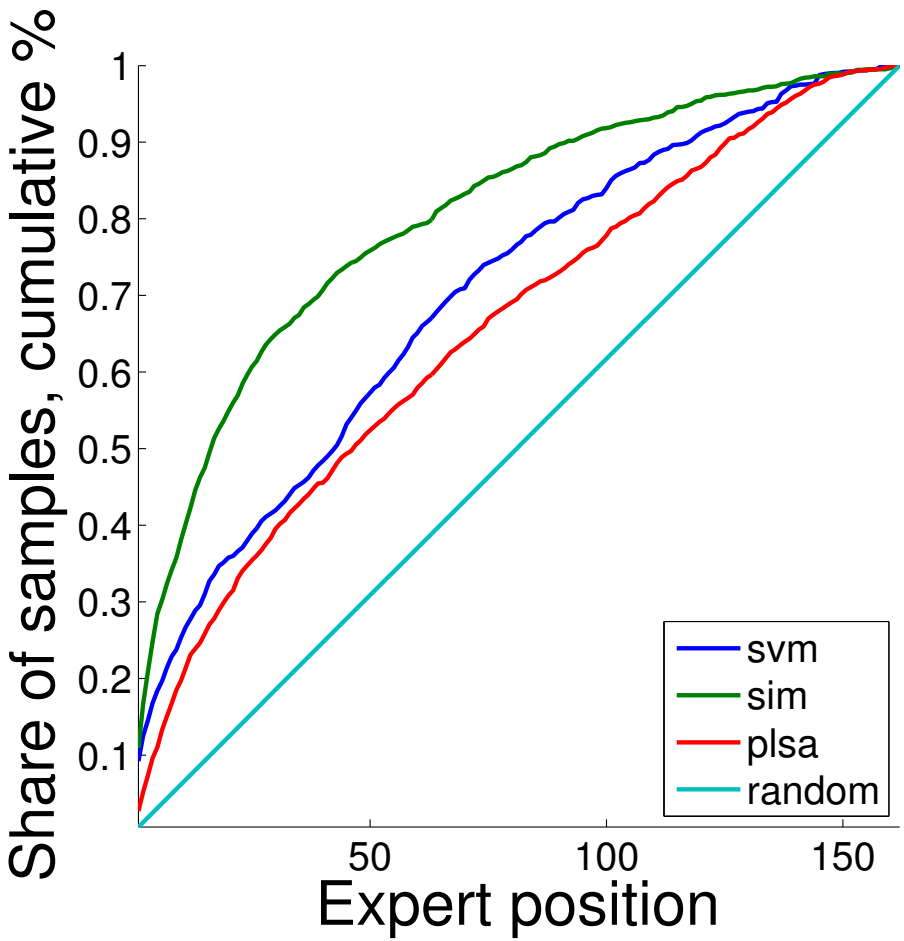


Площадь под гистограммой AUC CH(R): SVM - 0.80, SIM - **0.85**, PLSA - 0.81

Сравнение качества, уровень Stream



Средняя позиция экспертного кластера Q(R): SVM - 50.6, SIM - **32.9**, PLSA - 56.6



Площадь под гистограммой AUC CH(R): SVM - 0.69, SIM - **0.80**, PLSA - 0.65

Реализация

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)

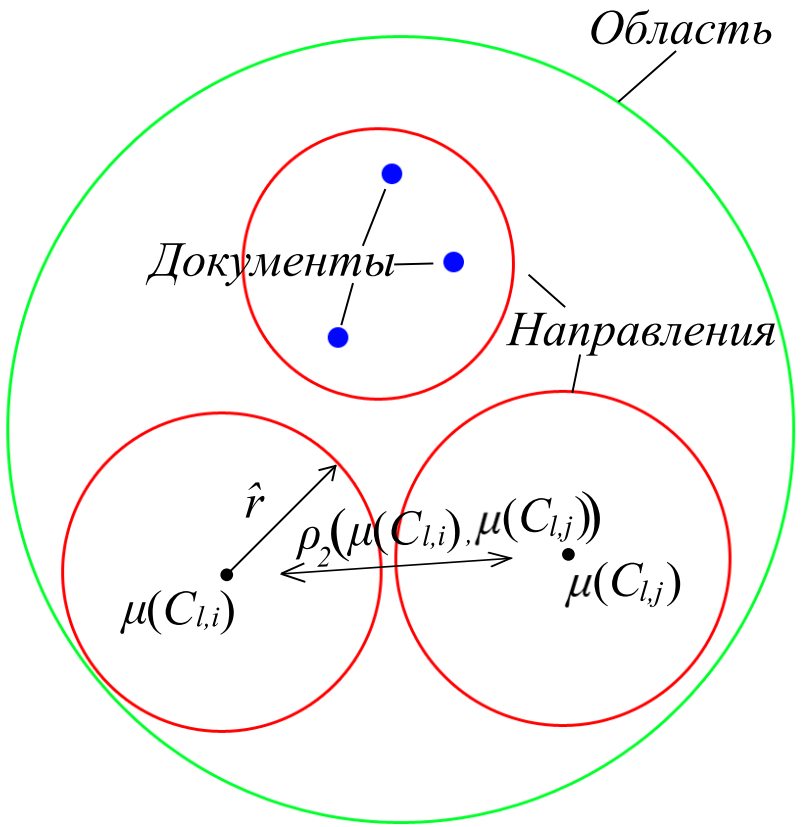
Area: Emerging Applications of OR **Stream:** Models of Embodied Cognition	Select
Area: OR in Health, Life Sciences & Sports **Stream:** Medical Decision Making	Select
Area: Discrete Optimization, Geometry & Graphs **Stream:** Graphs and Networks	Select
Area: Data Science, Business Analytics, Data Mining **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

Вложенная визуализация

Требования к визуализации

1. Вложенность визуализации.
2. Сохранность относительности расстояний.

- $\boldsymbol{\mu}(c_{\ell,i})$ — координаты центра кластера $c_{\ell,i}$.
- $\rho(\cdot,\,\cdot)$ — выбранное расстояние в исходном пространстве.
- $\rho_2(\cdot,\,\cdot)$ — соответствующее ему расстояние на плоскости.



Построение вложенной модели

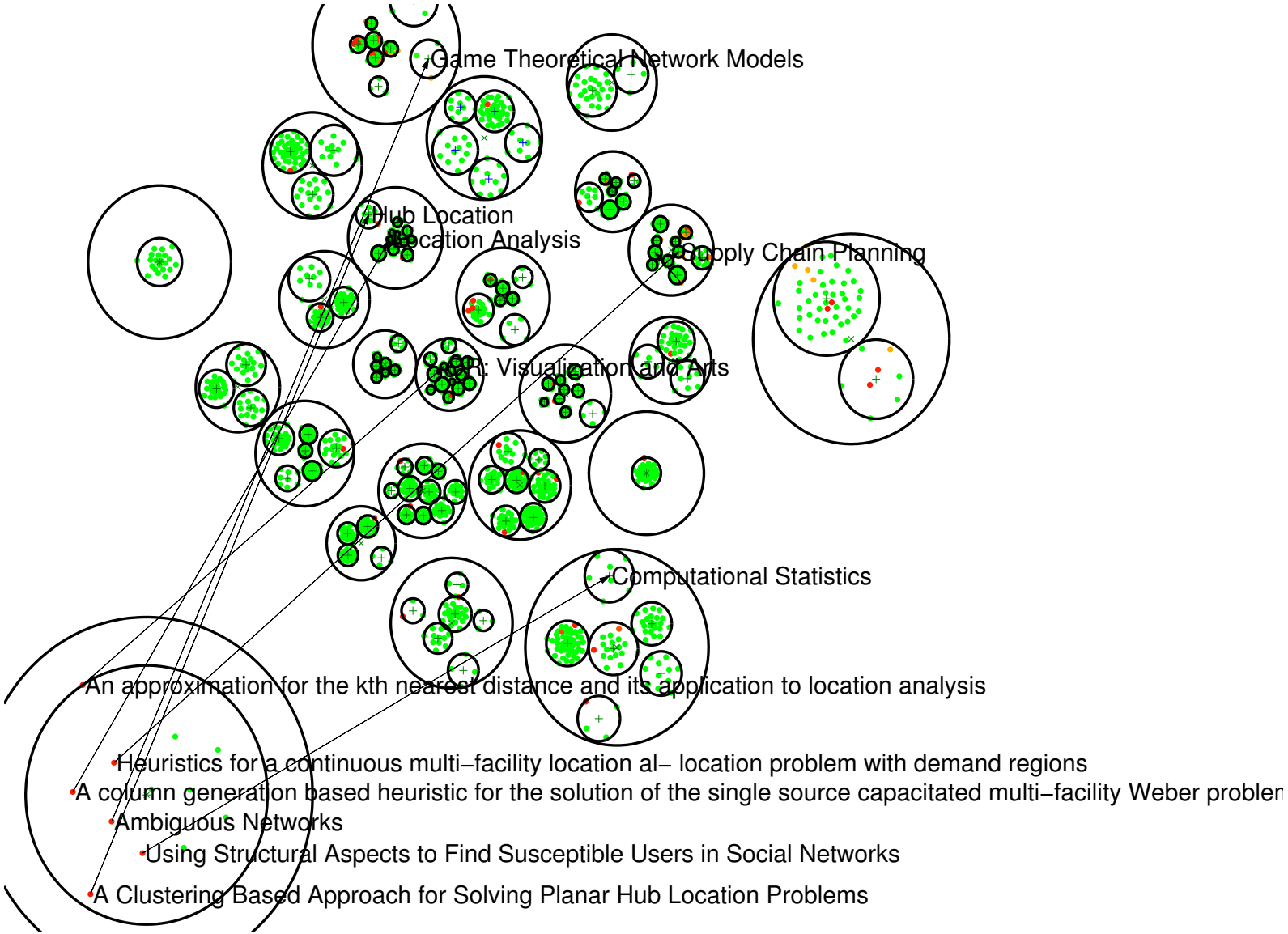
Пусть кластер $c_{\ell,i}$ с радиусом R уже размещен на плоскости;
$C_1, \ldots, C_q$ — кластеры уровня $\ell + 1$ содержащиеся в $c_{\ell,i}$,
$\boldsymbol{\mu}(C_1), \ldots, \boldsymbol{\mu}(C_q)$ — их центры, а $r_1, \ldots, r_q$ — их радиусы.

1. Проецируем на плоскость центры $\boldsymbol{\mu}(C_1), \ldots, \boldsymbol{\mu}(C_q)$ методом проекции Саммона.
2. Определяем радиусы $\hat{r}_1, \ldots, \hat{r}_q$ кластеров $C_1, \ldots, C_q$ по формуле:

$$\hat{r}_j = \min_{i \neq j} \frac{r_j}{r_j + r_i} \rho_2(\boldsymbol{\mu}(C_i), \boldsymbol{\mu}(C_j).$$

3. Находим $\hat{\rho} = \max\limits_{j \in \{1, \ldots, q\}} \rho_2(\boldsymbol{\mu}(C_j), \boldsymbol{\mu}_{\ell,i}) + \hat{r}_j$ — расстояние до границы полученной проекции, учитывающее размеры кластеров.
4. Гомотетия с коэффициентом $\dfrac{R}{\hat{\rho}}$ и центром $\boldsymbol{\mu}(c_{\ell,i})$

Визуализация несоответствий



Заключение

- Оператор, построенный на базе предложенной взвешенной иерархической функции сходства документа и кластера, ранжирует кластеры в порядке убывания релевантности новому документу не хуже известных методов;
- Предложен энтропийный метод оценки важности терминов и метод оптимизации весов $\boldsymbol{\Lambda}$;
- Предложен способ вложенной визуализации кластеров и верификации уже существующей экспертной модели;
- Разработана экспертная система, ранжирующая области и направления по убыванию сходства с заданным документом.