А.С. Златов, А.А. Кузьмин

Построение иерархической тематической модели крупной конференции[1]

Аннотация. Работа посвящена построению иерархической тематической модели тезисов крупной конференции. Используется разделяющая вероятностная модель для кластеризации тезисов на каждом уровне иерархии. Предложены адаптированные вероятностные модели, учитывающие сбалансированность структуры конференции. В адаптированных моделях снижено влияние мощности кластеров на построение тематической модели. Для построения тематической модели используется алгоритм кластеризации с частичным обучением. Строится плоская модель на каждом уровне иерархии. На основании плоских моделей строится иерархическая тематическая модель конференции. Для построения тематической модели тезисов конференции используется дивизимный иерархический алгоритм. Работа алгоритмов проиллюстрирована на коллекциях тезисов конференции EURO и сайтов индустриального сектора. Разделяющая вероятностная модель сравнивается с адаптированными моделями и иерархической моделью. Для оценки качества тематической модели используются модели, построенные экспертами.

Ключевые слова: иерархические модели, тематические модели, вероятностные тематические модели, иерархическая кластеризация, алгоритмы с частичным обучением.

Введение

В работе описан алгоритм построения иерархической модели тезисов конференции [1]. Конференция EURO содержит в себе 26 областей. Каждая область содержит 10-15 научных направлений. Участники конференции присылают тезисы, состоящие из не более, чем 600 символов. На основании содержания тезисов и ключевых слов эксперты распределяют работы по областям и направлениям.

На крупной конференции обычно более 2000 докладов. Доклады делятся на два типа. Первый тип – приглашенные доклады, для которых заранее известны научная область, направление. Второй тип – неразмеченные доклады, которые требуется распределить по иерархической структуре. Докладов обоих типов примерно равное количество, т.е. более 1000 докладов требуется распределять вручную. Для этого привлекается до 200 экспертов из различных областей. Автоматическое распределение докладов позволило бы значительно сократить число экспертов и их время. Создание такой автоматической системы и является целью данной работы.

Ранее были предложены алгоритмы текстовой кластеризации [2-5]. В Табл. 1 представлены четыре основных типа алгоритмов по тому, каким способом они описывают документ и тему документа в коллекции.

Жесткие модели. Данный способ текстовой кластеризации является адаптацией кластеризации произвольных объектов в метрическом или неметрическом [6] пространстве. Документы представляются в виде векторов.

Для решения задачи кластеризации документов применяется алгоритм k-means [2]. Он предполагает использование евклидовой метрики в качестве функции расстояния, так как только в этом случае вычисление центра кластера при помощи усреднения векторов, принадлежащих данному кластеру, ведет к неубыванию функционала качества на шаге пересчета координат центров кластеров, что гарантирует сходимость данного алгоритма за конечное число шагов. В статье [7] приводится обоб-

[1]Работа выполнена при поддержке РФФИ, проект 16-07-01160.

Табл.1. Основные типы алгоритмов текстовой кластеризации

Тип модели	Документ	Тема	Пример алгоритма
Жесткие	Вектор	Вектор	k-means [2]
Описательно-вероятностные	Вектор	Вероятность	DPM [3]
Смеси моделей	Вероятность	Вектор	mixture of Gaussian, vMF [4]
Вероятностные	Вероятность	Вероятность	LDA [5]

щенный вариант данного алгоритма, адаптирующий шаг пересчета координат центров кластеров для произвольной функции расстояния.

Рекомендуется несколько раз запустить данный алгоритм с различными начальными условиями. В статье [8] предлагается вместо того, чтобы присваивать каждому документу метку ближайшего кластера, некоторым документам присваивать метки случайных кластеров.

Для построения одномерной кластеризации требуется выбрать функцию расстояния векторов документов, при помощи которой документы попарно сравниваются и объединяются в кластеры. Функцией расстояния между документами является метрика Минковского. В [9] рассматривается способ применения взвешенных метрик Минковского в качестве функции расстояния.

В [8] функцией расстояния является взвешенная косинусная мера. При ее использовании расстояние не зависит от числа терминов в документах, и небольшое число элементов имеет вклад в ее значение. Недостатком этой функции расстояния является то, что она не является метрикой, так как не выполняется неравенство треугольника.

Вероятностные модели. В отличии от жесткого подхода, в вероятностном подходе каждый документ может состоять из произвольного количества тем. Это удобно, например, в задаче текстового анализа новостей, где каждую новость можно отнести к нескольким темам.

В модели вероятностного латентного семантического анализа (PLSA) [10] максимизируется логарифм правдоподобия коллекции. Недостатками данного подхода является наличие большого числа параметров [10], что приводит к переобучаемости. В [11] предлагается ввести различные варианты регуляризации для предотвращения переобучаемости. Альтернативным вариантом уменьшения числа параметров является априорное предположение о виде распределений. Так, в [5] предполагается, что векторы документов и тем порождаются распределениями Дирихле.

Для построения иерархической кластерной структуры предложены [12] два типа алгоритмов: агломеративные и дивизимные.

В работе [13] предлагается использовать иерархический вариант вероятностного латентного семантического анализа, основанный на введении классов, к которым принадлежат несколько тем.

В работе [14] предлагается метод Probabilistic boosting-tree для построения иерархической структуры. Строится решающее дерево, в каждом узле которого условные вероятности рассчитываются по вероятностям принадлежности объекта к классу, рассчитанным в соответствующем поддереве. Недостатком метода является переобучение, особенно в случае малого размера обучающей выборки.

Смеси моделей, vMF. В данном методе документы описываются векторами. В качестве сходства документов в статье [4] рассматривается корреляция Пирсона. Чтобы описать документ используется распределение фон Мизеса Фишера (vMF) [15, 16]. Документ представляется, как смесь кластеров в определенной пропорции. Для определения параметров данного алгоритма максимизируется логарифм правдоподобия выбранной коллекции документов с помощью EM алгоритма.

Вероятностно-описательные модели. В моделях данного типа документ представляется в виде вектора, как в жестких моделях, но принадлежит одновременно нескольким темам, как в вероятностных моделях. Алгоритм DPM показывает результаты по крайней мере не хуже (а в некоторых случаях лучше) вероятностного подхода [3].

В других работах [17-20] уже предлагались решения проблемы построения тематической модели крупной конференции. Однако ставились

другие задачи. В [17] решалась задача верификации модели. Использовалась готовая структура конференции и решалась задача оптимизации данной структуры путем перемещения небольшого числа документов в более подходящие кластеры. В работе [18] предлагается алгоритм выбора оптимальной меры сходства между документами при классификации новых документов. В работе [19] авторы предлагают использовать косинусную меру для оценки близости между документами и оценивать важность слов в каждом документе. С помощью метода, описанного в [20], строилась иерархическая модель конференции, однако метод не был приспособлен для классификации новых документов.

В данной работе мы предлагаем метод для построения иерархической структуры крупной конференции, который будет работать для нового набора документов. Для построения иерархической структуры требуется не вся экспертная модель, а лишь часть размеченных документов. Исследуется качество построения разделяющей вероятностной модели, которая ранее не применялась в данной задаче.

Требуется построить сбалансированную структуру конференции. Количество документов в разных научных областях и направлениях должно быть примерно одинаково. Особенностью DPM является то, что величина кластера влияет на вероятность принадлежности документа кластеру [3]. В результате многие документы оказываются в крупных кластерах, что нарушает требование сбалансированности.

Для кластеризации коллекции тезисов конференции EURO предлагается адаптировать модель DPM [3], снизив влияние размера кластера. Для иерархической кластеризации предлагается использовать дивизимный алгоритм. В работе адаптированные модели сравниваются с оригинальной моделью DPM. Также сравнивается качество иерархической и плоской кластеризации.

Для кластеризации документов проводится их предварительная обработка [21]. Слова приводятся к начальной лексической форме. Это позволяет уменьшить общее количество слов в словаре и основано на предположении, что форма слова не является определяющим признаком документа, в котором оно использовано. В [22] приведен код и подробное описание этой процедуры, а также обосновано использование лемматизации, а не стемминга. Используется критерий *TF-IDF* и словарь стоп-слов для отсева слов, встречающихся редко, а также слов, встречающихся в большинстве документов [22]. Документы представляются в виде “мешков слов” и каждому документу ставится в соответствие целочисленный вектор. Для предобработки признаков можно использовать различные варианты представления TF-IDF или BM25. Данные варианты сравнивались в работе [23]. Сравнение качества кластеризации проводилось на выборке из 1342 тезисов конференции EURO [1] и на выборке сайтов индустриального сектора.

1. Постановка задачи

Задан словарь $W = \{w_1, \ldots, w_n\}$, n - количество слов в словаре. Задана коллекция документов D Документом d из коллекции назовем неупорядоченное множество слов из W, $d = \{w_j\}$, где $j \in \{1, \ldots, n\}$.

Представим документ d_s в виде вектора $\mathbf{x}_s$ размерности n следующим образом: если слово w_j из словаря W встретилось в документе d_s k раз, то $x_{s,j} = k$, $k \geq 0$. Обозначим количество документов в коллекции через N. Получим матрицу $\mathbf{X}$ объект-признак, где каждая строка является описанием документа d_s

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

На Рис. 1 показана экспертная иерархическая модель конференции в виде дерева, с корнем — названием конференции. Уровнем l иерархии назовем множество всех узлов дерева, находящихся на глубине l. Каждый внутренний узел дерева обозначим c_{li}, где l — уровень, к которому принадлежит данный узел, а i — номер этого узла среди узлов на данной глубине. Документы являются листами этого дерева и имеют уровень четыре.

Документ d_s принадлежит кластеру, соответствующему узлу c_{li}, если путь к данному документу от вершины проходит через узел c_{li}. Обозначим c_{li} множество дочерних элементов данного узла. Множество кластеров нижнего уровня обозначим через C.

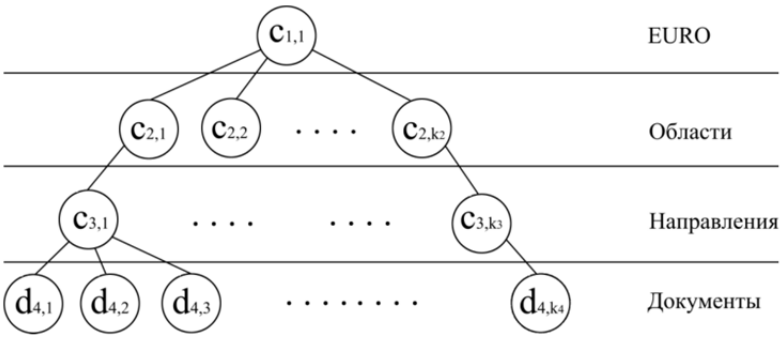


Рис. 1. Иерархическое представление тематической модели

Моделью назовем отображение множества документов во множество C кластеров нижнего уровня:

$$\mathbf{f}: D \mapsto C. \quad (1)$$

Экспертной назовем модель $\tilde{\mathbf{f}}$, составленную экспертами.

Используется функционал качества - доля документов, кластеризация которых совпадает с экспертной на уровне иерархии l:

$$S_l = \frac{1}{k_l}\sum_{i=1}^{k_l}[c_{li} = \tilde{c}_{li}], \quad (2)$$

где k_l - количество кластеров на l-ом уровне иерархии, $\tilde{c}_{li}$ - экспертный кластер для документа d_i.

Выражение $[c_{li} = \tilde{c}_{li}]$ равно 1, если кластер, которому отнесен документ, совпадает с экспертным, и равно 0 в противном случае. Функционал S показывает, насколько построенная модель $\mathbf{f}$ совпадает с экспертной $\tilde{\mathbf{f}}$.

Функционал качества (2) используется для кластеризации на каждом отдельном уровне иерархии, а также значение функционала на нижнем уровне ($l = h$) для оценки качества кластеризации с помощью иерархического алгоритма.

Требуется построить модель $\mathbf{f}$, при этом максимизировать функционал качества кластеризации $S \to \max$.

2. Описание алгоритма

Для кластеризации документов предлагается использовать разделяющую вероятностную модель, а также две ее адаптации: нормализованную разделяющую вероятностную модель (nDMP) и логарифмическую разделяющую вероятностную модель (lDMP).

DPM. Предполагается: все слова делятся на информативные и неинформативные, неинформативные слова не влияют на тему документа. Обозначим W - полный словарь слов w, $F \subseteq W$ - словарь информативных слов. Предполагается, что для каждого документа вероятность встретить неинформативное слово w одинакова:

$$\sum_{w \in W \setminus F} p(w \mid \mathbf{x}) = Z.$$

Тогда документ $\mathbf{x}$ принадлежит кластеру c_{li} с вероятностью

$$P(c_{li} \mid \mathbf{x}) = \frac{1}{1-Z}\sum_{w \in F} P(c_{li} \mid w) P(w \mid \mathbf{x}).$$

Введем обозначения для величин TF' и IDF':

$$\mathrm{TF}'(w_i, d) = \frac{x_i}{\#\mathbf{x}},\ \mathrm{IDF}'(w_i) = \sqrt{\frac{N}{\sum_{\mathbf{x} \in D}\frac{x_i}{\#\mathbf{x}}}}, \quad (3)$$

где N - число документов в коллекции (1), $\#\mathbf{x}$ - число слов в документе d. Отбор информативных слов осуществляется с помощью представления $TF'\text{-}IDF'$ (3). Традиционно, IDF описывает инверсию частоты, с которой некоторое слово встречается в документах коллекции. В нашем случае мы учитываем частоту слова $\frac{x_i}{\#\mathbf{x}}$ для определения IDF'. При этом редкое типичное слово (слово, которое встречается редко, но во многих документах) оправдано будет иметь высокое значение IDF', в отличии от низкого значения в случае традиционного IDF. Учитывая, что

информативные слова часто являются редкими типичными словами, наша модель (1) может эффективно повысить веса таких слов. В работе [24] делается подобный вывод при описании метрики Римана. В работе показано, что метрика Римана превосходит традиционный TF' - IDF' в задаче классификации текстов.

Перейдя к новому представлению документа d в виде вектора $\mathbf{x}$ с компонентами $x_i = \mathrm{TF}'(x_i, d) \cdot \mathrm{IDF}'(x_i)$ получим:

$$P(c_{li} \mid \mathbf{x}) = \frac{1}{1-Z} \cdot \frac{N(c_{li})}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{li}, \text{ где } \mathbf{c}_{li} = \frac{1}{N(c_{li})} \cdot \sum_{\mathbf{x}' \in c_{li}} \mathbf{x}'. \quad (4)$$

Для построения модели $\mathbf{f}$ предлагается использовать алгоритм кластеризации с частичным обучением. Предлагается использовать формулу (4) на каждой итерации для локальной оптимизации функционал (2). При этом выборку (1) делим на обучающую $\mathbf{X}_1$, для которой значение кластеров считаем известными, и контрольную $\mathbf{X}_2$. В начальном приближении для документов $\mathbf{x} \in \mathbf{X}_1$ полагаем

$$P(c_{li} \mid \mathbf{x}) = \begin{cases} 1 & \text{при } \mathrm{c}_{li} = \tilde{\mathrm{c}}_{li}, \\ 0 & \text{при } \mathrm{c}_{li} \neq \tilde{\mathrm{c}}_{li}. \end{cases}$$

Пересчитываем центры кластеров

$$\mathbf{c}_{li} = \frac{1}{N(c_{li})} \cdot \sum_{\mathbf{x}' \in c_{li}} \mathbf{x}'.$$

Рассчитываем новые вероятности $P^{\mathrm{new}}(c_{li} \mid \mathbf{x})$ по старым $P(c_{li} \mid \mathbf{x})$ и $\mathbf{c}_{li}$:

$$P^{\mathrm{new}}(c_{li} \mid \mathbf{x}) = \frac{1}{1-Z} \cdot \frac{N(c_{li})}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{li},$$

$$\text{где} \qquad N(c_{li}) = \sum_{x \in c_{li}} P(c_{li} \mid \mathbf{x}).$$

Присваиваем вероятности для документов из обучающей выборки по правилу

$$P(c_{li} \mid \mathbf{x}) = \begin{cases} 1 & \text{при } \mathrm{c}_{li} = \tilde{\mathrm{c}}_{li}, \\ 0 & \text{при } \mathrm{c}_{li} \neq \tilde{\mathrm{c}}_{li}. \end{cases}$$

Продолжаем итерации до тех пор, пока функционал качества (2) растет.

lDPM и nDPM. Для кластеризации документов предлагается также использовать модели lDMP и nDMP. Они отличаются от DPM изменением формулы (4). В этих случаях уменьшается влияние величины кластера $N(c_{li})$ на вероятность $P(c_{li} \mid \mathbf{x})$, что позволяет учесть сбалансированность структуры конференции. Для этого в lDMP предлагается использовать значение $\ln N(c_{li})$ вместо $N(c_{li})$ в формуле (4), т.е. используется формула:

$$P^{\mathrm{new}}(c_{li} \mid \mathbf{x}) = \frac{1}{1-Z} \cdot \frac{\ln N(c_{li})}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{li},$$

в алгоритме nDPM значение $P(c_{li} \mid \mathbf{x})$ нормируется на величину $N(c_{li})$. Используется формула:

$$P^{\mathrm{new}}(c_{li} \mid \mathbf{x}) = \frac{1}{1-Z} \cdot \frac{1}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{li}.$$

hDPM. Для кластеризации документов на низких ($l > 2$) уровнях иерархии предлагается дивизимный алгоритм иерархической кластеризации hierarhal DPM (hDPM). Сначала документы кластеризуются на втором уровне иерархии ($l = 2$) с помощью плоской модели DPM. При этом все документы распределяются по кластерам второго уровня. После чего для каждого кластера снова запускается алгоритм для построения плоской модели уже на третьем уровне и так далее. Таким образом строится иерархическая модель с заданным количеством уровней.

3. Дополнительные критерии качества

Результаты работы могут быть использованы экспертами для распределения докладов по структуре конференции. Важно, чтобы предсказанный класс документа был как можно ближе к экспертному. Для численной оценки этой характеристики введем дополнительные критерии качества.

Кроме функционала (2) для оценки качества модели используются критерий качества оператора релевантности $Q(R)$ и площадь под верхней огибающей кумулятивной гистограммы AUC(R). Обозначим P^{k_l} - множество перестановок порядка k_l. Определим оператор релевантности

$$R : \mathsf{R}^n \to P^{k_l},$$

ставящий в соответствие описанию документа $\mathbf{x} \in \mathsf{R}^n$, перестановку кластеров уровня l. При

этом кластеры c_{li} отсортированы по убыванию вероятности (4) того, что документ $\mathbf{x}$ принадлежит данному кластеру. Кластер c_{li} является наиболее релевантным для документа $\mathbf{x}$ относительно оператора релевантности R, если номер i данного кластера стоит на первом месте в перестановке, возвращаемой R.

Определим качество оператора релевантности R как среднюю позицию экспертного кластера $\tilde{c}_{li}$ для документа $\mathbf{x}_j$ в перестановке $R(x_j)$

$$Q(R) = \frac{1}{|D|}\sum_{j=1}^{|D|}\text{pos}(R(\mathbf{x}_j), \tilde{c}_{li}).$$

Чем меньше значение $Q(R)$, тем выше вероятности принадлежности документов $\mathbf{x}_j$ экспертным кластерам, в сравнении с не экспертными и тем меньше их позиции в перестановке, которую возвращает предложенный оператор релевантности R.

Для каждого уровня иерархии построим кумулятивную гистограмму следующим образом: пусть столбец гистограммы с номером i принимает значение

$$\#\text{pos}(R(\mathbf{x}_j), \tilde{c}_{li}) \le i,$$

где $\text{pos}(R(\mathbf{x}_j), \tilde{c}_{li})$ - множество всех документов, для которых номер позиции экспертного кластера меньше либо равен i в перестановке, возвращаемой R, а $\#\{\cdot\}$ - число элементов во множестве $\{\cdot\}$.

Альтернативным критерием качества служит AUC(R) - площадь под верхней огибающей кумулятивной гистограммы:

$$AUC(R) = \frac{1}{k_l |D|}\sum_{i=1}^{k_l}\#\text{pos}(R(\mathbf{x}_j), \tilde{c}_{li}) \le i.$$

$\text{AUC(R)} = 1$ соответствует случаю, когда экспертный кластер оказы вается в соответствии с R наиболее релевантным для каждого из документов коллекции D.

4. Вычислительный эксперимент

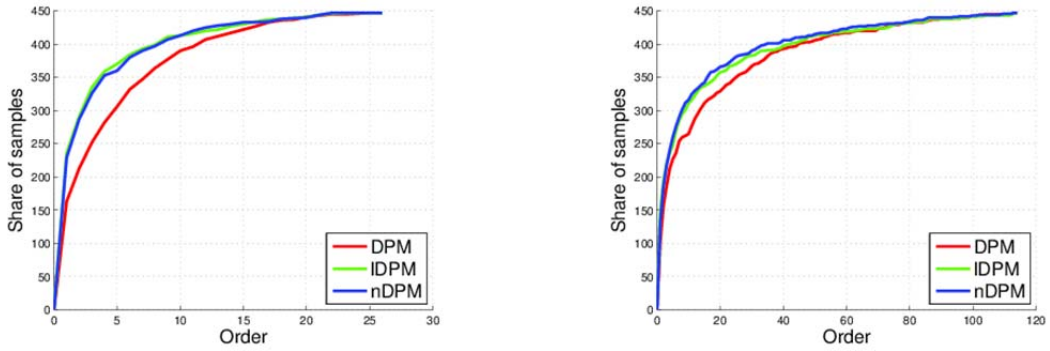
Для сравнения предложенных алгоритмов с оригинальным DPM проводилась кластеризация документов научной конференции EURO. В качестве исходных данных был взят набор из 1342 тезисов данной конференции и ее экспертная модель. В экспертной модели каждому тезису сопоставлена одна научная область и одно направление. После предобработки документов был получен словарь объемом $n = 3479$ слов.

Сравнивались модели lDPM и nDPM с моделью DPM на втором (26 областей) и третьем (114 направлений) уровнях иерархии. При этом вся выборка делилась на обучающую и тестовую в соотношении 2:1. Таким образом, объем обучающей выборки составил 895 тезисов. Для тезисов из обучающей выборки научная область и научное направление считались известными и брались из экспертной модели.

Сравнение моделей представлено в Табл. 2. Особенностью модели DPM является то, что большие кластеры притягивают документы (4) и начиная с какой-то итерации большинство документов попадают в крупные кластеры. Модели lDPM и nDPM менее чувствительны к мощности кластеров и с их помощью удалось получить лучшие показатели (Табл. 2). При этом наибольший процент правильно кластеризованных документов $S = 53.02\%$ и $S = 32,44\%$ для второго и третьего уровней иерархии соответственно был получен при использовании модели lDPM.

Табл. 2. Сравнение функционалов качества для алгоритмов DPM, lDPM и nDPM

Модель	DPM		lDPM		nDPM	
Критерий качества	область	направление	область	направление	область	направление
S	36,7%	22,8%	**53,0 %**	**32,4%**	51,5%	29,5%
Q	4,79	16,17	**3,35**	13,89	3,42	**12,79**
AUC	0,854	0,867	**0,910**	0,887	0,907	**0,897**



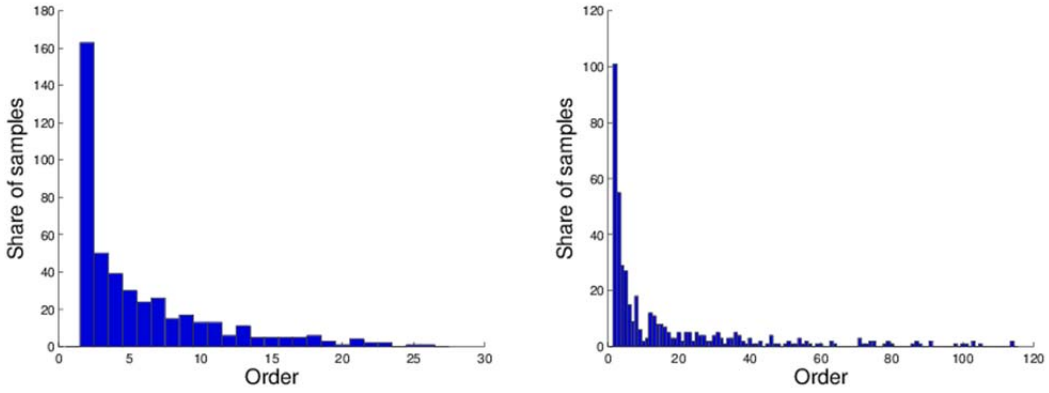
а) б)

Рис. 2. Верхние огибающие AUC(R): а) – области; б) – направления

На графиках огибающих кумулятивных гистограмм (Рис. 2) для кластеризаций на втором и третьем уровнях иерархии по осям абсцисс и ординат отложен номер i кластера и количество документов, для которых их экспертный кластер занимает место $\le i$ по релевантности, соответственно. Кривая для DPM проходит ниже соответствующих кривых для lDPM и nDPM. Следовательно, по показателю AUC(R) модели lDPM и nDPM превосходят DPM.

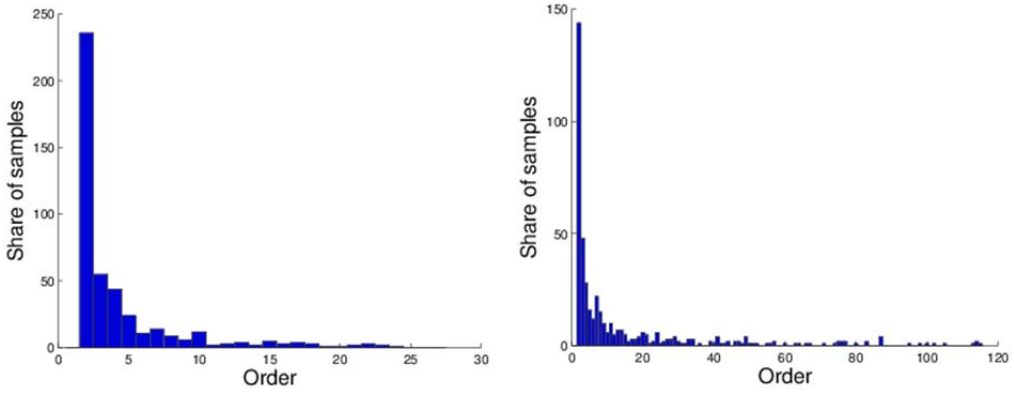
Приведем также гистограммы распределения $\#\text{pos}(R(\mathbf{x}_j),\tilde{c}_{li}) = i$, показывающие количество объектов, для которых их экспертный кластер занимает i-ое место по релевантности по оператору R (Рис. 3-Рис. 5). Из гистограмм видно, что в случае использования lDPM и nDPM экспертный кластер оказывается наиболее релевантным в большем числе случаев, чем при использовании модели DPM. Важным показателем также является процент документов, для которых экспертный кластер оказался в числе первых по релевантности, пусть даже и не самым релевантным. Например, экспертная область оказалась в первой пятерке по релевантности для 70,1%, 83,1% и 81,6% документов при использовании моделей DPM, lDPM и nDPM соответственно. Эти документы были распределены в область, совпадающую с экспертной, или в близкую к ней.

Модель lDPM использовалась для дивизимного иерархического алгоритма кластеризации hDPM. С помощью модели lDPM проводилась кластеризация на втором уровне иерархии (26 областей), а затем отдельно для документов, отнесенных к одной и той же области для определения кластеров на уровне направлений. При этом информация об экспертном направлении считалась известной для документов из обучающей выборки $\mathbf{X}_l$. В этом случае размер обучающей выборки также составлял 895 тезисов. Заметим, что данный алгоритм позволил получить небольшое улучшение в проценте пра-



а) б)

Рис. 3. Распределение документов по релевантности их экспертного кластера для DPM: а) – области; б) – направления



а) б)

Рис. 4. Распределение документов по релевантности их экспертного кластера для lDPM:

а) – области; б) – направления

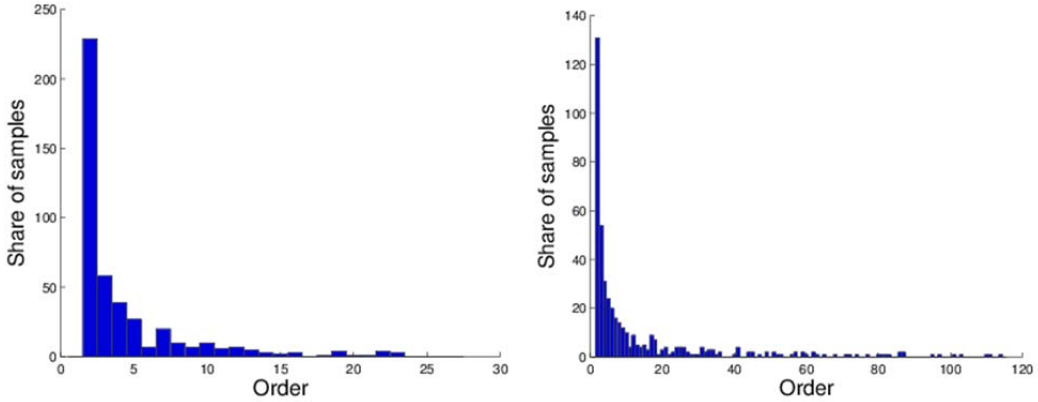


Рис. 5. Распределение документов по релевантности их экспертного кластера для nDPM:
а) – области; б) – направления

Табл.3. Процент правильно кластеризованных документов в иерархической модели

Модель	lDPM	hDPM
S	32,4%	**32,9%**

Табл. 4. Сравнение функционалов качества для алгоритмов DPM, lDPM и nDPM

Модель	DPM		lDPM		nDPM	
Критерий качества	область	направление	область	направление	Область	направление
S	29,3%	24,2%	**65,4 %**	**49,1%**	64,8%	48,1%

вильно кластеризованных документов на третьем уровне иерархии (Табл. 3).

Работа предложенных алгоритмов также проверялась на наборе сайтов индустриального сектора. В качестве исходных данных был взят набор из 1076 сайтов. Каждый из них был отнесен экспертами к одной из 11 областей и одному из 78 направлений. Для представления сайта в виде текстового документа из него удалялись все специальные тэги разметки, а все его страницы объединялись в одну. После предобработки полученных документов был получен словарь объемом $n = 20278$ слов.

Использовались модели DPM, lDPM и nDPM на втором (11 областей) и третьем (77 направлений) уровнях иерархии. При этом вся выборка делилась на обучающую и тестовую в соотношении 2:1. Для сайтов из обучающей

выборки область и направление считались известными.

Построенные модели обладают лучшим качеством (Табл.4), чем модели для научной конференции, что можно объяснить меньшим количеством областей и направлений.

Заключение

В данной работе решалась задача построения иерархической тематической модели крупной конференции. Модель DPM была адаптирована для кластеризации тезисов конференции. В новых моделях lDPM и nDPM снижено влияние мощности кластеров на построение тематической модели. На основании плоских моделей построен дивизимный иерархический алгоритм кластеризации. Качество кластеризации сравнивалось на коллекции тезисов конференции EURO и коллекции сайтов индустриального сектора. Вычислительный эксперимент показал, что модели lDPM и nDPM превосходят базовую модель DPM по проценту правильно кластеризованных документов, релевантности экспертного кластера и по значению AUC. Иерархический алгоритм с использованием модели lDPM также улучшил качество кластеризации.

Предложенный алгоритм может быть использован организаторами крупных конференций для автоматического распределения неразмеченных докладов по иерархической структуре. Алгоритм можно использовать для первой итерации распределения докладов. В этом случае экспертам будет необходимо выбрать окончательную научную область не из всего списка, а лишь из малого количества близких областей, что позволит значительно сократить время.

Литература

1. Тезисы конференции EURO. URL: https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/EURO_data/дата обращения: 27.06.2016.
2. Hartigan J.A., Wong M. A. Algorithm AS 136: A k-means clustering algorithm // Applied Statistics. 1979. Vol. 28, no. 1. Pp. 100–108.
3. Qi He, Kuiyu Chang, Ee-Peng Lim, Arindam Banerjee Keep It Simple with Time: A Reexamination of Probabilistic Topic Detection Models // IEEE Transactions on Pattern Analysis & Machine Intelligence, vol.32, no. 10, pp. 1795-1808, October 2010, doi:10.1109/TPAMI.2009.203.
4. Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Suvrit Sra Generative Model-based Clustering of Directional Data // Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, NY, USA: ACM, 2003. Pp. 19–28.
5. David M. Blei, Andrew Y. Ng, Michael I. Jordan Latent dirichlet allocation // The Journal of Machine Learning Research. Vol. 3, 2003. Pp. 993-1022.
6. Ackermann Marcel R., Blomer Johannes, Sohler Christian. Clustering for Metric and Nonmetric Distance Measures // ACM Trans. Algorithms. 2010. Vol. 6, no. 4. Pp. 1:59. http://doi.acm.org/10.1145/1824777.1824779.
7. Hand DJ, Krzanowski WJ. Optimising k-means clustering results with standard software packages // Computational statistics and Data analysis. 2005. Vol. 49. Pp. 969–973.
8. Leisch Friedrich. A Toolbox for K-centroids Cluster Analysis // Comput. Stat. Data Analysis. 2006. Vol. 51, no. 2. Pp. 526–544.
9. Yih Wen-tau. Learning Term-weighting Functions for Similarity Measures // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2, EMNLP '09 Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. Pp. 793–802. http://dl.acm.org/citation.cfm?id=1699571.1699616.
10. Hofmann Thomas. Probabilistic Latent Semantic Indexing // Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. SIGIR '99. New York, NY, USA: ACM, 1999. Pp. 50–57.
11. Vorontsov Konstantin, Potapenko Anna, Plavin Alexander. Additive Regularization of Topic Models for Topic Selection and Sparse Factorization // Statistical Learning and Data Sciences / edited byAlexander Gammerman, Vladimir Vovk, Harris Papadopoulos. Springer International Publishing, 2015. Vol. 9047 of Lecture Notes in Computer Science. Pp. 193–202.
12. Hao Pei-Yi, Chiang Jung-Hsien, Tu Yi-Kun. Hierarchically SVM classification based on support vector clustering method and its application to document categorization // Expert Systems with Applications. 2007. Vol. 33, no. 3. Pp. 627–635.
13. Eric Gaussier, Cyril Goutte, Kris Popat, Francine Chen. A Hierarchical Model for Clustering and Categorising Documents // Advances in Information Retrieval Proceedings of the 24th BCS-IRSG European Colloquium on IR Research (ECIR-02), 2002.
14. Tu Z. Probabilistic boosting-tree: Learning discriminative models for classification, recognition, and clustering // Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on. – IEEE, 2005. – T. 2. – C. 1589-1596.
15. Dhillon Inderjit S., Sra Suvrit. Modeling Data using Directional Distributions: Tech. Rep. TR-03-06: The University of Texas, Department of Computer Sciences, 2003. January.
16. Kuzmin A.A., Aduenko A.A., Strijov V.V. Thematic Classification for EURO/IFORS Conference Using Expert Model // Conference of the International Federation of Operational Research Societies, 2014.
17. Адуенко А.А., Кузьмин 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.

18. Aduenko A. A., Kuzmin A. A., Strijov V. V. Adaptive thematic forecasting of major conference proceedings May 4, 2014.
19. Kuznetsov M.P., Clasel M., Amini M.-R., Gaussier E., Strijov V.V. Supervised topic classification for modeling a hierarchical conference structure // in S. Arik et al. (Eds.): International conference on neural information processing, Part 1, LNCS, 2015, 9489 : 90–97.
20. Kuzmin A.A., Aduenko A.A., Strijov V.V. Thematic Classification for EURO/IFORS Conference Using Expert Model // Conference of the International Federation of Operational Research Societies, 2014 : 175.
21. Кузьмин А.А. Предобработка тезисов конференции EURO. https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/EURO_data/Data%20preparation/ дата обращения: 27.06.2016.
22. Mr. V.K. Bhalla Deepika Sharma, Dr. Deepak Garg. Improved stemming approach used for text processing in infirmation retrieval // Computer science and engineering department Thapar University Patiala – 147004. 2012.
23. Puurula A., Read J., Bifet A. Kaggle LSHTC4 winning solution // arXiv preprint arXiv:1405.0546. – 2014.
24. Lebanon G. Learning Riemannian Metrics // Proc. of the 19thConference on Uncertainty in Artificial Intelligence 2003.

Златов Александр Сергеевич. Студент Московского физико-технического института. Область научных интересов: машинное обучение, тематическое моделирование. E-mail: zlatov_a_s@mail.ru

Кузьмин Арсентий Александрович. Финансовая группа Лайф, аналитик отдела математического моделирования. Окончил Московский физико-технический институт в 2015 году. Автор 4 печатных работ. Область научных интересов: машинное обучение, тематическое моделирование, кредитный скоринг. E-mail: arsentii.kuzmin@gmail.com