Московский физико-технический институт
(Государственный университет)

Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 974 ГРУППЫ

«Проверка адекватности тематических моделей коллекции документов»

Выполнил:
студент 4 курса 974 группы
Кузьмин Арсентий Александрович

Научный руководитель:
к.ф-м.н., н.с. ВЦ РАН
Стрижов Вадим Викторович

Москва, 2013

Содержание

Аннотация

Работа посвящена верификации иерархической тематической модели коллекции текстов, построенной экспертами. Рассматривается коллекция документов с экспертной тематической моделью. Для проверки адекватности экспертной модели предлагается построить алгоритмическую модель путем иерархической кластеризации коллекции текстов. Используется функция сходства между документами. Определяется степень несоответствия экспертной модели и предлагаемой. Предлагается метод визуализации построенной иерархической модели. Алгоритм построения тематической модели проиллюстрирован кластеризацией коллекции тезисов конференции EURO 2012.

Ключевые слова: коллекция документов, тематические модели, иерархические модели, кластеризация, иерархическая визуализация.

1 Введение

Перед программным коммитетом конференции с большим числом участников встает задача проверки корректности построения иерархической модели тезисов конференции. Рассмотрим процесс построения такой модели на примере конференции EURO 2012. Конференция содержит в себе 26 главных тем (далее область), определяемых председателем программного комитета. Каждая главная тема содержит в себе примерно 10 больших подтем (далее направление), каждая из которых делится на сессии (далее сессия), содержащие в себе ровно 4 документа. К каждой из главных тем приекрепляется группа экспертов в соответствующей области. Для подачи заявки авторы отправляют аннотацию (далее документ) к своей работе, состоящую из не более чем 600 символов, и выбирают из предложенного им списка 3 ключевых слова, наиболее сильно коррелирующих по их мнению с тематикой их работы. Все присланные заявки хранятся в общей базе. После окончания срока подачи заявок, эксперты начинают отбирать присланные документы из общей базы в свои темы, основываясь на ключевых словах, указанных авторами, и содержании документов. Затем эксперты распределяют отобранные ими документы по иерархической структуре их главной темы, исходя из своей стратегии организации докладов конференции.

В силу большого числа экспертов, субъективности экспертной кластеризации и отсутствия эталонной модели, оценить качество экспертной иерархической модели сложно. Анализ экспертно заданной тематической модели конференции необходим для решения следующих организационных задач:

1. Выявление дубликатов областей, направлений, сессий.
2. Выявление взаимосвязей между прикладными и теоретическими работами в фиксированной области.
3. Выявление нарушений иерархической модели: использокание элемента верхнего уровня иерархии в качестве нижнего и наоборот.

4. Выявление направлений/сессий, не представляющих интереса со стороны исследователей (научного сообщества).

5. Выявление наборов документов, для которых необходимо создать новую область/направление.

Поэтому предлагается построить иерархическую модель коллекции тезисов, основыванную на их терминологическом сходстве.

Кластеризация текстов при построении каждого уровня иерархической модели может проводиться с помощью метрических алгоритмов кластеризации, например, K-means [7], FOREL [8], C-means [9], STOLP [10], FRiS-STOLP [11], BoostML, DANN [12] и другие [13, 14], и с помощью вероятностных методов [16], например с помощью вероятностного латентного семантического анализа (англ. PLSA — probabilistic latent semantic analysis) [5], или латентного размещения Дирихле (англ. LDA — latent Dirichlet allocation) [2].

Для построения иерархической кластерной модели существует два основных типа алгоритмов — дивизимные и агломеративные [3].

Для кластеризации множества документов, необходимо ввести способ оценки степени сходства между двумя документами. Чтобы оставить в документе только те слова, которые несут информацию о его сходстве и отличии от других документов, предварительно проводится предобработка документов. Слова приводятся к начальной лексической форме [4], удаляются знаки препинания. Часто для отсева неинформативных слов, встречающихся малое количество раз, а также слов, встречающиеся в большинстве документов используется критерий $\mathrm{tf} \cdot \mathrm{idf}$ (англ. tf — term frequency, idf — inverse document frequency) [6], а также словарь стоп-слов. Но для повышения качества работы алгоритма кластеризации в данной работе словарь терминов составляется экспертно.

После отсева неинформативных слов из словаря и документов, документы представляются в виде «мешков слов» [5] и каждому документу ставится в соответствие булевый вектор. В работе [4] сравниваются способы построения вектора — описания документа.

В данной работе для построения тематической модели используется неметрический алгоритм кластеризации, перераспределяющий объекты между кластерами так, чтобы улучшить среднее сходство между объектами из одного кластера и увелиить среднее различие между объектами из разных кластеров. Требуется построить иерархическую модель, сохранив ее схожесть с экспертной, поэтому при кластеризации учитывается модель, предложенная экспертами, с определенным весом. При построении иерархической модели проводится «жесткая» кластеризация, согласно которой объект принадлежит только к одному из кластеров, так как каждый документ может принадлежать только одной теме, что соответствует правилам проведения конференции.

Для анализа предлагается методика, позволяющая визуализировать иерархическую модель на плоскости. Эта методика учитывает как экспертную, так и алгоритми-

ческую модели. Она также позволяет выявлять и составлять ранжированный список тезисов, которые необходимо переместить в другие направлени/сессии/области.

Основными требованиями к визуализации являются: сохранение связи между схожестью элементов иерархической структуры одного уровня и расстоянием между ними на плоскости, наглядное отображение несоответствий в экспертной модели и возможность посмотреть варианты их устранения.

После предобработки документов и построения модели все элементы иерархической структуры представлены булевыми векторами в многомерном пространстве. Существует множество способов понижения размерности пространства. Целенаправленное проецирование, многомерное шкалирование, методы главных многообразий, методы топологически непрерывных отображений и нейросетевые методы [17–19]. В данной работе внимание уделяется методам, позволяющим отобразить матрицу парных расстояний между документами на плоскость, для наглядной визуализации тематической схожести и разлиячия. Одним из наиболее распространенных методов для этого являются самоорганизующиеся карты Кохонена [20, 21], однако в данной работе особый интерес представляет не отображение всех совокупностей документов, а визуализация относительного расположения документов и центров иерархической структуры. Для этого используется проекция Саммона [22, 23]. Основным критерием качества работы данного алгоритма является не столько сохранение расстояний между объектами по абсолютной величине, сколько сохранение их относительности. Еще одной особенностью поставленной задачи является иерархичность. Для иерархической визуализации можно использовать как обычные методы [24], так и методы, явно отображающие древовидную структуру полученной модели [25, 26].

Таким образом, основными задачами, рассматриваемыми в данной статье являются:

1. Построение иерархии неметрическим методом, учитывающим существующую экспертную модель.

2. Визуализация экспертной модели на плоскости.

3. Отображение выявленных некорректностей на шаге 1. на экспертной модели, построенной на шаге 2.

2 Построение алгоритмической модели

2.1 Постановка задачи верификации тематической модели

Пусть $W = \{w_1, \ldots, w_n\}$ — заданное множество слов (словарь), где n — количество слов в словаре. Документом d из коллекции D назовем неупорядоченное множество слов из $W, d = \{w_j\}$, где $j \in \{1, \ldots, n\}$.

Поставим в соответствие каждому документу d его описание — вектор $\mathbf{x}$ размерности n следующим образом: если слово w_j из словаря W встретилось в документе d_s

k раз, то $x_{s,j} = k$, $k \geq 0$. Получим матрицу $\mathbf{X}$ «объект-признак», где каждая строка $\mathbf{x}_s = [x_{s,1}, \ldots, x_{s,n}]$ — признаковое описание документа d_s.

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

Для удобства дальнейшего изложения нормируем все строки полученной матрицы $\mathbf{X}$ следующим образом:

$$\mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^{\mathsf{T}} \mathbf{x}_s}}. \tag{2}$$

Представим иерархическую тематическую модель в виде дерева (см. рис. 1). Глубину дерева обозначим h. Уровнем l иерархии назовем множество всех узлов дерева, находящихся на глубине l. Документы $d_s \in D$ являются листами этого дерева и имеют уровень h. Кластером c будем называть подмножество коллекции документов D. Сопоставим каждому узлу i уровня l дерева, эти два индекса объединим в пару (l, i), кластер $c_{l,i}$, состоящий из документов d_s, путь до которых от вершины $c_{1,1}$ проходит через узел (l, i).

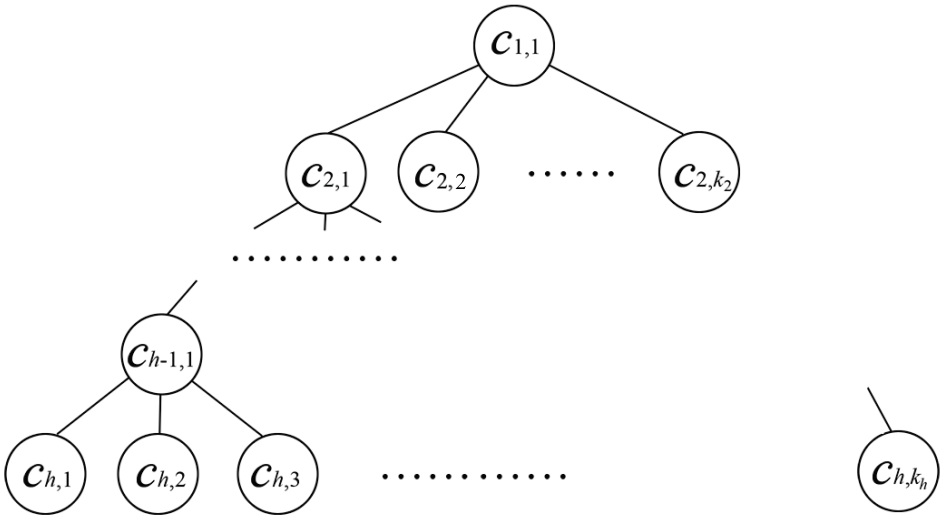


Рис. 1: Иерархическое представление тематической модели.

2.2 Функция сходства документов

Предполагается, что каждый документ в коллекции может быть описан небольшим набором признаков — ключевых слов. В рассматриваемой в данной работе коллек-

ции [29] каждый документ описывается 10-15 признаками. При этом словарь состоит из более 1000 слов.

Предлагается ввести функцию сходства $s(\mathbf{x}_i,\ \mathbf{x}_j)$ двух документов:

$$s(\mathbf{x}_i,\ \mathbf{x}_j) = \frac{\mathbf{x}_i^\intercal \mathbf{x}_j}{\|\mathbf{x}_i\|_2 \|\mathbf{x}_j\|_2} = \mathbf{x}_i^\intercal \mathbf{x}_j. \tag{3}$$

В (3) учтена нормировка (2), позволяющая документам $\mathbf{x}_i$ и $\mathbf{x}_j$ иметь разную длину в словах при сравнении. Так как все компоненты векторов $\mathbf{x}_i$, $\mathbf{x}_j$ неотрицательны, то $s(\mathbf{x}_i, \mathbf{x}_j) \in [0, 1]$, причем $s = 1$ достигается для документов, словарный состав которых одинаков.

Введем функцию сходства $S(c_{l,i}, c_{l,j})$ двух кластеров $c_{l,i}$ и $c_{l,j}$ уровня l будем понимать среднее сходство $s(\mathbf{x}, \mathbf{y})$ между документами $\mathbf{x} \in c_{l,i}, \mathbf{y} \in c_{l,j}$, содержащимися в них (4). Среднее сходство $S(\cdot, \cdot)$ внутри одного кластера для каждого документа d_s определяется как среднее сходство $s(\cdot, \cdot)$ с остальными документами данного кластера.

$$S(c_{l,i},\ c_{l,j}) = \frac{1}{|A|} \sum_{(\mathbf{x},\ \mathbf{y}) \in A} s(\mathbf{x},\ \mathbf{y}), \tag{4}$$

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

Сравним введенную функцию сходства (3) с Евклидовым расстоянием (5), расстоянием Хеллингера (6) и расстоянием Дженсона-Шеннона (7).

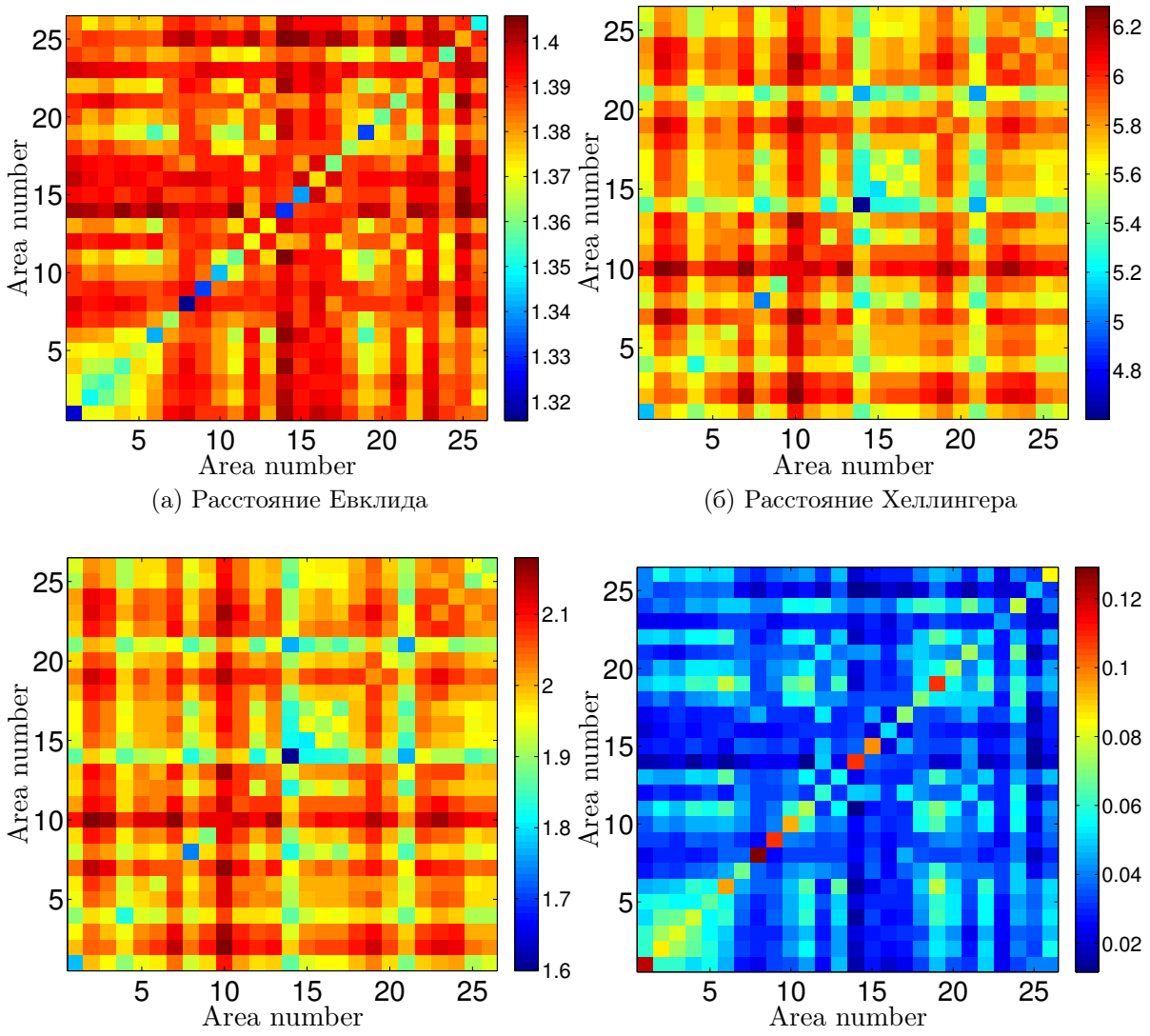
$$\rho(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{k=1}^{n} (x_k - y_k)^2} \tag{5}$$

$$H(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{2}} ||\sqrt{\mathbf{x}} - \sqrt{\mathbf{y}}||_2 \tag{6}$$

$$JSD(\mathbf{x}||\mathbf{y}) = \frac{1}{2} D(\mathbf{x}||M) + \frac{1}{2} D(\mathbf{y}||M),\ M = \frac{1}{2}(\mathbf{x} + \mathbf{y}),\ \text{где}$$
$$D(\mathbf{x}||\mathbf{y}) = \sum_i \ln\left(\frac{x_i}{y_i}\right) x_i \ \text{— расстояние Кульбака Лейблера} \tag{7}$$

На рис. 2 а)-в) приведены значения средних расстояний между документами разных областей для указанных трех типов расстояния для экспертной кластеризации. Рис. 2 г) соответствует введенной функции сходства. По осям отложены номера областей, цвет элемента (x, y) соответствует среднему расстоянию между документами области с номером x и области с номером y. Таким образом элементы диагонали (x, x) соответствуют внутрикластерному расстоянию, а элементы $(x, y),\ x \neq y$ — межкластерному. В случае г) под средним расстоянием понимается сходство двух кластеров.

Показателем качества кластеризации являются большие межкластерные расстояния по сравнению с внутрикластерными, что соответствует пику на диагонали. Можно заметить, что отличия в среднем внутрикластерном расстоянии и среднем межкластерном расстоянии для а)-в) почти отсутствуют, а потому рассчитывать на хороший результат при применении метрических алгоритмов кластеризации вряд ли возможно. Значительные отличия между сходством или расстоянием внутри одной области и между разными областями наблюдается только на рис. 2 г), что подтверждает целесообразность использования введенной функции сходства (4).



(а) Расстояние Евклида

(б) Расстояние Хеллингера

(в) Расстояние Дженсона-Шеннона

(г) Введенная функция сходства

Рис. 2: Значения расстояния а)-в) и функции сходства г) для разных областей в экспертной модели.

Определим $\bar{\mathbf{x}}_i$ как средний вектор в кластере $c_{l,i}$

$$\bar{\mathbf{x}}_i = \frac{1}{|c_{l,i}|} \sum_{\mathbf{x} \in c_{l,i}} \mathbf{x}. \tag{8}$$

Тогда в соответствии с (4) и (3) при $i \neq j$

$$S(c_{l,i},\ c_{l,j}) \ = \ \frac{1}{|c_{l,i}||c_{l,j}|} \sum_{\mathbf{x} \in c_{l,i}} \sum_{\mathbf{y} \in c_{l,j}} \mathbf{x}^\intercal \mathbf{y} \ = \ \left(\frac{1}{|c_{l,i}|} \sum_{\mathbf{x} \in c_{l,j}} \mathbf{x} \right)^\intercal \left(\frac{1}{|c_{l,j}|} \sum_{\mathbf{y} \in c_{l,i}} \mathbf{y} \right) \ = \ \bar{\mathbf{x}}_i^\intercal \bar{\mathbf{x}}_j.$$

Аналогично

$$S(c_{l,i},\ c_{l,i}) = \frac{1}{|c_{l,i}|} \sum_{\mathbf{x} \in c_{l,i}} \frac{1}{|c_{l,i}| - 1} \sum_{\mathbf{y} \neq \mathbf{x},\, \mathbf{y} \in c_{l,i}} \mathbf{x}^\intercal \mathbf{y} = \frac{1}{|c_{l,i}|} \sum_{\mathbf{x} \in c_{l,i}} \frac{1}{|c_{l,i}| - 1} \mathbf{x}^\intercal (|c_{l,i}| \bar{\mathbf{x}}_i - \mathbf{x}) =$$

$$= \frac{1}{|c_{l,i}|} \sum_{\mathbf{x} \in c_{l,i}} \frac{|c_{l,i}|}{|c_{l,i}| - 1} \mathbf{x}^\intercal \bar{\mathbf{x}}_i - \frac{1}{|c_{l,i}| - 1} = \frac{|c_{l,i}|}{|c_{l,i}| - 1} \bar{\mathbf{x}}_i^\intercal \bar{\mathbf{x}}_i - \frac{1}{|c_{l,i}| - 1}.$$

В последнем выражении учтена нормировка $\mathbf{x}^\intercal \mathbf{x} = 1$. Таким образом, и сходство между кластерами, и внутри кластеров определяются только средними векторами кластеров, что позволяет их эффективно считать и пересчитывать при изменении состава кластеров. Введем далее функционал качества кластеризации и опишем алгоритм.

2.3 Функционал качества кластеризации документов

В качестве функционала качества кластеризации будем использовать комбинацию внутри- и межкластерных сходств следующего вида

$$Q(\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_k) = \sum_{l=2}^{h-1} \left[\frac{1 - \alpha}{k_l} \sum_{i=1}^{k_l} |c_{l,i}| S(c_{l,i},\ c_{l,i}) - \alpha \frac{2}{k_l(k_l - 1)} \sum_{i<j} S(c_{l,i},\ c_{l,j}) \right] \to \max, \tag{9}$$

где $\alpha \in [0, 1]$ – весовой коэффициент, отвечающий за приоритет при максимизации, а k — общее количество кластеров уровня l. При $\alpha \ \to \ 0$ алгоритм будет максимизировать внутрикластерное сходство, вне зависимости от межкластерного, и наоборот при $\alpha \ \to \ 1$. Весовой множитель $|c_{l,i}|$ позволяет считать среднее внутрикластерное сходство не по кластерам, а по документам. Если считать среднее внутрикластерное сходство по кластерам, то возникает один кластер, собирающий множество документов, мало сходных друг с другом. Остальные кластеры малочисленны и обладают высоким внутрикластерным сходством. При усреднении по документам эта особенность исчезает.

2.4 Алгоритм кластеризации

В качестве начального приближения кластеризации распределим документы d_s по кластерам $c_{l,i}$ согласно экспертной модели. Затем на каждом шаге алгоритма по очереди выбираем по одному документу $\mathbf{x} \in c_{l,i}$ из коллекции D и переносим его в другой кластер так, чтобы значение функционала качества Q определенного как (9) возросло. Пусть для этого его требуется перенести в кластер $c_{l,j}$ и это перенесении дает максимальный рост функционала качества. Заметим, что из всех членов в сумме для Q изменяются только $S(c_{l,i}, c_{l,i})$, $S(c_{l,i}, c_{l,j})$, $S(c_{l,j}, c_{l,j})$. Новые средние векторы кластеров $c_{l,i}$ и $c_{l,j}$ определяются по $\mathbf{x}$ и старым средним векторам как

$$\bar{\mathbf{x}}_i \to \frac{|c_{l,i}|}{|c_{l,i}|-1}\bar{\mathbf{x}}_i - \frac{1}{|c_{l,i}|-1}\mathbf{x},$$
$$\bar{\mathbf{x}}_j \to \frac{|c_{l,j}|}{|c_{l,j}|+1}\bar{\mathbf{x}}_i + \frac{1}{|c_{l,j}|+1}\mathbf{x}.$$

Далее вычисляем изменение функционала качества Q. Находя тот кластер $c_{l,j}$, в который перенесение $\mathbf{x}$ дает наибольший эффект, переносим $\mathbf{x}$ в $c_{l,j}$, если есть улучшение. Повторяем эти шаги пока кластеризация не стабилизируется в терминах Q (9).

3 Верификация и визуализация экспертной модели

3.1 Верификация тематической модели

Для верификации тематической модели получим кластеризацию, сходную с экспертной. Для этого модифицируем алгоритм кластеризации, описанный выше для учета экспертной модели.

Рассмотрим операцию перенесения объекта $\mathbf{x}$ из одного направления в другое. Пусть экспертные область и направление для объекта $\mathbf{x}$ $C_1(\mathbf{x})$ и $C_2(\mathbf{x})$ соответственно. На очередном шаге алгоритма объект $\mathbf{x}$ находится в области $\hat{C}_1(\mathbf{x})$ и направлении $\hat{C}_2(\mathbf{x})$ и рассматривается его перенос из направления $\hat{C}_2(\mathbf{x})$ в направление $\tilde{C}_2(\mathbf{x})$, которое находится в области $\tilde{C}_1(\mathbf{x})$. В силу вложенности направления в область возможны 9 разных ситуаций относительно изменения отличий экспертной и алгоритмической кластеризации при переносе объекта $\mathbf{x}$

- $\hat{C}_1(\mathbf{x}) = C_1(\mathbf{x}) = \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) = C_2(\mathbf{x}) = \tilde{C}_2(\mathbf{x})$: $(+,\ +) \mapsto (+,\ +)$,
- $\hat{C}_1(\mathbf{x}) = C_1(\mathbf{x}) = \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) = C_2(\mathbf{x}) \neq \tilde{C}_2(\mathbf{x})$: $(+,\ +) \mapsto (+,\ -)$,
- $\hat{C}_1(\mathbf{x}) = C_1(\mathbf{x}) = \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) \neq C_2(\mathbf{x}) = \tilde{C}_2(\mathbf{x})$: $(+,\ -) \mapsto (+,\ +)$,
- $\hat{C}_1(\mathbf{x}) = C_1(\mathbf{x}) = \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) \neq C_2(\mathbf{x}) \neq \tilde{C}_2(\mathbf{x})$: $(+,\ -) \mapsto (+,\ -)$,
- $\hat{C}_1(\mathbf{x}) = C_1(\mathbf{x}) \neq \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) = C_2(\mathbf{x}) \neq \tilde{C}_2(\mathbf{x})$: $(+,\ +) \mapsto (-,\ -)$,

- $\hat{C}_1(\mathbf{x}) = C_1(\mathbf{x}) \neq \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) \neq C_2(\mathbf{x}) \neq \tilde{C}_2(\mathbf{x})$: $(+, -) \mapsto (-, -)$,
- $\hat{C}_1(\mathbf{x}) \neq C_1(\mathbf{x}) = \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) \neq C_2(\mathbf{x}) = \tilde{C}_2(\mathbf{x})$: $(-, -) \mapsto (+, +)$,
- $\hat{C}_1(\mathbf{x}) \neq C_1(\mathbf{x}) = \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) \neq C_2(\mathbf{x}) \neq \tilde{C}_2(\mathbf{x})$: $(-, -) \mapsto (+, -)$,
- $\hat{C}_1(\mathbf{x}) \neq C_1(\mathbf{x}) \neq \tilde{C}_1(\mathbf{x})$ и $\hat{C}_2(\mathbf{x}) \neq C_2(\mathbf{x}) \neq \tilde{C}_2(\mathbf{x})$: $(-, -) \mapsto (-, -)$.

Положению $\mathbf{x}$ до переноса и после сопоставлено символьное обозначение, которое приведено в списке. $(+, +) \mapsto (+, -)$, например, означает, что происходит перенос объекта $\mathbf{x}$ из экспертной области и направления в ту же область, но в направление, не совпадающий с экспертным. Сопоставим каждому указанному выше случаю штраф δ за осуществление такого переноса. Пусть Q_1 – значение оптимизируемой функции Q (3), когда объект $\mathbf{x}$ находится в области $\hat{C}_1(\mathbf{x})$ и направлении $\hat{C}_2(\mathbf{x})$, а Q_2 – значение той же функции, когда объект $\mathbf{x}$ находится в области $\tilde{C}_1(\mathbf{x})$ и направлении $\tilde{C}_2(\mathbf{x})$. Перенос объекта $\mathbf{x}$ из направления $\hat{C}_2(\mathbf{x})$ в направление $\tilde{C}_2(\mathbf{x})$ будем теперь осуществлять при выполнении условия

$$Q_2 - Q_1 \geq \delta,$$

где δ – штраф, соответствующий переносу. Так как разных случаев переноса объекта $\mathbf{x}$ девять и каждому соответствует свой штраф δ, составим матрицу штрафов $\mathbf{F}$ (см. табл. 1). При этом предполагаем $\delta_{11} = \delta_{22} = \delta_{33} = 0$, так как переносы объекта $\mathbf{x}$ такого

Таблица 1: Матрица штрафа $\mathbf{F}$.

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

вида не добавляют отличий кластеризации, построенной алгоритмом, и экспертной кластеризации.

3.2 Визуализация модели на плоскости

Опишем как визуализировать имеющуюся экспертную иерархическую модель на плоскости. Определим координаты центров кластеров на l-м уровне иерархии $\boldsymbol{\mu}_l^1, \ldots, \boldsymbol{\mu}_l^{k_l}$, где k_l – количество кластеров на l-м уровне иерархии. Обозначим $r_l^1, \ldots, r_l^{k_l}$ радиусы соответствующих кластеров, понимаемые как расстояния от центра соответствующего до самого далекого документа из этого кластера. В качестве расстояния можно использовать евклидову метрику (для получения границы кластера в виде многомерной сферы) или расстояния городских кварталов (для получения границы в виде

границы многомерного куба). Соответствующее выбранному расстоянию двумерное расстояние будет использоваться и для определения расстояния между проекциями центров кластеров на плоскости. Обозначим $\rho(\cdot,\ \cdot)$ выбранное расстояние в исходном многомерном пространстве, а $\rho_2(\cdot,\ \cdot)$ – соответствующее ему расстояние на плоскости.

Начиная с верхнего уровня иерархии выполняем проекцию центров кластеров на плоскость, например, с помощью многомерного шкалирования, метода главных компонент или проекции Саммона. Получаем двумерные координаты проекций центров кластеров $\hat{\boldsymbol{\mu}}_l^1,\ \ldots,\ \hat{\boldsymbol{\mu}}_l^{k_l}$. Затем определяем радиус кластера на плоскости как

$$\hat{r}_l^j = \min_{i \neq j} \frac{r_l^j}{r_l^j + r_l^i} \rho_2(\hat{\boldsymbol{\mu}}_l^i,\ \hat{\boldsymbol{\mu}}_l^j). \tag{10}$$

Для вложения следующего уровня иерархии в предыдущий предлагается делать следующее. Рассмотрим произвольный кластер с номером i на l-м уровне иерархии. Рассмотрим кластеры $C_1,\ \ldots,\ C_q$ $l+1$-го уровня, содержащиеся в нем. Осуществим аналогично предыдущему проекцию их центров на плоскость и определим радиусы кластеров $r_1,\ \ldots,\ r_q$. Совместим центр масс системы проекций с центром рассматриваемого кластера l-го уровня и обозначим $\rho_1,\ \ldots,\ \rho_q$ расстояния от центров кластеров $l+1$-го уровня до центра рассматриваемого кластера l-го уровня. Пусть R – радиус рассматриваемого кластера l-го уровня. Определим расстояние от кластера $l+1$-го уровня, наиболее удаленного от центра кластера l-го уровня как

$$\hat{\rho} = \max_{i \in \{1,\ \ldots,\ q\}} \rho_i + r_i. \tag{11}$$

Далее проводим гомотетию (стягивание, если $\hat{\rho} > R$ и растяжение, если $\hat{\rho} < R$) с коэффициентом $\dfrac{R}{\hat{\rho}}$ и центров в центре рассматриваемого кластера l-го уровня. После этого даже самые далекие документы кластеров $l+1$ уровня будут находиться внутри рассматриваемого кластера l-го уровня. Проводя эту операцию для всех кластеров сверху вниз получим визуализацию иерархии, обладающую свойством вложенности, то есть разные кластеры не пересекаются и кластеры более низкого уровня лежат рядом и внутри общего кластера более высокого уровня.

4 Вычислительный эксперимент

Для проверки работы предложенных алгоритмов проводилась верификация иерархической тематической модели конференции EURO 2012. В качестве исходных данных был взят набор из 1342 тезисов данной конференции и модель, построенная экспертами. Модель состояла из 4-х уровней иерархии $h = 4$ (см. рис. 3).

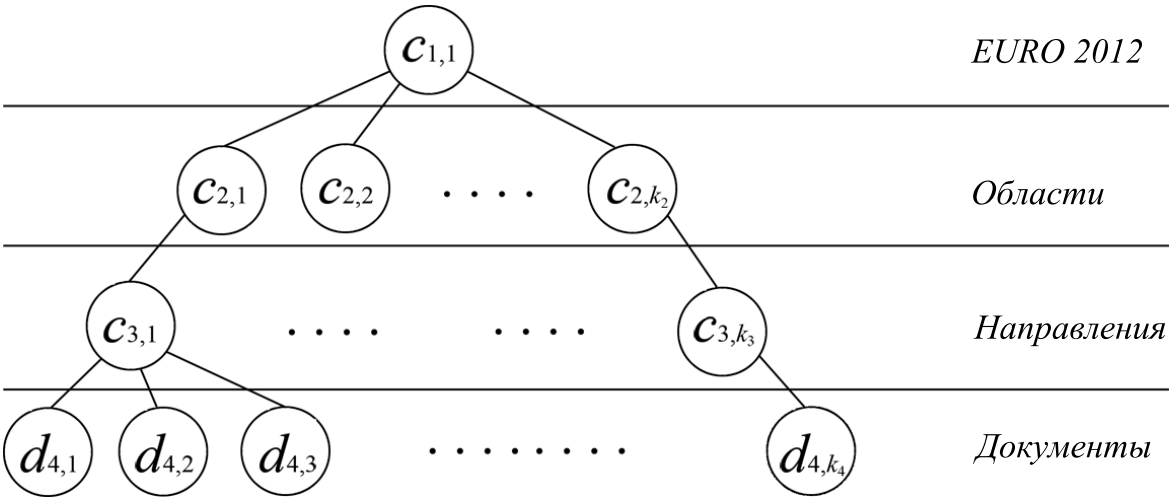


Рис. 3: Иерархическая структура конференции EURO 2012

4.1 Предобработка данных

После приведения слов в документах к начальной форме, был получен словарь конференции, состоящий примерно из 7000 слов. Чтобы оценить качество полученного словаря использовался следующий функционал:

$$\mathcal{D} = \frac{\sum\limits_{\mathbf{x}^i,\mathbf{x}^j \in c_{2,k}, i<j} \sum\limits_{t=1}^{n} |x_t^i - x_t^j|}{\sum\limits_{\mathbf{x}^i,\mathbf{x}^j \in c_{2,k}, i<j} \sum\limits_{t=1}^{n} (x_t^i + x_t^j)}, \tag{12}$$

где k —номер произвольного кластера уровня иерархии 2. Таким образом, если $\mathcal{D} \approx 1$, то пересечений в словах, принадлежащих одному кластеру, почти нет, что противоречит нашему предположению, что схожие документы имеют схожий терминологический состав. При использовании критерия $\mathrm{tf} \cdot \mathrm{idf}$ для отсева неинформативных слов, значение $\mathcal{D}$ равнялось 0.99. Это вызвано значительным количеством шумовых слов в словаре и существенным разбросом их встречаемостей. Поэтому отбор терминов проводился экспертно. При этом не только отбрасывались неинформативные слова, но и схожие с экспертной точки зрения термины объединялись в один. Это позволило уменьшить значение $\mathcal{D}$ до 0.96 и сократить словарь до 1063 терминов.

4.2 Построение алгоритмической иерархии

Кластеризуем коллекцию D алгоритмом, описанным в разделе 3.1 с параметром оптмизируемой функции Q (9) $\alpha = 0.1$. Задавая различные штрафы, мы с различным весом учитываем существующую экспертную модель. Если требуется выявить небольшое число наиболее сильных тематических противоречий, то штрафы на перемещение

документа из его экспертного кластера следует задавать большие. Если же целью является построить модель не основываясь на экспертной, то штрафы следует устремить к нулю. В табл. 2 приведена матрица, использованная для построения модели. При выборе значений элементов данной матрицы использовалась следующая логика:

1. Если документ не был перенесен, то и штрафа не должно быть. Поэтому на диагонале будут 0.

2. Чем больше различий с экспертной моделью перемещая документ мы добавим, тем на большую величину должен вырасти функционал качества при данном переносе, чтобы мы его совершили.

3. Должно выполняться свойство транзитивности. Иными словами порог у действия $(-,-) \to (-,+),\ (-,+) \to (+,+)$ должен совпадать с порогом действия $(-,-) \to (+,+)$.

4. Переносы, возвращающие документы в экспертные кластеры поощряются.

5. Чтобы избежать циклов, сумма порогов для круговых переносов документа вида $(+,+) \to (+,-),\ (+,-) \to (+,+)$ должна быть больше нуля.

Матрица штрафов $\mathbf{F}$ выражается через $\tilde{\mathbf{F}}$ как

$$\mathbf{F} = \gamma\tilde{\mathbf{F}},$$

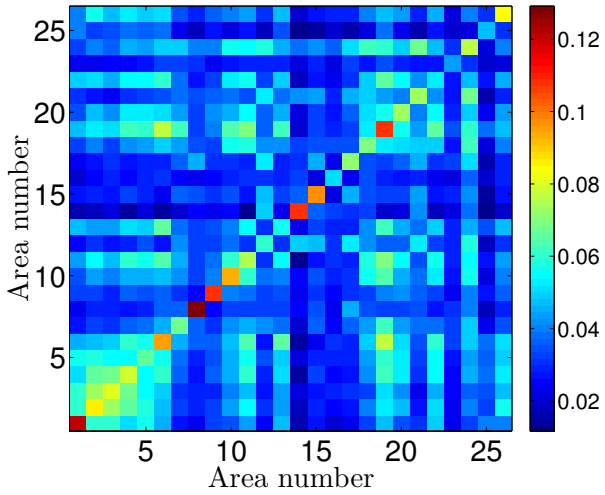
где $\gamma \geq 0$ – весовой множитель, регулирующий допустимую степень несоответствия построенной кластеризации и экспертной. Результаты кластеризации, соответствующие матрице штрафов $\mathbf{F}$ для трех разных значений γ приведены в табл. 3.

Таблица 2: Матрица $\tilde{\mathbf{F}}$, задающая штраф.

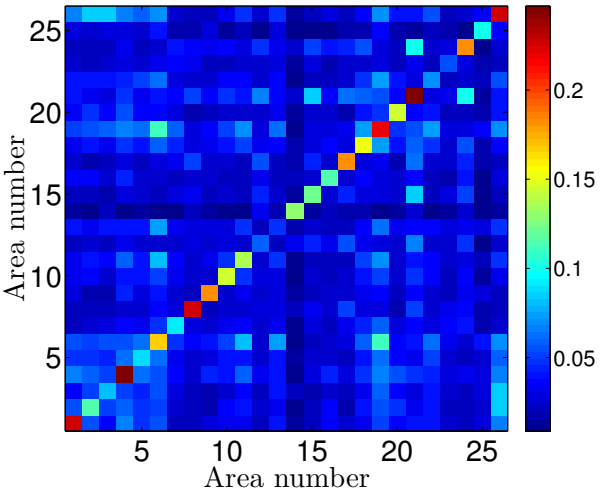
Из \ В	$(+,+)$	$(+,-)$	$(-,-)$
$(+,+)$	0	0.002	0.005
$(+,-)$	-0.001	0	0.003
$(-,-)$	-0.003	-0.002	0

Приведем далее более подробнее результаты для $\gamma = 0.7$. При таком значении параметра штрафа 878 из 1342 документов оказались в экспертных областях и 730 в экспертных направлениях (см. табл. 3). В табл. 3 также приведены значения средних внутри- и межкластерных сходств для областей и направлений в экспертной и построенной кластеризациях. Произошло значительное увеличение внутрикластерного сходства как на уровне областей, так и на уровне направлений и снижение межкластерного сходства на обоих уровнях. Приведем далее графики, иллюстрирующие

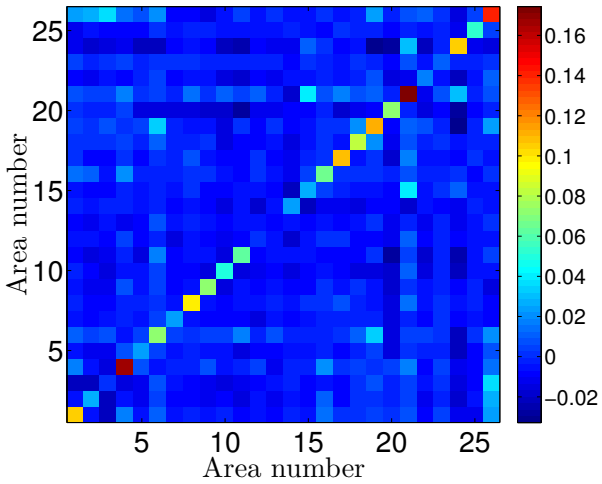
изменение сходства между всеми кластерами для областей (см. рис. 4) и направлений (см. рис. 5), а также процентное распределение документов экспертных кластеров по построенным алгоритмом (см. рис. 6). Графики на рис. 4–6 показывают, что произошло перераспределение документов, но все же каждый кластер, построенный алгоритмом, содержит больше всего документов экспертного кластера, который ему соответствует. При этом почти для всех кластеров произошло значительное увеличение внутрикластерного сходства.



(а) Экспертная кластеризация

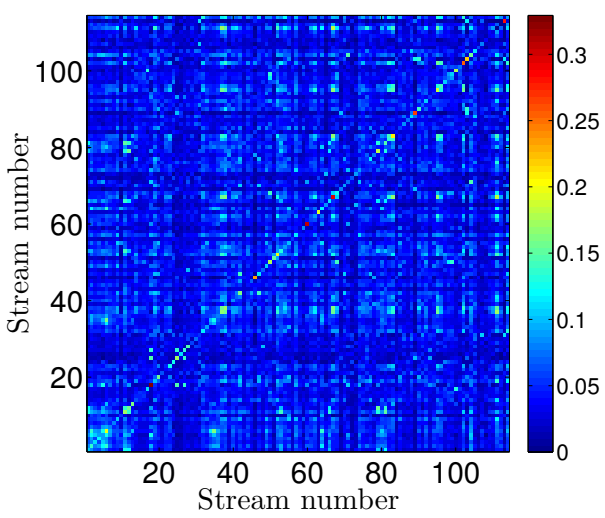


(б) Построенная кластеризация

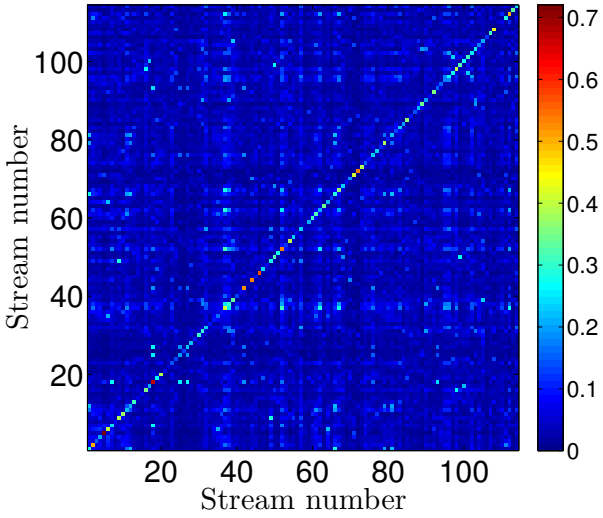


(в) Разность сходств областей для исходной и построенной кластеризаций

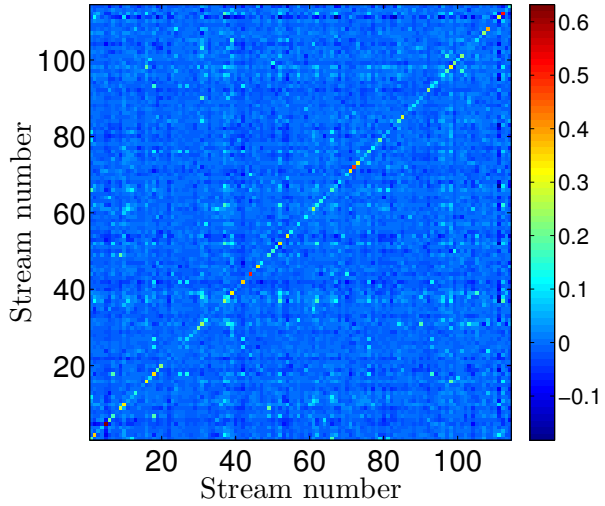
Рис. 4: Сравнение среднего сходства по областям.



(а) Экспертная кластеризация

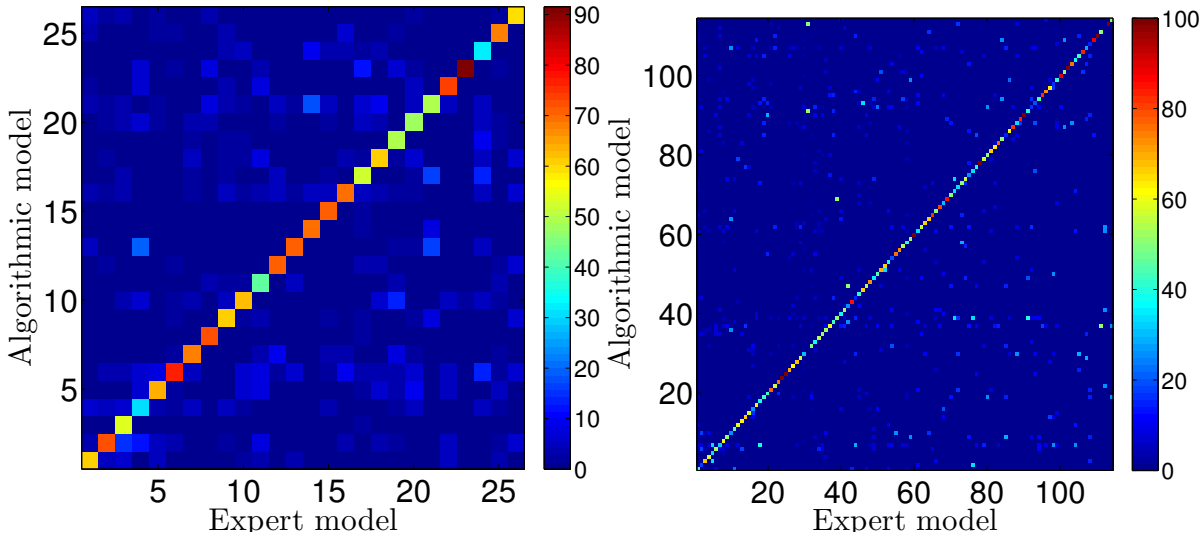


(б) Построенная кластеризация



(в) Разность сходств направлений для исходной и построенной кластеризаций

Рис. 5: Сравнение среднего сходства по направлениям.



(а) Распределение по областям (б) Распределение по направлениям

Рис. 6: Процентное распределение документов по областям и направлениям.

Таблица 3: Результаты кластеризации, соответствующие штрафам из табл. 2.

	Среднее внутрикластерное сходство		**Среднее межкластерное сходство**		**Число совпадений с экспертной моделью**	
Вес экспертной модели	Области	Направления	Области	Направления	Области	Направления
Экспертная модель ($\gamma = \infty$)	108	147	56.5	57.4	1342	1342
Сильно учитывается ($\gamma = 1.25$)	123	190	56.8	58.4	1272	1225
Средне учитывается ($\gamma = 0.7$)	185	309	54.9	53.4	878	730
Слабо учитывается ($\gamma = 0.5$)	204	349	55.0	53.8	689	508

4.3 Визуализация экспертной модели и выявленных некорректностей

На рис. 7, 8, 9 показан результат визуализации экспертной модели алгоритмом, описанным в разделе 3.2. Центры уровня областей отмечены меткой «x», уровня направлений — «+», а документы d отображаются метками «o». Вокруг центра каждого кластера проводится его граница, и все объекты лежащие внутри границы данного кластера принадлежат ему.

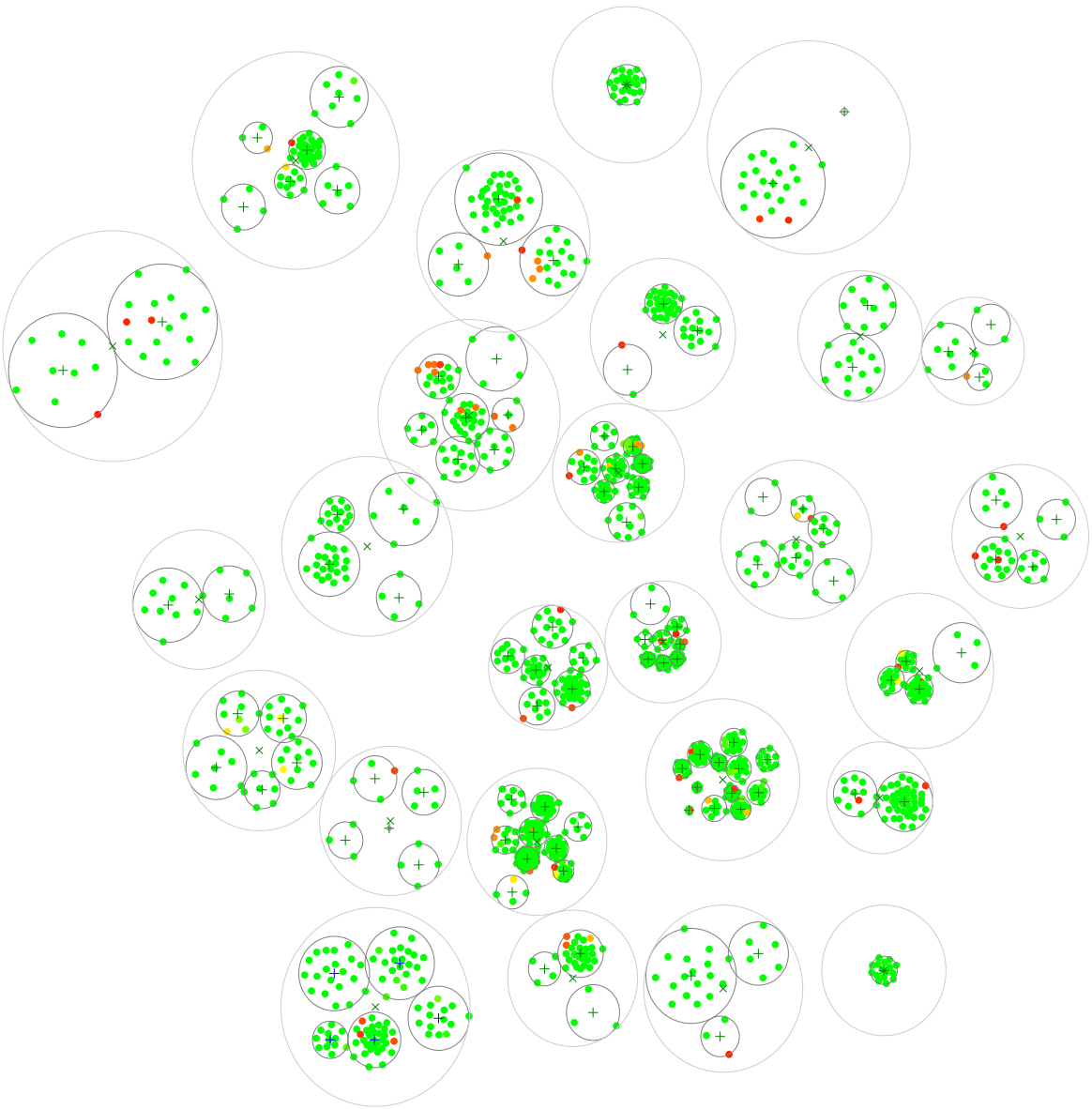


Рис. 7: Иерархическая визуализация несоответствий с большими штрафами.

Цвет документа d^s определяется степенью некорректности его экспертной кластеризации с тематической точки зрения. Для отображения некорректностей будем использовать цветовую шкалу от RGB $(255; 0; 0)$ — красный (документ, для которого алгоритмическая и экспертная кластеризации отличаются сильнее всего) до RGB $(0; 255; 0)$ — зеленый (документы, для которых экспертная и алгоритмическая кластеризации совпадают). Отобразим на шкалу полученный диапазон значений некорректностей, сопоставив таким образом каждому числовому значению некорректности цвет, и этим цветом раскрасим точку на плоскости, соответствующую d^s.

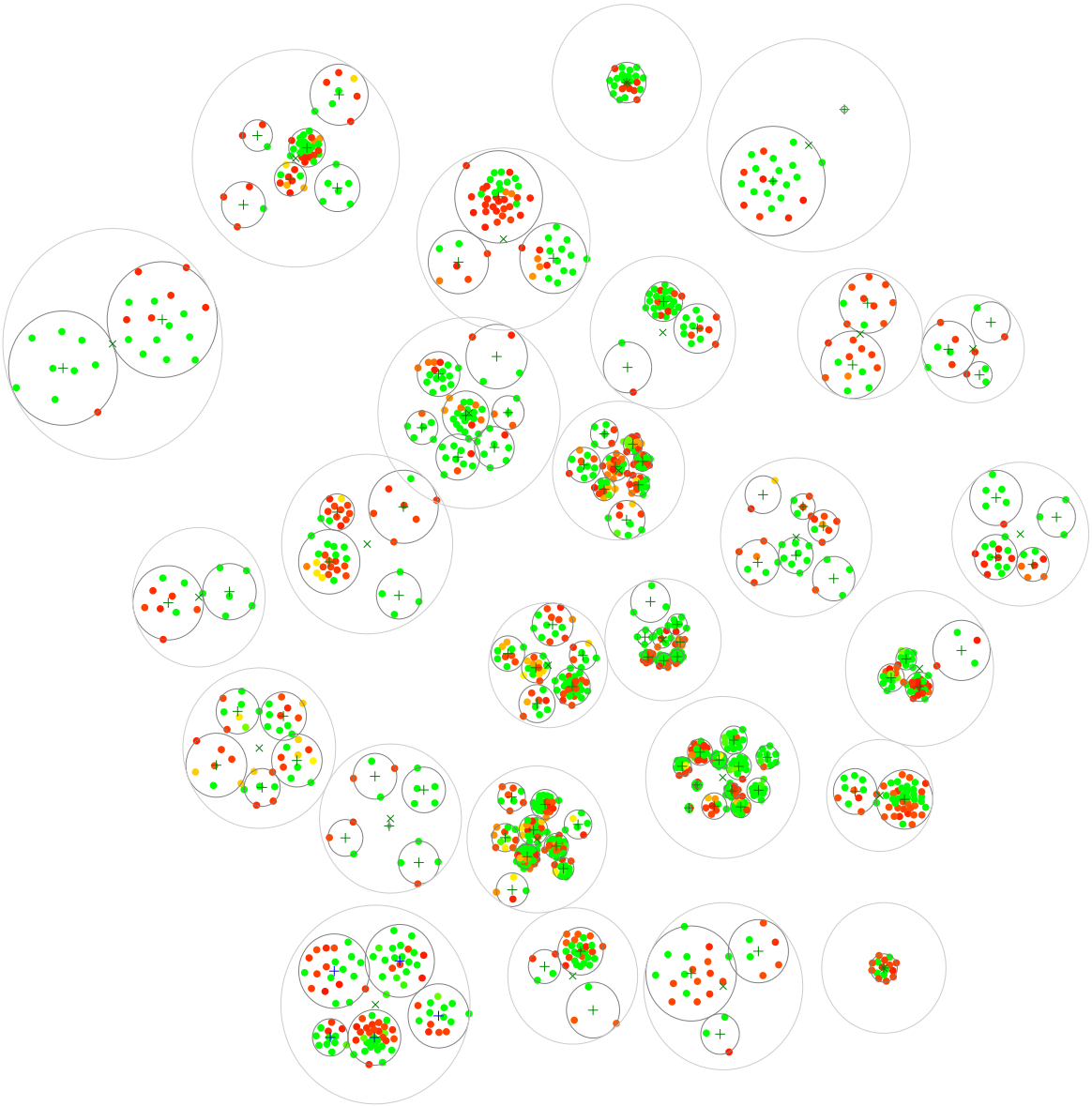


Рис. 8: Иерархическая визуализация несоответствий со средними штрафами.

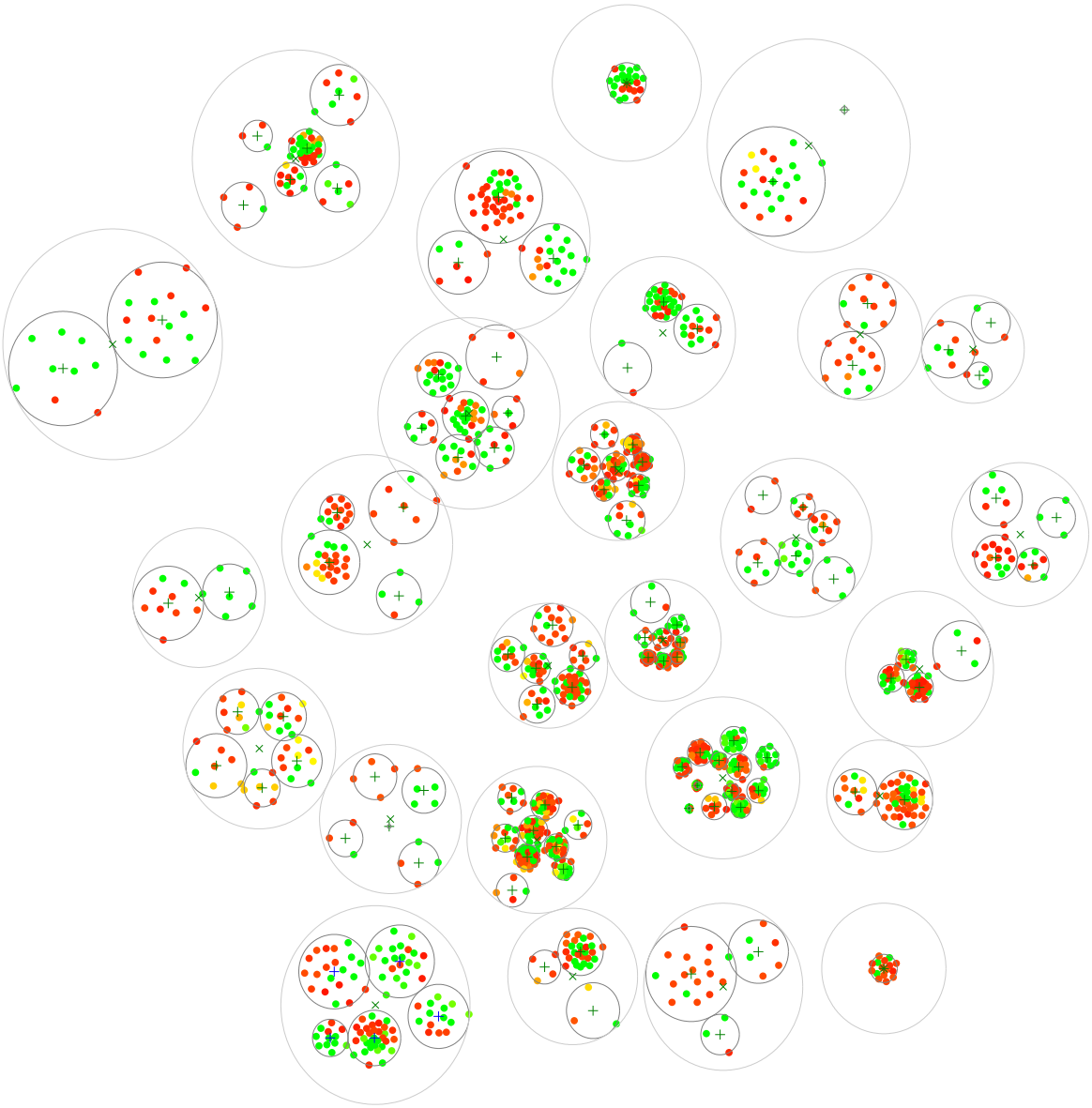


Рис. 9: Иерархическая визуализация несоответствий с малыми штрафами.

По полученной визуализации модели на плоскости можно судить о следующих параметрах экспертной модели:

1. Тематическом расстоянии между подкластерами одного уровня, лежащих в одном кластере. Например тематически схожие области лежат на плоскости рядом. Аналогично два тематически схожих направления, лежащих внутри одной области, также будут находиться рядом на плоскости.

2. Качестве кластеризации конкретного документа. Если документ находится вда-

ли от других документов из данного кластера или на краю кластера, и алгоритм выделяет его красным цветом, то данный документ отличается по терминологическому составу от остальных документов, лежащих в данном кластере.

3. Корректности кластера. Если кластер содержит больше половины «красных» документов, то целесообразно проверить оправданность существования темы, соответствующей данному кластеру.

Однако по построенной визуализации нельзя судить о тематической близости объектов, принадлежащих разным кластерам, например по изображению нельзя выяснить, близки ли два направления лежащие в разных областях.

5 Заключение

В данной работе предлагался метод анализа и верификации экспертной тематической модели крупной конференции. Был предложен метод составления терминологического словаря конференции, метод построения иерархической тематической модели, с различным весом учитывающей экспертную, алгоритм визуализации экспертной модели на плоскости и способ отображения выявленных несоответствий между алгоритмической и экспертной моделями. Работа предложенных алгоритмов продемонстрированна верификацией экспертной тематической модели конференции EURO 2012.

Список литературы

[1] *Воронцов К. В.* Вычислительные методы обучения по прецедентам. http://www.machinelearning.ru/wiki/images/6/6d/Voron-ML-1.pdf

[2] *Blei D. M., Ng A. Y., Jordan M. I.* Latent dirichlet allocation. // Journal of Machine Learning Research, 2003. Vol. 3. Pp. 993-1022.

[3] *Кузьмин А. А., Стрижов В. В.* Проверка адекватности тематических моделей коллекции документов. // Программная инженерия, 2013, 4 — 16-20.

[4] *Кузьмин А. А., Адуенко А. А., Стрижов В. В.* Выбор признаков и оптимизация метрики при кластеризации коллекции документов. // Известия Тульского государственного университета, Естественные науки, 2012, 3 — 119-131.

[5] *Hofmann T.* Probabilistic latent semantic indexing. // Proceedings of the 22nd annual interanational ACM SIGIR conference on research and development in information retrieval. New York: ACM, 1999. Pp. 50–57.

[6] *Blei D. M., Lafferty J. D.* Topic Models. // Text Mining: Classifcation, Clustering, and Applications. Chapman & Hall/CRC Press, 2009.

[7] *Hartigan J. A., Wong M. A.* Algorithm as 136: A k-means clustering algorithm. // Applied statistics, 1978. Vol. 28. Pp. 100–108.

[8] *Загоруйко Н. Г., Ёлкина В. Н., Лбов Г. С.* Алгоритмы обнаружения эмпирических закономерностей. Новосибирск: Наука, 1985.

[9] *Pal N. R.,Bezdek J. C.* On cluster validity for the fuzzy c-means model. // IEEE Transactions on Fuzzy Systems, 1995. Vol. 3(3). Pp. 370–379.

[10] *Загоруйко Н. Г.* Прикладные методы анализа данных и знаний. // Новосибирск: Издательство И.М., 1999.

[11] *Борисова И. А.* Использование fris-функции для построения решающего правила и выбора признаков (задача комбинированного типа dx). //Новосибирск. Знания. Онтологии. Теории. Материалы Всероссийской Конференции, 2007. Т. 1. С. 37–44.

[12] *Tibshirani R., Hastie T.* Discriminative adaptive nearest neighbor classification. // IEEE transactions on pattern analysis and machine intelligence, Vol. 18 Issue 6, June 1996. Pp. 607–616.

[13] *Peng J., Gunopulos D., Domenciconi C.* An adaptive metric machine for pattern classification. // Advances in Neural Information Processing Systems 13, MIT Press, 2000. Pp. 458–464.

[14] *Zagoruiko N. G.* Methods of recognition based on the function of rival similarity. // Pattern recognition and image analysis, 2008. Vol. 18(1). Pp. 1–6.

[15] *Manning C. D., Raghavan P., Schütze H.* Introduction to information retrieval. Cambridge: Camdridge University Press, 2008.

[16] *Daud A., Li J., Zhou L., Muhammad F.* Knowledge discovery through directed probabilistic topic models. // Frontiers of computer science in China, 2010. Vol. 4(2). Pp. 280–301.

[17] *Miguel A.* A Review of Dimension Reduction Techniques. Dept. of Computer Science University of Sheffield, 1997.

[18] *Зиновьев А. Ю.* Визуализация многомерных данных. Красноярск: Изд-во КГТУ, 2000.

[19] *Marghescu D.* Evaluating Multidimensional Visualization Techniques in Data Mining Tasks. Painosalama Oy – Turku, Finland, 2008

[20] *Wu H., Tien Y., Chen C.* GAP: A graphical environment for matrix visualization and cluster analysis. Elsevier, 2008.

[21] *Hamel L., Brown C.* Improved interpretability of the unified distance matrix with connected components. // Proc. 7th International Conference on Data Mining 2011, CSREA Press, Las Vegas. Pp. 338–343.

[22] *Haiyan Li* Data visualization of asymmetric data using sammon mapping and applications of self-organizing maps. Dissertation, Faculty of the Graduate School of the University of Maryland, 2005.

[23] *Condon E., Golden B., Lele S., Raghavan S., Wasil E.* A visualization model based on adjacency data. Elsevier, 2001.

[24] *van der Laan M. J., Pollard K. S.* A new algorithm for hybrid hierarchical clustering with visualization and the bootstrap. // Journal of Statistical Planning and Inference, 2003. Vol. 117, No. 2., Pp. 275-303.

[25] *Gilles Bisson, Renaud Blanch* Improving Visualization of Large Hierarchical Clustering. // 16th International Conference on Information Visualisation, 2012. Pp. 220–228.

[26] *Stevens J. R.* Clustering with Gene Expression Data. Utah State University, 2013.

[27] *Константинов Р. В.* Функциональный анализ. Курс лекций. Долгопрудный, 2009.

[28] *Loohach R., Garg K.* Effect of distance functions on simple k-means clustering problem. // International Journal of Computer Applications, 2012. Vol. 49. No. 6.

[29] Тезисы конференции EURO 2012. https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/Kuzmin2013HierarchicalVisualization/data/