Kuzmin A.A., Aduenko A.A., Strijov V.V. Thematic classification using expert model for major conference abstracts // Information Technologies, 2014, 6 : 22-26.

УДК 519.256

А. А. Кузьмин, студент, Московский физико-технический институт
А. А. Адуенко, студент, Московский физико-технический институт
В. В. Стрижов, к.ф.-м.н., Вычислительный Центр РАН

Тематическая классификация тезисов крупной конференции с использованием экспертной модели[1]

Работа посвящена определению тем, научных направлений и сессий тезисов крупной научной конференции. Рассматривается коллекция тезисов конференции с экспертной тематической моделью. Строится терминологический словарь конференции. Предлагается функция сходства двух тезисов. Методом неметрической иерархической кластеризации строится алгоритмическая модель конференции, с заданным весом учитывающая существующую экспертную модель. Выявляются несоответствия между экспертной моделью и предлагаемой. Алгоритм построения тематической модели проиллюстрирован кластеризацией коллекции тезисов конференции EURO 2013.

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

A. A. Kuzmin, A. A. Aduenko, V. V. Strijov

Thematic classification using expert model for major conference abstracts

This paper is devoted to the thematic verification of the areas, streams and sessions of a major conference. An abstract collection and its expert thematic model are considered. The terminological dictionary of the conference is constructed. A similarity function between two abstracts is proposed. A non-metric clustering algorithm is used to construct the hierarchical thematic model of the conference. The expert model is considered in this algorithm and its parameters are optimized, to make the algorithmic model similar to the expert model. The expert model is compared with algorithmic model. The algorithm is illustrated by clustering of the EURO 2013 abstracts.

Key words: document collection, thematic model, hierarchical model, clustering.

1 Введение

При организации крупной конференции, возникает задача построения ее тематической модели. Рассмотрим процесс построения тематической модели тезисов (далее документов) на примере конференции EURO 2013. Документом является аннотация к докладу участника конференции, состоящая не более чем

[1] Подано в журнал «Информационные технологии, 2013»
E-mail: strijov@ccas.ru

из 600 символов. Конференция содержит в себе 24 главные темы (далее область), определяемые председателем программного комитета. Каждая главная тема содержит в себе примерно 10 больших подтем (далее направление), каждая из которых делится на сессии, содержащие в себе ровно четыре документа. Эксперты, исходя из своей стратегии организации докладов конференции и основываясь на содержании документа, относят его к одной из областей, затем к одному из направлений данной области. После этого все документы, попавшие в одно направление, разбиваются на сессии.

В силу большого числа экспертов и отсутствия эталонной модели, оценить качество экспертной тематической модели сложно. Поэтому предлагается построить алгоритмическую тематическую модель коллекции документов, основанную на их терминологическом сходстве, и сравнить ее с экспертной.

Для этого, сначала составляется экспертным образом терминологический словарь конференции, и из всех документов удаляются слова, не являющиеся терминами. После отсева неинформативных слов, документы представляются в виде «мешков слов» [1] и каждому документу ставится в соответствие целочисленный вектор. В работе [2] сравниваются способы построения вектора — описания документа.

Кластеризация текстов выполняется с помощью метрических алгоритмов кластеризации, например, K-means [3], FOREL [4], C-means [5], STOLP [6], FRiS-STOLP [7], BoostML, DANN [8] и другие [9, 10], и с помощью вероятностных методов [11], например, с помощью вероятностного латентного семантического анализа [1], или латентного размещения Дирихле [12].

Для построения иерархической кластерной модели существует два основных типа алгоритмов — дивизимные и агломеративные [13]. В данной работе для построения тематической модели предлагается неметрический алгоритм кластеризации (метрический алгоритм для аналогичной задачи был предложен в [13]), перераспределяющий объекты между кластерами так, чтобы улучшить среднее сходство между объектами из одного кластера и увеличить 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 такая кластеризация, согласно которой объект принадлежит только к одному из кластеров, так как каждый документ может принадлежать только одной теме, что соответствует правилам проведения конференции.

2 Математическое представление конференции

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, $d = \{w_j\}$, $j \in \{1, \ldots, n\}$.

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

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

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

$$\mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^{\mathrm{T}} \mathbf{x}_s}}. \qquad (2)$$

Представим иерархическую тематическую модель в виде дерева, см. рис. 1. Глубину дерева обозначим h, на рис. 1 глубина $h = 4$. 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).

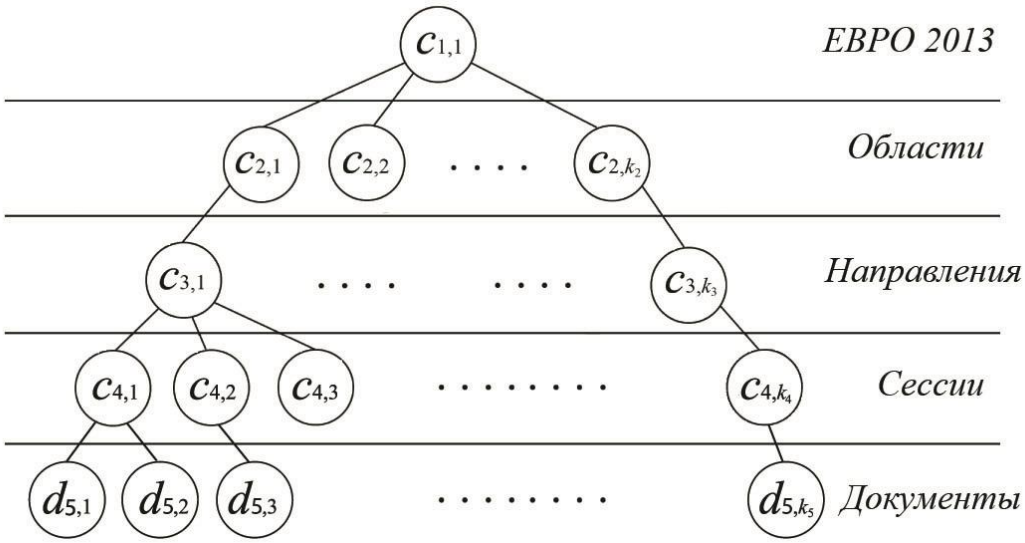


Рис. 1. Иерархическое представление тематической модели

3 Функция сходства документов

Предполагается, что каждый документ в коллекции может быть описан небольшим набором признаков — ключевых слов. В рассматриваемой в данной работе коллекции [14] 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:

$$s(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^{\mathrm{T}} \mathbf{x}_i}\sqrt{\mathbf{x}_j^{\mathrm{T}} \mathbf{x}_j}} = \mathbf{x}_i^{\mathrm{T}} \mathbf{x}_j \qquad (3)$$

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.

Под сходством $S(c_{l,i}, c_{l,j})$ 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:

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

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):

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

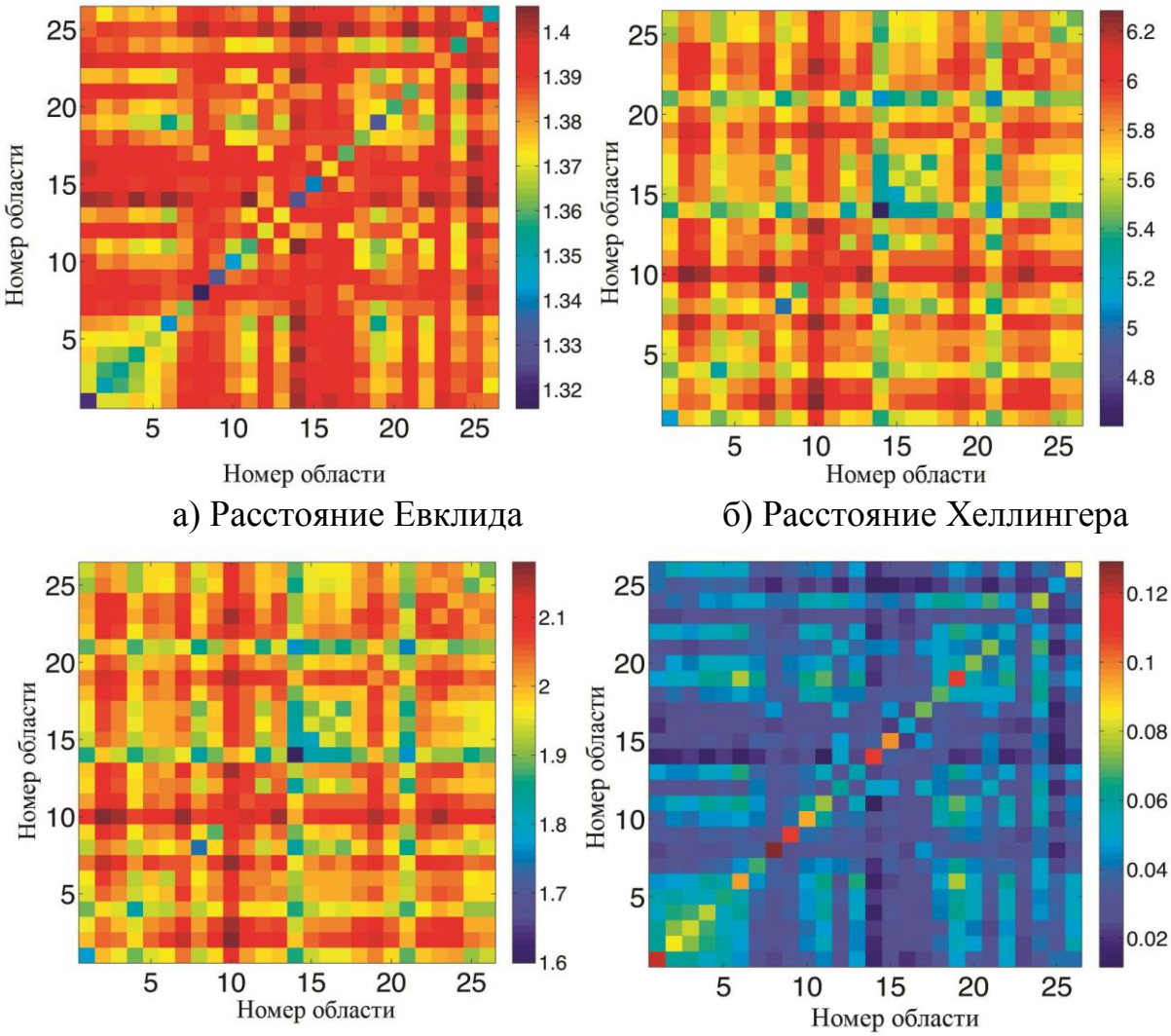
$$H(\mathbf{x}, \mathbf{y}) = \frac{1}{\sqrt{2}} \| \sqrt{\mathbf{x}} - \sqrt{\mathbf{y}} \|_2, \qquad (6)$$

$$JSD(\mathbf{x} \| \mathbf{y}) = \frac{1}{2} D(\mathbf{x} \| M) + \frac{1}{2} D(\mathbf{y} \| M), M = \frac{1}{2}(\mathbf{x} + \mathbf{y}), \quad \text{где}$$

$$D(\mathbf{x} \| \mathbf{y}) = \sum_i \ln\left(\frac{x_i}{y_i}\right) x_i \quad \text{— расстояние Кульбака Лейблера.} \qquad (7)$$

На рис. 2а)-в) приведены значения средних расстояний между документами разных областей для указанных трех типов расстояния для экспертной кластеризации. Рис. 2г) 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. Элементы диагонали (x, x) соответствуют внутрикластерному расстоянию, а элементы $(x, y), x \neq y$ — межкластерному. В случае г) средним расстоянием является сходство двух кластеров. Показателем качества кластеризации являются большие межкластерные расстояния по сравнению с внутрикластерными, что соответствует пику на диагонали. Отличия в среднем внутрикластерном расстоянии и среднем межкластерном расстоянии для а)-в) почти отсутствуют.

Поэтому применение метрических алгоритмов кластеризации в данной задаче необоснованно. Значимые отличия между сходством или расстоянием внутри одной области и между разными областями наблюдается только на рис. 2г), 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)

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

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

для разных областей в экспертной модели. В соответствии с (4) и (3) при $i \neq j$

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

Аналогично,

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

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

В последнем выражении учтена нормировка $\mathbf{x}^{\mathrm{T}}\mathbf{x}=1$. 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.

4 Функционал качества и алгоритм кластеризации документов

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

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

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.

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 эта операция дает максимальный увеличение 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

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

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).

5 Построение словаря

Рассмотрим подробнее процесс составления терминологического словаря конференции W. Сначала проводится предобработка данных. После приведения слов в документах к начальной форме, получается полный словарь конференции. На следующем шаге из словаря требуется исключить все слова, не являющиеся терминами. Чтобы оценить качество терминологического словаря в данной работе использовался следующий функционал:

$$\mathsf{D} = \frac{\sum\limits_{\mathbf{x}^i,\mathbf{x}^j \in c_{2,k}, i<j} \sum\limits_{t=1}^{n} |x_t^i - x_t^j|}{\sum\limits_{\mathbf{x}^i,\mathbf{x}^j \in c_{2,k}, i<j} \sum\limits_{t=1}^{n} (x_t^i + x_t^j)}, \qquad (10)$$

где k — номер произвольного кластера уровня иерархии 2. Если $\mathsf{D} \approx 1$, то пересечений в терминах документов, 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 D равнялось 0.99. Это вызвано значительным количеством шумовых слов в словаре и существенным разбросом их встречаемости. 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 D до 0.96 и сократить словарь до 1063 терминов.

6 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. Сопоставим документу $\mathbf{x}$ пару — (область, направление), в которой на месте слова «область» стоит знак $+$, если у документа $\mathbf{x}$ область совпадает с экспертной, и знак $-$ иначе. Аналогично, вместо слова «направление» стоит знак $+$, если для этого документа направление совпадает с экспертным, и знак $-$ иначе. Например, выражение $(+,-)$ означает, что документ находится в экспертной области, но не в экспертном направлении.

Обозначение операции переноса $(+,+) \mapsto (+,-)$ объекта $\mathbf{x}$ означает, что объект $\mathbf{x}$ из экспертной области и направления переносится в экспертную область (ту же область), но в направление, не совпадающий с экспертным. Сопоставим каждому варианту переноса штраф δ за осуществление такого

переноса, см. таб. 0. При этом предполагаем $\delta_{11} = \delta_{22} = \delta_{33} = 0$, так как переносы объекта $\mathbf{x}$ такого вида не добавляют отличий кластеризации, построенной алгоритмом, и экспертной кластеризации. Переносы вида $(-,+) \mapsto (+,+)$ не рассматриваются, так как имеет место свойство вложенности: экспертное направление не может находится вне экспертной области.

Пусть Q_1 — значение оптимизируемой функции Q (3) до переноса документа $\mathbf{x}$, а Q_2 — ее значение после переноса. Перенос объекта $\mathbf{x}$ будем теперь осуществлять только при выполнении условия:

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

где δ — штраф, соответствующий переносу.

Табл. 1. Матрица штрафа $\mathbf{F}$

Из \ В	$(+,+)$	$(+,-)$	$(-,-)$
$(+,+)$	$\delta_{11} = 0$	$\delta_{12} = 0.002$	$\delta_{13} = 0.005$
$(+,-)$	$\delta_{21} = -0.001$	$\delta_{22} = 0$	$\delta_{23} = 0.003$
$(-,-)$	$\delta_{31} = -0.003$	$\delta_{32} = -0.002$	$\delta_{33} = 0$

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 требуется построить алгоритмическую модель без учета экспертной модели, то штрафы следует устремить к нулю. В табл. 1 приведена матрица, использованная для построения модели. Значения элементов δ определялись исходя их следующих эвристических правил.

1. Чем больше отличий от экспертной кластеризации вносит перенос документа, тем больше порог для осуществления переноса.
2. Должно выполняться свойство транзитивности. Иными словами порог у действия $(-,-) \to (-,+), (-,+) \to (+,+)$ должен совпадать с порогом действия $(-,-) \to (+,+)$.
3. Переносы, возвращающие документы в экспертные кластеры, поощряются.
4. Чтобы избежать циклов, сумма порогов для циклических переносов документа вида $(+,+) \to (+,-), (+,-) \to (+,+)$ должна быть больше нуля.

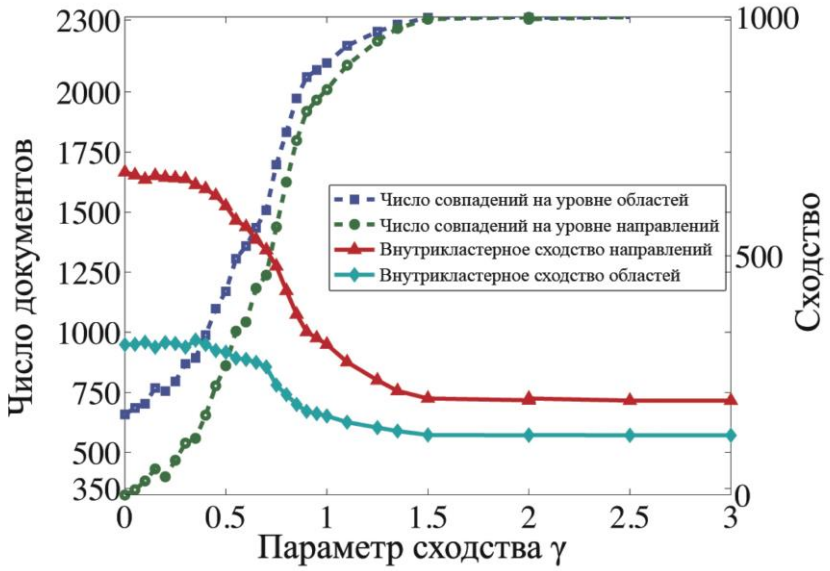


Рис. 3. Зависимость внутри- и межкластерного сходства на уровнях областей и направлений от параметра штрафа γ

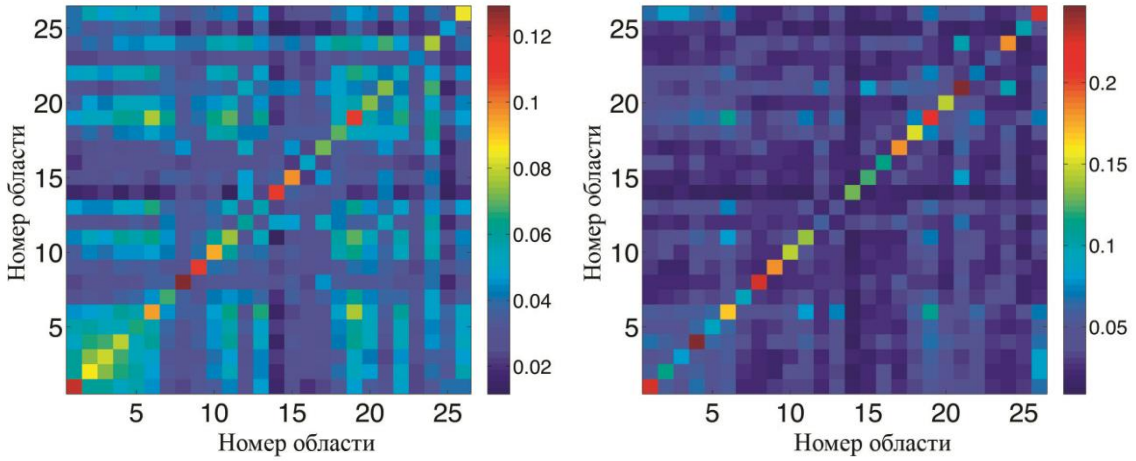
Матрица штрафов $\mathbf{F}$ выражается через $\tilde{\mathbf{F}}$ как

$$\mathbf{F} = \gamma\tilde{\mathbf{F}}, \qquad (11)$$

где $\gamma \geq 0$ — весовой множитель, регулирующий допустимую степень несоответствия построенной кластеризации и экспертной.

7 Вычислительный эксперимент

Для проверки работы предложенных алгоритмов проводилась верификация тематической модели конференции EURO 2013. В качестве исходных данных был взят набор из 2313 тезисов данной конференции и модель, построенная экспертами. Модель включала 4 уровня иерархии $h = 4$, см. рис. 1.



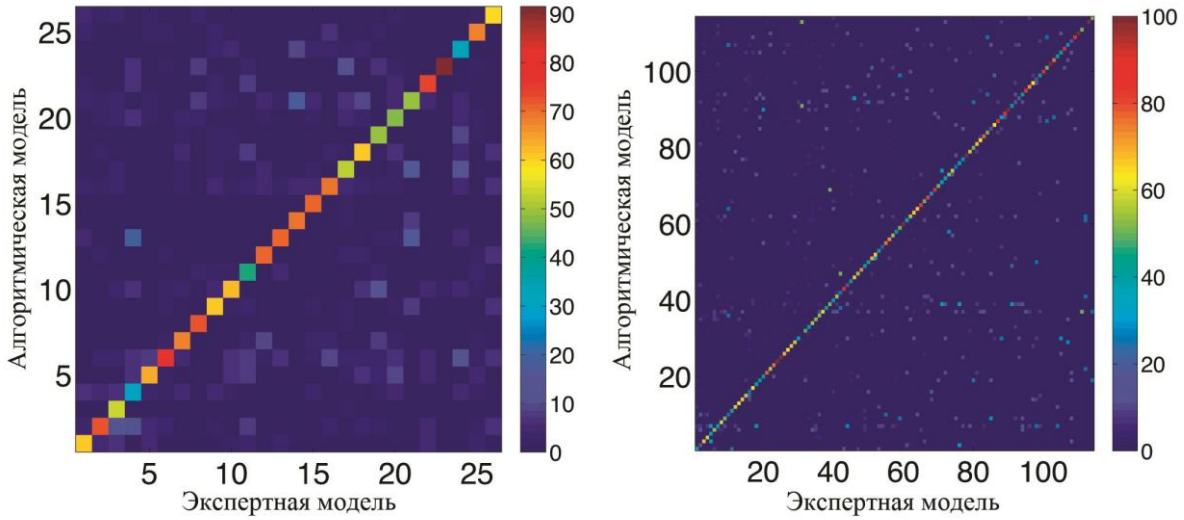
а) Экспертная кластеризация б) Построенная кластеризация

Рис 4. Сравнение среднего сходства по областям

Коллекция D кластеризовалась алгоритмом, описанным в разделе 6 с параметром $\alpha = 0.1$ оптимизируемой функции Q (9). Результаты кластеризации,

соответствующие матрице штрафов $\mathbf{F}$ (11) для разных значений γ приведены на рис. 3. По левой оси отложено количество документов, для которых экспертная и алгоритмическая кластеризации совпали, по правой оси значения среднего внутрикластерного сходства (4), а по нижней оси отложено соответствующее значение параметра γ. Чем больше γ, тем меньше документов попадают в чужие кластеры, но и внутрикластерное сходство становится меньше. Так при значении $\gamma > 2$, 99 % документов попадают в свои экспертные кластеры.

Приведем графики для значения параметра штрафа $\gamma = 0.7$, иллюстрирующие изменение сходства между всеми кластерами для уровня областей, см. рис. 4, и распределение документов из экспертных кластеров по алгоритмическим, см. рис. 5.



а) Распределение по областям б) Распределение по направлениям

Рис 5. Процентное распределение документов по областям и направлениям

На рис. 4 по осям отложены номера областей. Цвет клетки (x, y) соответствует значению среднего сходства между документами области с номером x и области с номером y. Клетки диагонали (x, x) соответствуют внутрикластерному сходству, а клетки $(x, y), x \neq y$ — межкластерному. Из графиков видно, что внутрикластерное сходство увеличилось в два раза, см. рис. 3 при $\gamma = 0.7$ и при $\gamma = 2.5$. При этом увеличение внутрикластерного сходства происходит для большинства кластеров (см. рис. 4).

На рис. 5 клетка с координатами (x, y) показывает количество документов, которые эксперт отнес к кластеру с номером x, а алгоритм к кластеру с номером y. Большие значения на диагонали $x = y$ и маленькие вне диагонали показывают, что 67 % документов остались в исходных (экспертных) кластерах. Таким образом, построенная модель схожа с экспертной.

8 Заключение

В данной работе предлагался метод анализа и верификации экспертной тематической модели крупной конференции. Был предложен метод составления

терминологического словаря конференции. Был предложен метод построения алгоритмической иерархической тематической модели с учетом экспертной тематической модели. Работа предложенных методов продемонстрирована верификацией экспертной тематической модели конференции EURO 2013.

Список литературы

[1] *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. P. 50-57.

[2] *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, С. 119-131.

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

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

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

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

[7] *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.

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

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

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

[11] *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). P. 280-301.

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

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

[14] Тезисы конференции EURO 2013, URL: http://euro2013.org/wp-content/uploads/program-euro26.pdf, дата обращения: 26.12.2013