Проверка адекватности тематических моделей коллекции документов

Кузьмин Арсентий

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

Научный руководитель к.ф.-м.н., н.с. ВЦ РАН В. В. Стрижов

Москва,
2013 г.

Цель работы

Задача

Построить тематическую модель конференции и выявить несоответствия с экспертной моделью. Визуализировать экспертную модель и найденые несоответствия.

Методы

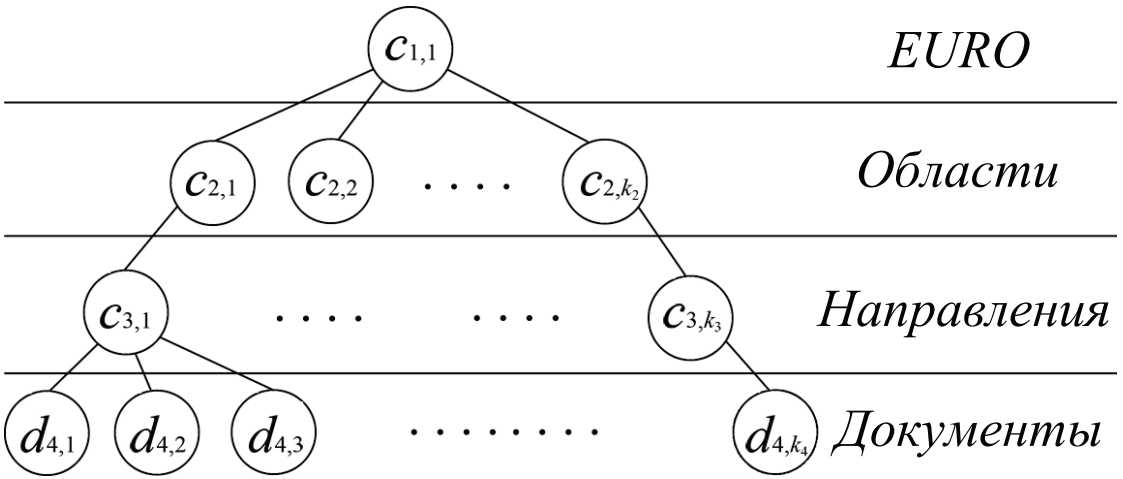
- Неметрическая кластеризация
- Иерархическая визуализация

Требования к алгоритмической модели

- Схожесть с экспертной
- Ранжированность несоответствий
- Сохранность относительности расстояний при визуализации

Структура крупной конференции на примере EURO

Структура:



Построение экспертной тематической модели:

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 (4 доклада).

Возникающие проблемы

Особенности задачи

1. Большое число экспертов (более 200).
2. Субъективность экспертной классификации.
3. Отсутствие эталонной тематической модели.

Возникающие задачи

1. Верификация тематической целостности
2. Выявление нарушений иерархической модели
3. Выявление направлений/сессий, не представляющих интереса со стороны участников конференции.
4. Оценка качества экспертной иерархической модели.

Признаковое описание документа

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

Документ — мешок слов

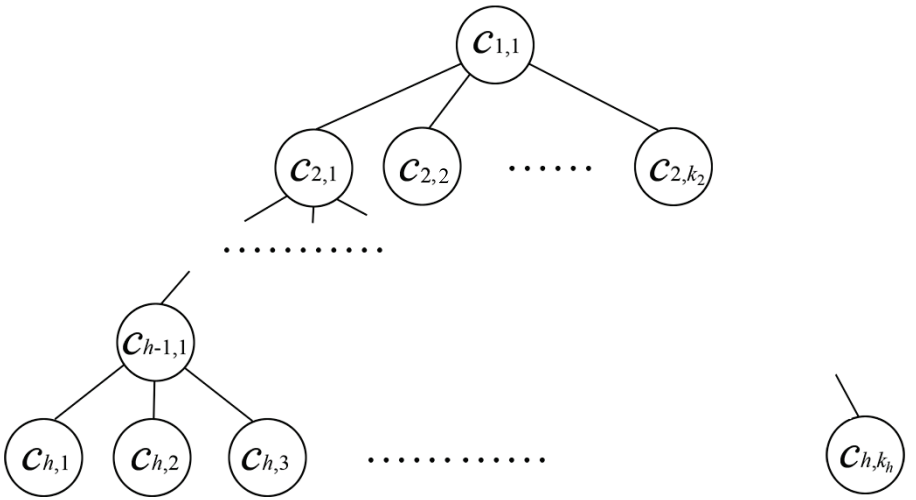
Документом d из коллекции D назовем неупорядоченное множество слов из W, $d = \{w_j\}$, где $j \in \{1, \ldots, n\}$. Будем предполагать, что принадлежность данного документа к какой-либо теме определяется набором терминов, содержащихся в документе.

$$\mathbf{X} = \begin{pmatrix} x_{1,1} & \ldots & x_{1,n} \\ \ldots & \ldots & \ldots \\ x_{|D|,1} & \ldots & x_{|D|,n} \end{pmatrix}.$$

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 документов

Нормируем все вектора признаков документов d_s: $\mathbf{x}_s \mapsto \dfrac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^{\mathsf{T}}\mathbf{x}_s}}$.

Сходство двух документов с учетом нормировки

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j}{\|\mathbf{x}_i\|_2\|\mathbf{x}_j\|_2} = \mathbf{x}_i^{\mathsf{T}}\mathbf{x}_j.$$

Функция сходства кластеров

Под сходством $S(\cdot,\cdot)$ двух кластеров $c_{\ell,i}$ и $c_{\ell,j}$ уровня ℓ будем понимать среднее сходство $s(\mathbf{x},\mathbf{y})$ между документами $\mathbf{x} \in c_{\ell,i}, \mathbf{y} \in c_{\ell,j}$, содержащимися в них.

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

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

Среднее сходство $S(\cdot,\cdot)$ внутри одного кластера для каждого документа d_s определяется как его среднее сходство $s(\cdot,\cdot)$ с документами данного кластера.

Качество кластеризации

$$F_0 = \frac{1}{k_\ell} \sum_{i=1}^{k_\ell} S(c_{\ell,i},\ c_{\ell,i}) \text{ — среднее внутрикластерное сходство.}$$

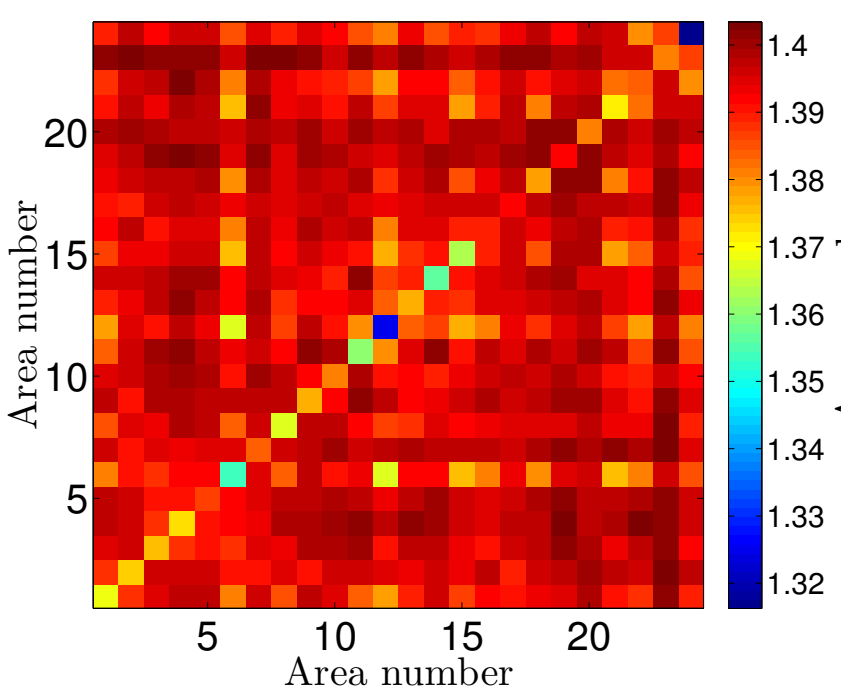
$$F_1 = \frac{2}{k_\ell(k_\ell - 1)} \sum_{i<j} S(c_{\ell,i},\ c_{\ell,j}) \text{ — среднее межкластерное сходство.}$$

Критерий качества кластеризации

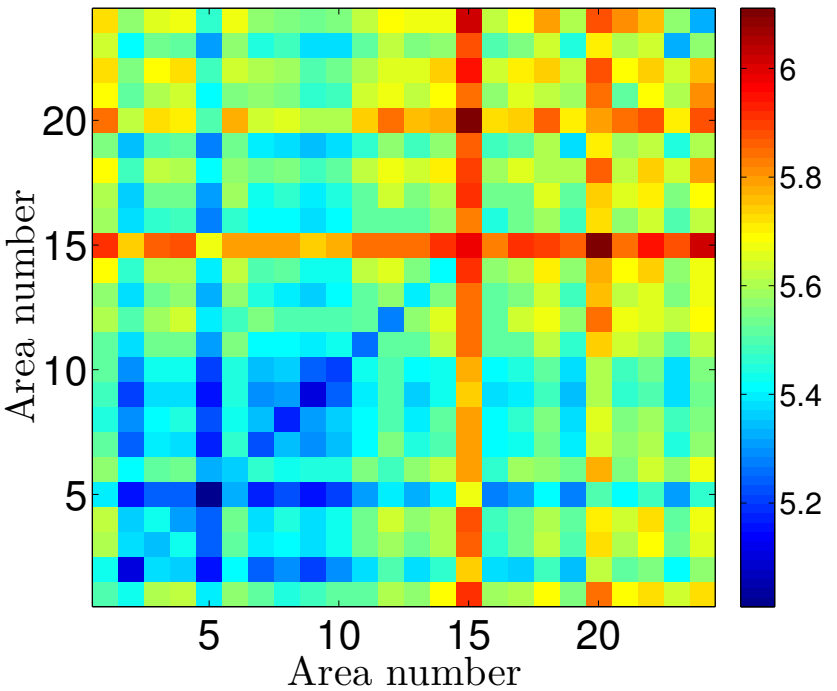
$F = \frac{F_1}{F_0} \to \min.$

Экспертная иерархическая кластеризация документов является базовой при построении тематической модели.

Сравнение различных функций сходства

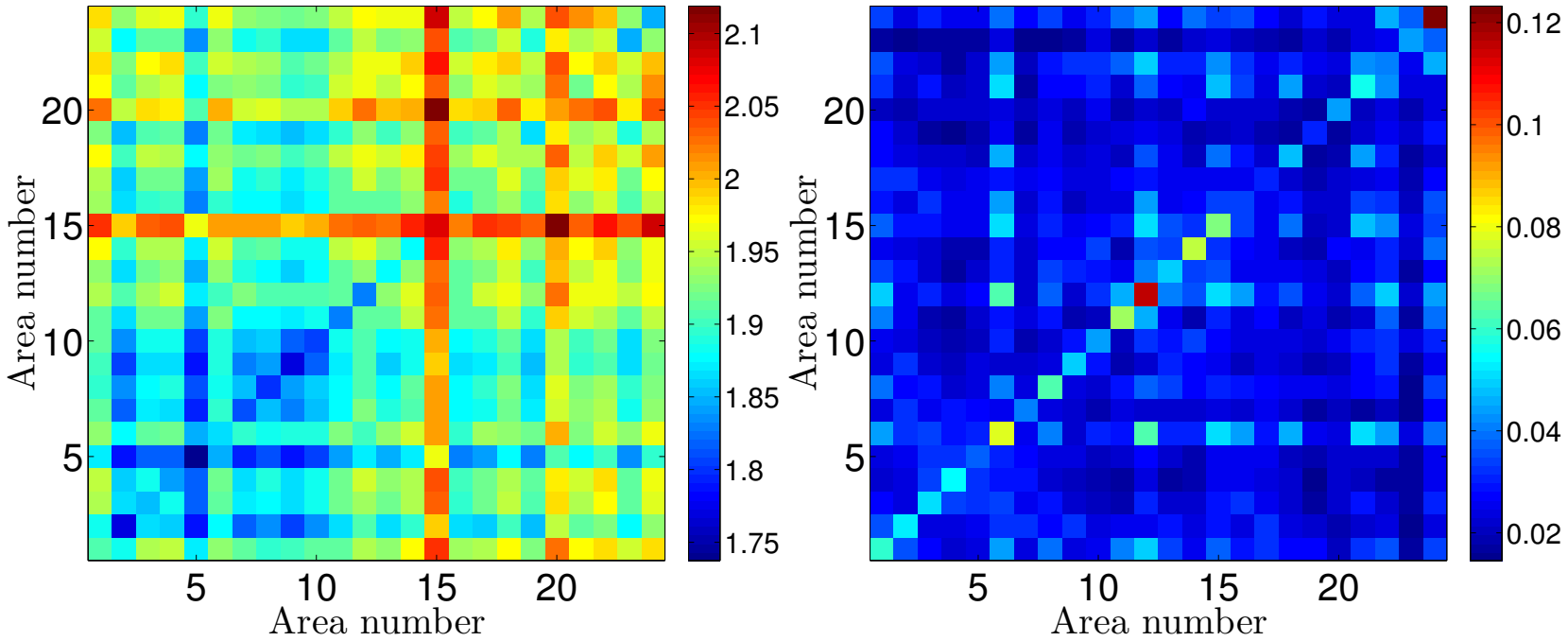


(а) Расстояние Евклида



(б) Расстояние Хеллингера

Сравнение различных функций сходства



(в) Расстояние Дженсона-Шеннона (г) Предложенная функция сходства

Функция качества иерархической модели

Будем использовать комбинацию внутри- и межкластерных сходств следующего вида:

$$Q(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_k) =$$

$$= \sum_{\ell=2}^{h-1} \left[\frac{1-\alpha}{k_\ell} \sum_{i=1}^{k_\ell} |c_{\ell,i}| S(c_{\ell,i}, c_{\ell,i}) - \frac{2\alpha}{k_\ell(k_\ell-1)} \sum_{i<j} S(c_{\ell,i}, c_{\ell,j}) \right] \to \max$$

$\alpha \in [0,1]$ — весовой коэффициент, определяющий приоритет кластеризации.
k_ℓ — общее количество кластеров уровня ℓ.

Построение иерархической модели схожей с экспертной

Введем штрафы на перенос объекта из экспертного кластера.

Таблица: Матрица штрафа

Из \ В	$(+,+)$	$(+,-)$	$(-,-)$
$(+,+)$	δ_{11}	δ_{12}	δ_{13}
$(+,-)$	δ_{21}	δ_{22}	δ_{23}
$(-,-)$	δ_{31}	δ_{32}	δ_{33}

Будем переносить объект, только если будет выполнено условие:

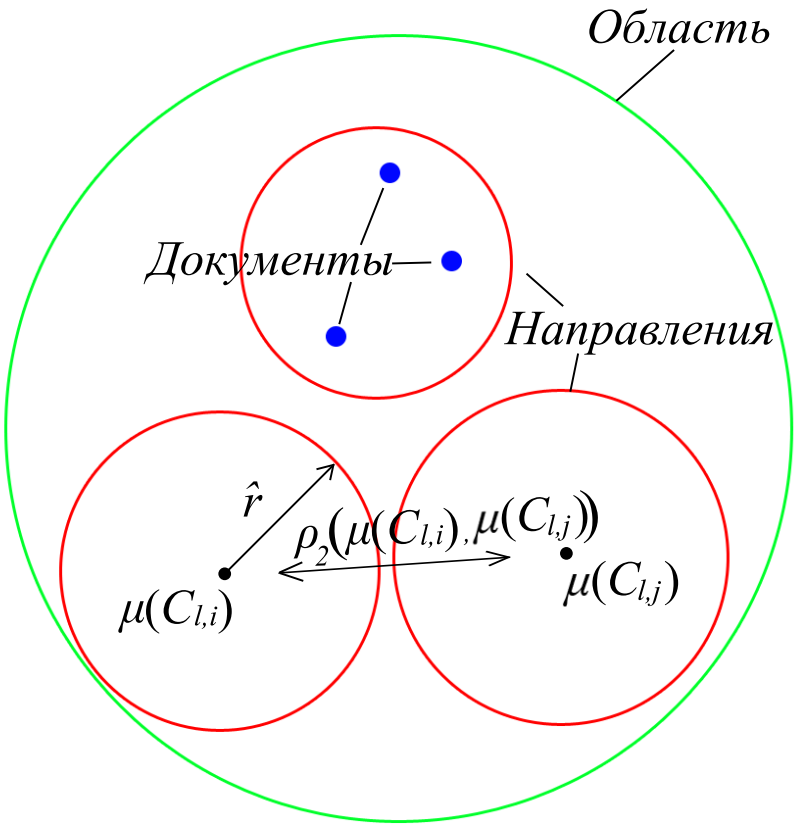
$$Q_2 - Q_1 \geq \delta.$$

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 $\boldsymbol{\mu}(c_{\ell,i})$

Коллекция документов

Цель вычислительного эксперимента

Визуализация иерархической тематической модели EURO.

- Число документов: $|D| = 2313$.
- Размер словаря $|W| = 1063$.
- Штрафы задаются параметром несоответствия с экспертной моделью $\gamma \geq 0$, $\mathbf{F} = \gamma\tilde{\mathbf{F}}$.

Таблица: Матрица $\tilde{\mathbf{F}}$, задающая штраф.

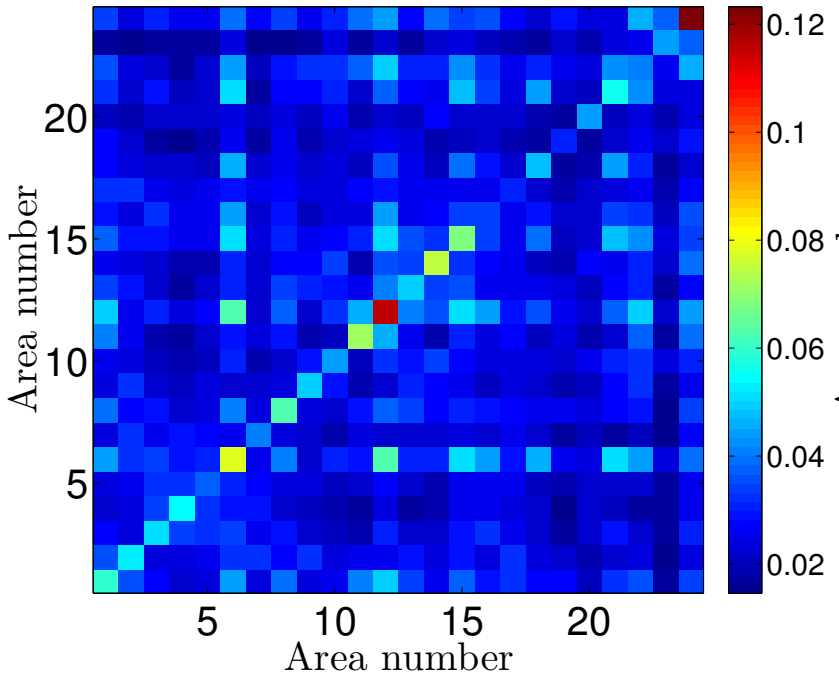
Из \ В	$(+,+)$	$(+,-)$	$(-,-)$
$(+,+)$	0	0.002	0.005
$(+,-)$	-0.001	0	0.003
$(-,-)$	-0.003	-0.002	0

Результаты кластеризации при различных штрафах

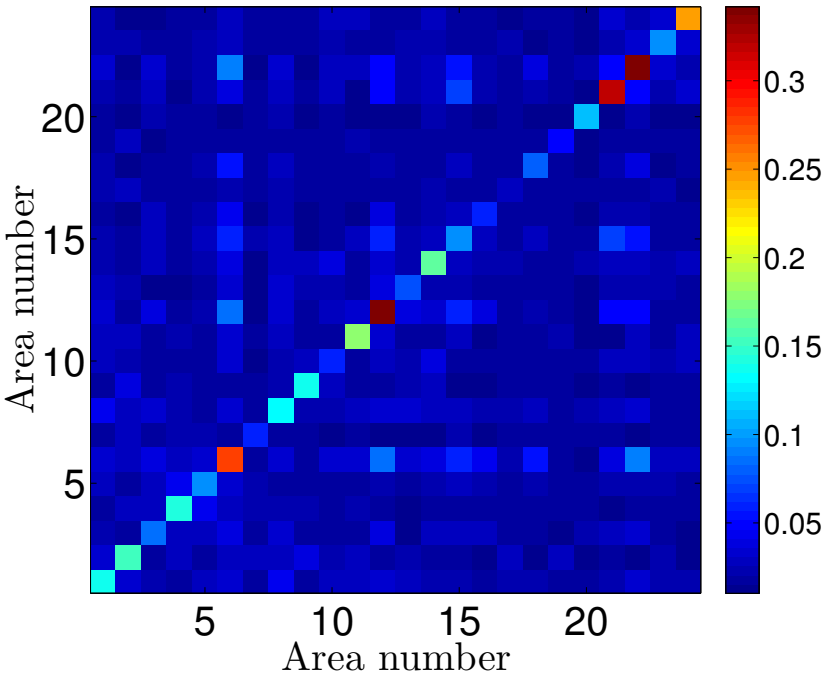
	Среднее внутрикластерное сходство		Среднее межкластерное сходство		Число совпадений	
Вес экспертной модели	Области	Направления	Области	Направления	Области	Направления
Экспертная модель ($\gamma = \infty$)	124.58	197.26	70.08	68.7	2313	2313
Сильно учитывается ($\gamma = 1.25$)	140.66	239.83	69.62	67.54	2252	2212
Средне учитывается ($\gamma = 0.7$)	267.59	512.19	70.08	59.09	1508	1238
Слабо учитывается ($\gamma = 0.5$)	298.33	604.4	69.85	60.41	1170	861

При уменьшении штрафов γ качество модели растет, но число совпадений с экспертной моделью уменьшается.

Сравнение средней близости кластеров по областям, $\gamma = 0.7$

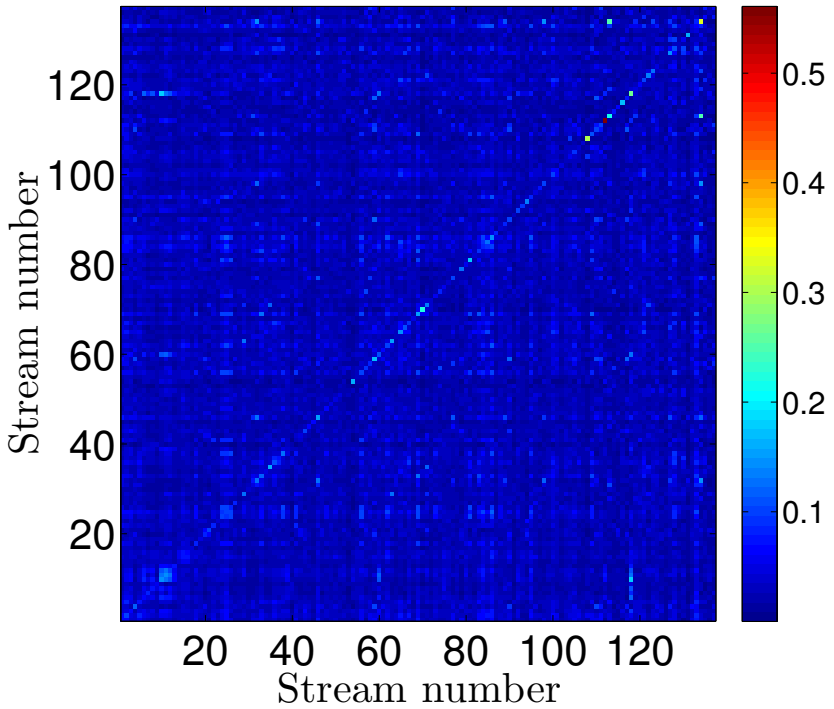


(д) Экспертная кластеризация

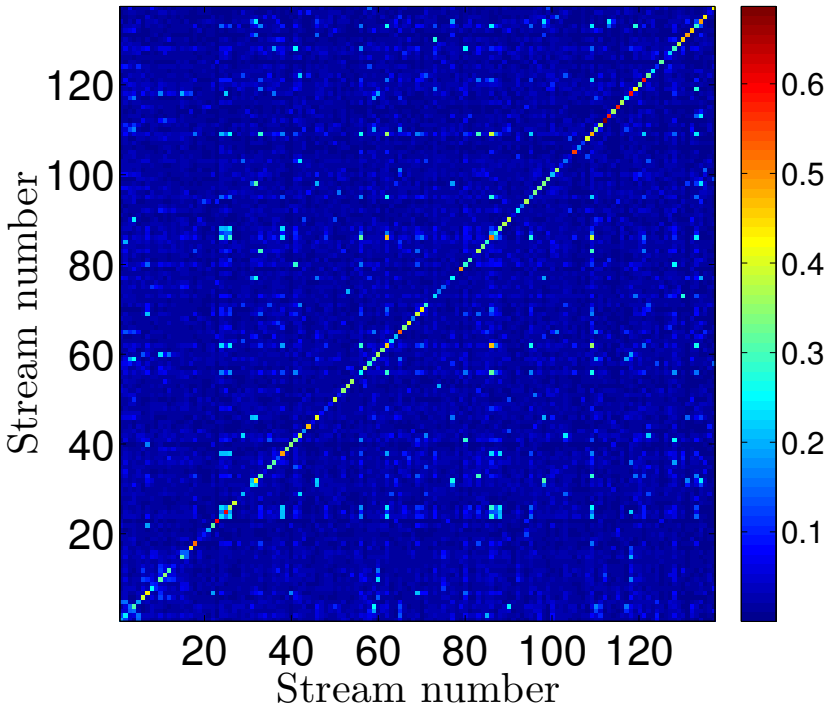


(е) Построенная кластеризация

Сравнение средней близости кластеров по направлениям, $\gamma = 0.7$

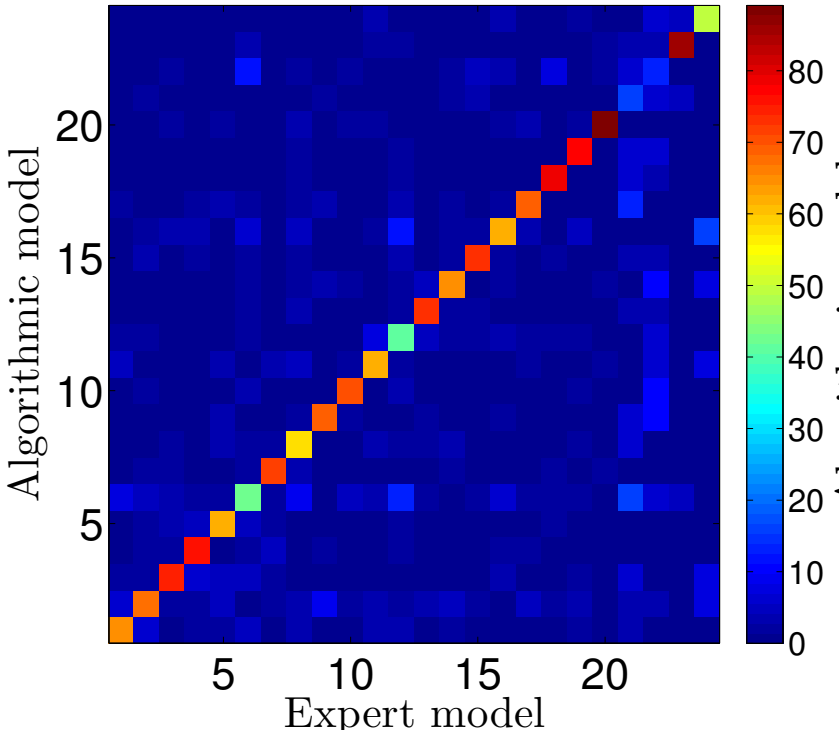


(ж) Экспертная кластеризация

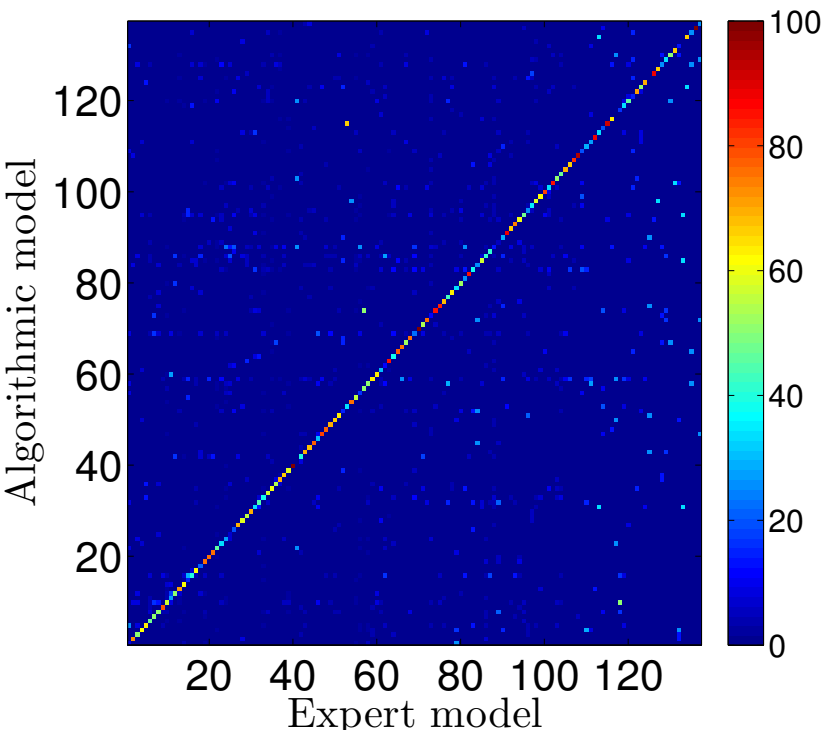


(з) Построенная кластеризация

Процентное распределение документов по областям и направлениям. $\gamma = 0.7$



(и) Распределение по областям



(к) Распределение по направлениям

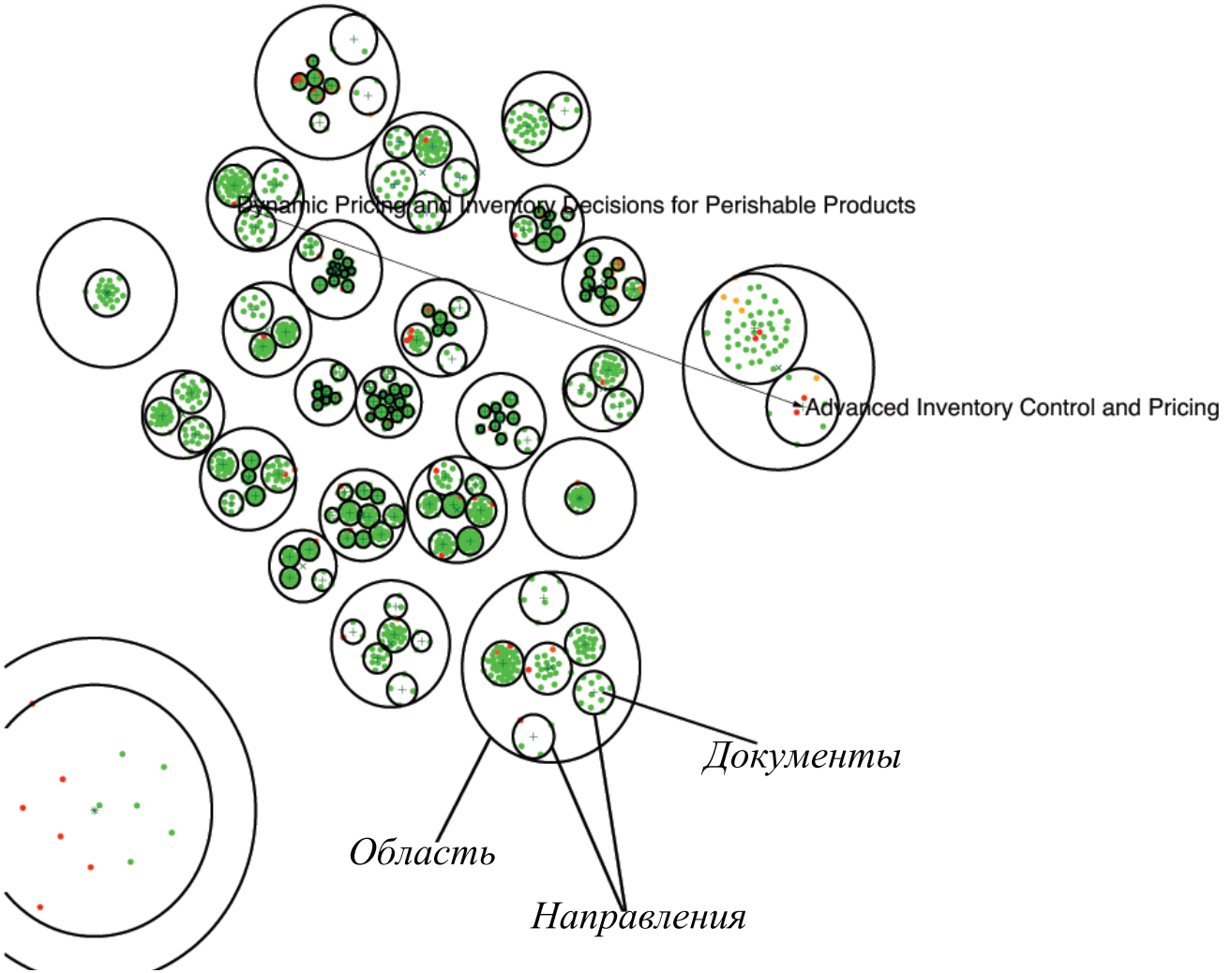
Степень несоответствия

Степень несоответствия построенной кластеризации и экспертной для докуменета определяется

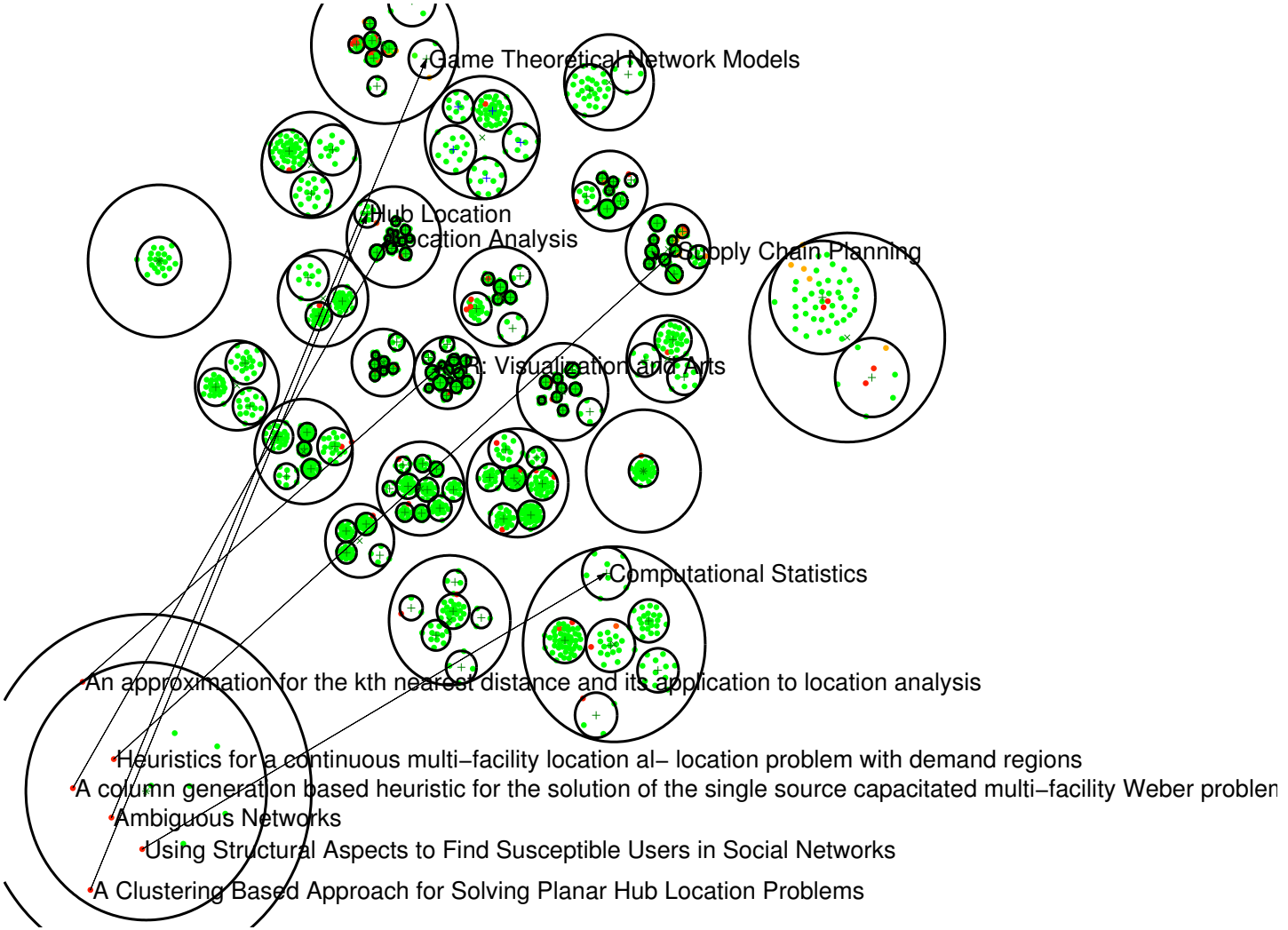
1. Количеством уровней, в которых кластеризации отличаются
2. Значимостью уровня
3. Расстоянием между экспертным и алгоритмическим кластерами

Полученная степень несоответствия отображается в цветовую шкалу $[RGB(255, 0, 0)RGB(0, 255, 0)]$.

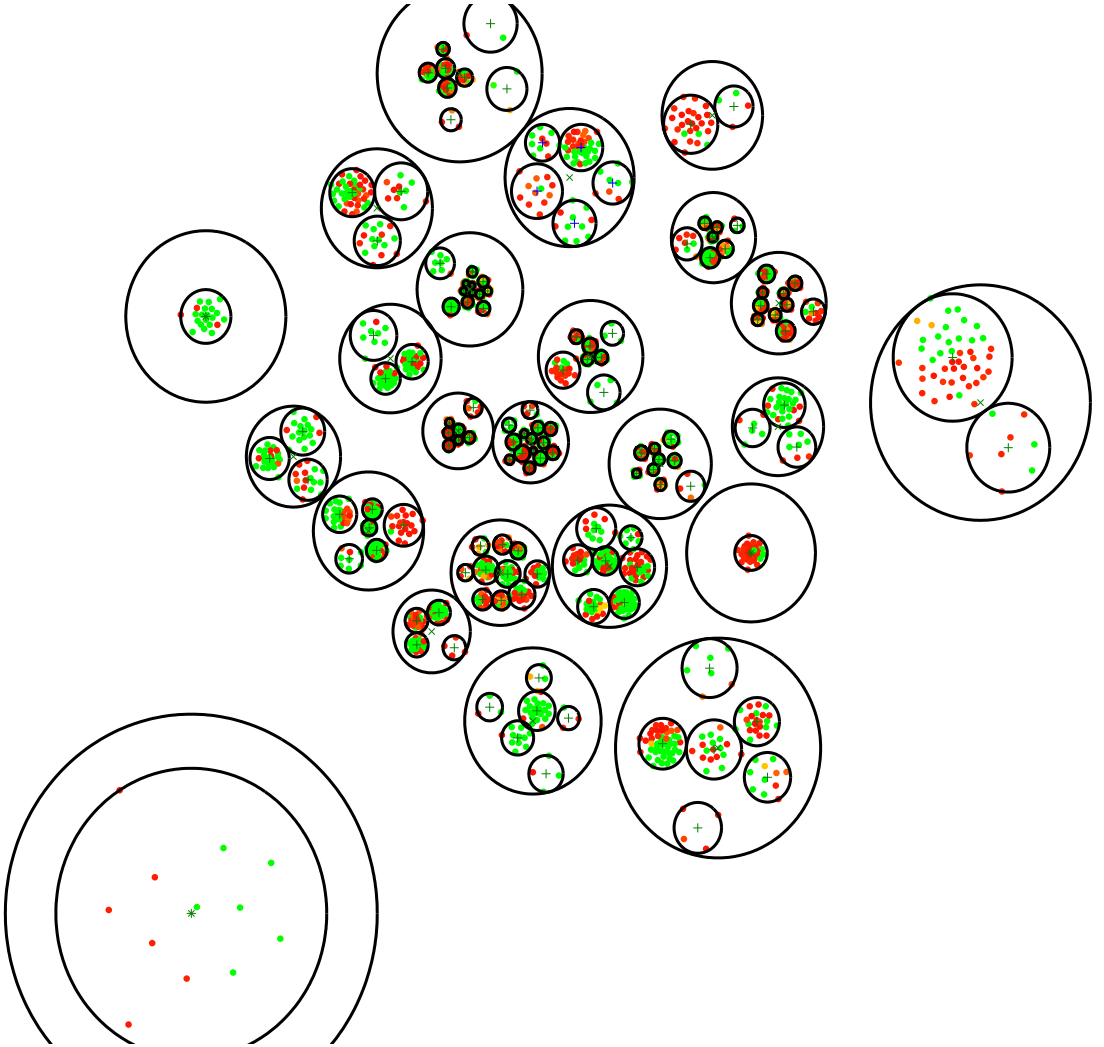
Визуализация несоответствий, $\gamma = 1.25$



Визуализация несоответствий, $\gamma = 1.25$



Визуализация несоответствий, $\gamma = 0.5$



Публикации по теме

- Кузьмин А. А., Стрижов В. В. Многоуровневая классификация при обнаружении движения цен // Машинное обучение и анализ данных. — 2012. — № 3. — С. 318-327.
- Кузьмин А. А., Адуенко А. А., Стрижов В. В. Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ.. — 2012. — № 3. — С. 119-131.
- Кузьмин А. А., Стрижов В. В. Проверка адекватности тематических моделей коллекции документов. // Программная инженерия, 2013, 4 — 16-20.

Заключение

- Предложена функция близости документов и кластеров.
- Предложен способ построения алгоритмической иерархической тематической модели, учитывающий существующую экспертную модель с заданым весом.
- Предложен способ визуализации иерархической модели, состоящей из большого числа документов, на плоскости.
- Предложен способ визуализации степени несоответствия экспертной кластеризации для документа с алгоритмической.