Построение иерархической тематической модели крупной конференции

Златов А. С., Кузьмин А. А.

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. На основании плоских моделей строится иерархическая тематическая модель конференции. Для построения тематической модели тезисов конференции используется дивизимный иерархический алгоритм. Работа алгоритмов проиллюстрирована на коллекции тезисов конференции EURO-2012. Разделяющая вероятностная модель сравнивается с адаптированными моделями и иерархической моделью. Для оценки качества тематической моедли используется модель этой же конференции, построенная экспертами.

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

Thematic model of major conference proceedings

Aleksander Zlatov, Arsentii Kuzmin

Annotation

The aim of this paper is to construct a hierarchical thematic model for abstracts of a major conference. We propose to adapt Discriminative Probabilistic Model to the balanced structure of

the conference for document clustering at each level of hiererchical structure. Semi-supervised learning is used for document clustering. We also propose the hierarchal divisive clustering algorithm to construct the hierarchical thematic model. The hierarchical model is based on models for each level of hiererchical structure. The algorithms are applied to the collection of abstructs of conference EURO-2012. The constructed model is compered with the thematic model of EURO-2012 constructed by experts.

Keywords: *thematic model, hiererchical model, probabilistic thematic model, document clustering, semi-supervised learning.*

1 Введение

Ежегодно программный комитет крупной конференции решает задачу построения иерархической модели тезисов конференции. Рассмотрим такую модель на примере конференции EURO 2012 [5]. Конференция содержит в себе 26 главных областей. Каждая область содержит в себе $10-15$ научных направлений. Участники конференции для подачи заявки присылают тезисы документов, состоящие из не более, чем 600 символов, и выбирают три ключевых слова, наиболее связанных по их мнению с тематикой работы. На основании содержания документов и ключевых слов эксперты распределяют работы по иерархической структуре.

В крупной конференции обычно более 2000 докладов. Доклады делятся на два типа. Первый тип – приглашенные доклады, для которых заранее известны научная область, направление и так далее. Второй тип – неразмеченные доклады, которые нужно распределить по иерархической структуре. Докладов двух типов примерно равное количество, то есть, более 1000 докладов требуется распредилять вручную. Для этого привлекается до 200 экспертов из различных областей. В среднем, на каждый доклад эксперт тратит 2 дня. Автоматическое распределение неразмеченных докладов позволило бы значительно сократить затраченное время. Создание такой автоматической ситсемы и является целью данной работы.

Ранее для кластеризации текстов использовались плоские и иерархические методы. Из плоских методов использовались метрические: K-means [1, 9], FOREL [12], C-means [7], STOLP [13], а также вероятностные методы [10], например, вероятностный латентный семантический анализ [2] или латентное размещение Дирихле [6]. Используются также смеси метрических и вероятностных методов, например, Разделяющая вероятностная модель (англ. Discriminative Probabilistic Model, DPM) [4]. Для построения иерархической кластерной структуры существует два типа алгоритмов [11]: агломеративные и дивизимные.

В работе [32] прадлагается использовать иерархический вариант вероятностного латентгого семантического анализа, основанный на введении классов, к которым принадлежат сразу несколько тем. На тестовых коллекциях документов показано, что иерархическая модель превосходит плоскую.

В работе [31] предлагается метод Вероятностное бустинг-дерево (англ. Probabilistic boosting-tree) для построения иерархической структуры. Строится решающее дерево, в каждом узле которого условные вероятности расчитываются по вероятностям принадлежности объекта к классу, расчитанным в соответствующем поддереве. Недостатком метода является переобучение, особенно в случае малого размера обучающей выборки.

В других работах уже предлагались решения проблемы построения тематической модели крупной конференции. Однако, ставились другие задачи. В работе [33] решалась задача верификации модели. Использовалась готовая структура конференции и решалась задача перемещения документов для улучшения модели. Предлагался алгоритм выбора оптимальной меры сходства между документами [36]. В работе [35] авторы предлагают использовать косинусную меру для оценки близости между документами и оценивать важность слов в каждом документе. Последний метод находится в состоянии доработки. С помощью другого метода, описанного в [34], строилась иерархическая модель конференции, однако, метод не приспособлен для классификации новых документов.

В данной работе, в отличии от предыдущих, мы предлагаем метод для построения

иерархической структуры крупной конференции, который будет работать для нового набора документов. Помимо этого, исследуется качество построения Разделяющей вероятностной модели, которая ранее не применялась в данной задаче.

В моделе DPM документ представляется в виде вектора. Однако, в отличии от метрических методов, эта модель ставит документ в соответствие не одной, а сразу нескольким темам. При этом каждый документ принадлежит темам с разными вероятностями, как в вероятностных методах. Модель DPM работает не хуже (в некоторых случаях лучше) [4] вероятностных методов.

Предполагается, что структура конференции должна быть сбаланированной. То есть количество документов в разных научных областях и направлениях должно быть примерно одинаково. Особенностью DPM является то, что величина кластера влияет на вероятность документа принадлежать этому кластеру [4]. В результате многие документы оказываются в крупных кластерах, что не согласуется с идеей сбалансированности структуры конференции. Для кластеризации коллекции тезисов конференции EURO 2012 предлагается адаптировать модель DPM [4], снизив роль мощности кластера. Для иерархической кластеризации предлагается использовать дивизимный алгоритм. В работе адаптированные модели сравниваются с оригинальной моделью DPM. Также сравнивается качество иерархической и плоской кластеризации.

Сравнение качества кластеризации проводилось на выборке из 1342 тезисов конференции EURO 2012 [5]. Для кластеризации документов проводится их предварительная предобработка [8]. Слова приводятся к начальной лексической форме. Это позволяет уменьшить общее количество слов в словаре и основано на предположении, что форма слов не влияет на его смысл и не является определяющим признаком тезиса, в котором оно использовано. Наряду с лемматизацией часто используется стемминг, при котором слова приводятся не к начальной форме, а от них оставляют только основу. Эксперимент с данными конференции показал, что стемминг почти никак не влияет на количество слов в словаре. Поэтому, решено было использовать только лемматизацию. В

[29] приведен код и подробное описание этой процедуры. Используется критерий TF-IDF и словарь стоп-слов для отсева слов, встречающихся малое количество раз, а также слов, встречающихся в большинстве документов [3]. 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 [30].

2 Постановка задачи

Пусть $W = \{w_1, \ldots, w_n\}$ — заданное множество слов. Назовем его словарем терминов, 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_s вектор $\mathbf{x}_s$ 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 $\#D$. Получим матрицу $\mathbf{X}$ объект-признак, где каждая строка является описанием документа d_s

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

Представим экспертную иерархическую модель в виде дерева (см. рис. 1), 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.

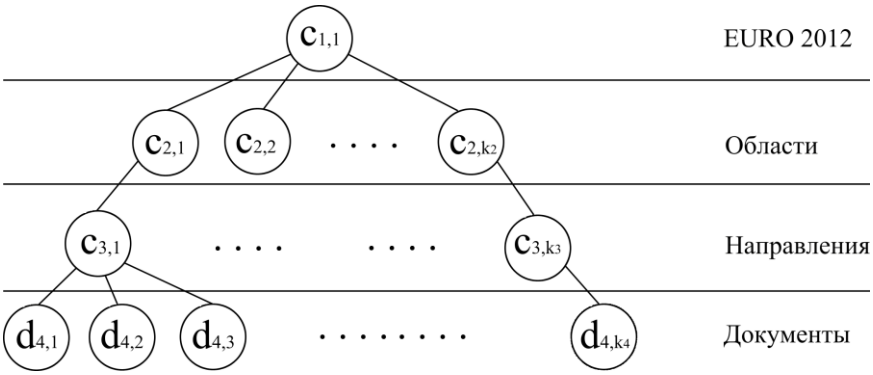


Рис. 1. Иерархическое представление тематической модели

Будем говорить, что документ d_s принадлежит кластеру, соответствующему узлу c_{li}, если путь к данному документу от вершины проходит через узел c_{li}. Под c_{li} будем понимать множество дочерних элементов данного узла. Множество кластеров нижнего уровня обозначим через C.

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

$$\mathbf{f} : D \mapsto \mathsf{C}.$$

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

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

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

где k_l — количество кластеров на l-ом уровне иерархии, $\tilde{c}_{li}$ — экспертный кластер для данного документа. Выражение $[c_{li} = \tilde{c}_{li}]$ равно 1, если кластер, к которому отнесен документ, совпадает с экспертным, и равно 0 в противном случае. Функционал S показывает, насколько построенная модель $\mathbf{f}$ совпадает с экспертной $\tilde{\mathbf{f}}$.

Будем использовать функционал качества (2) для кластеризации на каждом отдельном

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

Ставится задача построить модель $\mathbf{f}$, при этом максимизировать функционал качества кластреризации $S \to \max$.

3 Описание алгоритма

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

DPM.

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

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

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

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

Введём обозначения для величин TF' (от англ. tf —term frequency) и IDF' (от англ. idf —inverse document frequency):

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

где $N = \#D$ — число документов в коллекции (1), $\#\mathbf{x}$ — число слов в документе d. 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:

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

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 частоя являются редкими типичными словами, наша модель может эффективно повысить веса таких слов. В работе [28] делается подобный вывод при описании метрики Римана. В работе показано, что метрика Римана превосходит традиционный $TF' \text{-} IDF'$ в задаче классификации текстов.

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 всю 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

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

Шаг 1.

Пересчитываем центры кластеров

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

Шаг 2.

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

$$P^{\text{new}}(c_{li} \mid \mathbf{x}) = \frac{1}{1-R} \cdot \frac{N(c_{li})}{N} \mathbf{x}^{\mathrm{T}} \mathbf{c}_{li}, \qquad \text{где} \qquad N(c_{li}) = \sum_{x \in c_{li}} P(c_{li} \mid \mathbf{x}).$$

Шаг 3.

Присваиваем вероятности для документов из обучающей выборки по правилу

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

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) на Шаге 2. 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). То есть на Шаге 2 алгоритма lDPM используется формула

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

в алгоритме nDPM значение $P(c_{li} \mid \mathbf{x})$ нормируется на величину $N(c_{li})$. На Шаге 2 алгоритма nDPM используется формула

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

hDPM.

Для кластеризации документов на низких ($l > 2$) уровнях иерархии предлагается использовать дивизимный алгоритм иерархической кластеризации hierarhal DPM (hDPM). Сначала документы кластеризуются на втором уровне иерархии ($l = 2$) с помощью модели DPM. При этом все документы распределяются по кластерам второго уровня. После чего для каждого кластера запускается аналогичный алгоритм, разделяющий документы уже по кластерам третьего уровня и так далее.

4 Вычислительный эксперимент

Для проверки работы предложенных алгоритмов проводилась кластеризация документов научной конференции EURO- 2012 . В качестве исходных данных был взят набор из 1342 тезисов данной конференции и ее экспертная модель. В экспертной моделе каждому тезису сопоставлена одна научная область и одно направление.

После предобработки документов был получен словарь объемом $n = 3479$ слов.

Использовались модели DPM, lDPM и nDPM 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.

Для оценки качества построенной модели f кроме функционала (2) применим также 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

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

ставящий в соответствие описанию документа $\mathbf{x} \in \mathbb{R}^n$, перестановку кластеров уровня l. При этом кластеры c_{li} отсортированы по убыванию вероятности (4) того, что документ $\mathbf{x}$ принадлежит данному кластеру. Будем называть кластер c_{li} наиболее релевантным для документа $\mathbf{x}$ относительно оператора релевантности R, если номер i данного кластера стоит на первом месте в перестановке, возвращаемой R.

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

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

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

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

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

где $\mathrm{pos}(R(\mathbf{x}_j), \tilde{c}_{li})$ — 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:

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

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

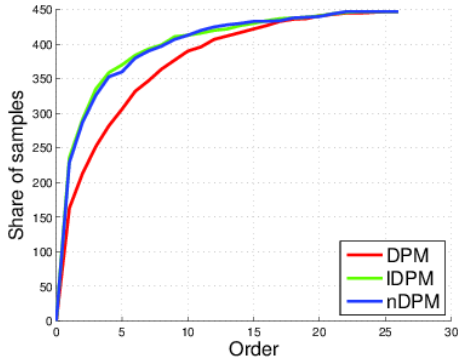
Сравнение моделей DPM, lDPM и nDPM представлены в табл. 0. Особенностью модели

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. 0). 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.

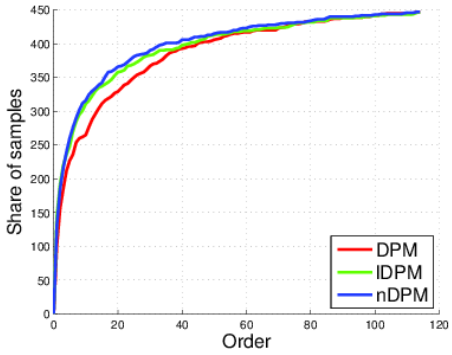
Таблица 1. Сравнение функционалов качества для алгоритмов DPM, lDPM и nDPM.

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**

На графиках огибающих кумулятивных гистограм, 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 $\leq i$ 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.



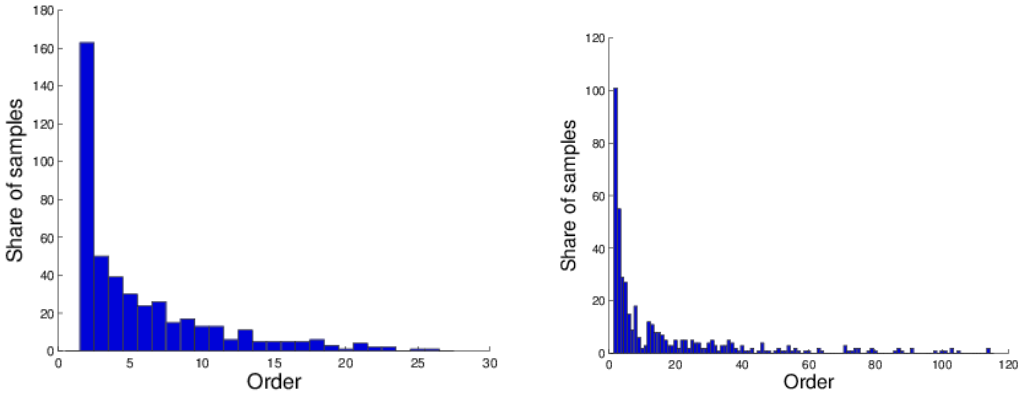
а) Области



б) Направления

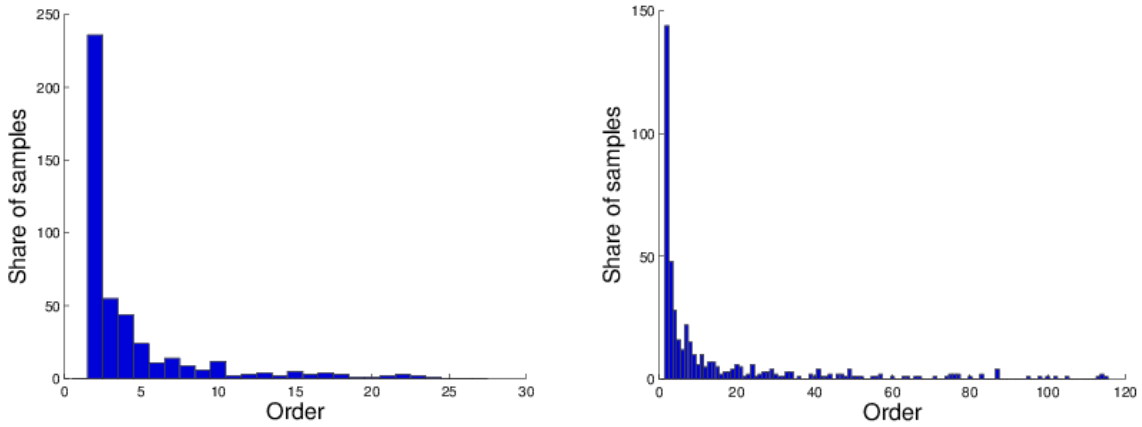
Рис. 2. Верхние огибающие AUC(R).

Приведем также гистограммы распределения $\#\mathrm{pos}(R(\mathbf{x}_j), \tilde{c}_{li}) = i$, показывающие количество объектов, для которых их экспертный кластер занимает i -ое место по релевантности по оператору R (см. рис. 3, 4, 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.



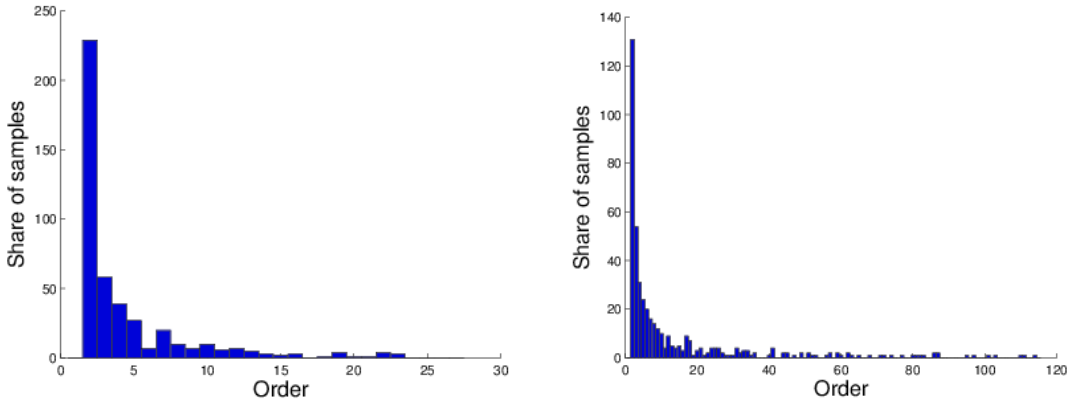
а) Области б) Направления

Рис. 3. Распределение документов по релевантности их экспертного кластера для DPM.



а) Области б) Направления

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.

Модель lDPM использовалась для дивизимного иерархического алгоритма кластеризации hDPM. С помощью модели lDPM проводилась кластеризация на 2-ом уровне иерархии (26 областей), а затем отдельно для документов, отнесённых 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 $\mathbf{X}_1$. В этом случае размер обучающей выборки также составлял 895 тезисов. Заметим, что данный алгоритм позволил получить небольшое улучшение в проценте правильно кластеризованных документов на третьем уровне иерархии (табл. 1).

Таблица 2. Процент правильно кластеризованных документов.

Модель	lDPM	hDPM
S	32,4%	**32,9%**

5 Заключение

В данной работе решается 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-2012 по проценту правильно кластеризованных документов, релевантности экспертного кластера и показателю AUC. Вычислительный эксперимент показал, что модели lDPM и nDPM превосходят базовую модель DPM. Иерархический алгоритм с использованием модели lDPM также незначительно улучшил качество кластеризации тезисов конференции.

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.

Список литературы

[1] J. A. Hartigan and M. A. Wong. Algorithm AS136. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[2] Thomas Hofmann. Probabilistic latent semantic indexing. pages 50–57.

[3] Mr. V. K. Bhalla Deepika Sharma, Dr. Deepak Garg. Improved stemming approach used for text processing in infirmation retrieval. 2012.

[4] Ee-Peng Lim Arindam Banerjee Qi He, Kuiyu Chang. Keep it simple with time: A re-examination of probabilistic topic detection models. 2009.

[5] Тезисы конференции euro-2012. https://sourceforge.net/p/mlalgorithms/code/head/tree/group274/zlatov2015conferencemodel/

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] Nikhil R. Pal and James C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE T. Fuzzy Systems*, 3(3):370–379, 1995.

[8] Стрижов В.В. Кузьмин, А. А. Тематическое моделирование. 2015.

[9] Richa Loohach and Kanwal Garg. Effect of distance functions on simple K-means clustering algorithm. (6/), 2012.

[10] Воронцов К.В. Вероятностное тематическое моделирование. 2013.

[11] Ланс Д. Н. Уиллиамс У. Т. *Методы иерархической классификации // Статистические методы для ЭВМ*. Наука, 1986.

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

[13] Н. Г. Загоруйко. *Прикладные методы анализа данных и знаний*. Новосибирск: ИМ СО РАН, 1999.

References

[1] J. A. Hartigan and M. A. Wong. Algorithm AS136. A K-means clustering algorithm. *Applied Statistics*, 28:100–108, 1979.

[2] Thomas Hofmann. Probabilistic latent semantic indexing. pages 50–57.

[3] Mr. V. K. Bhalla Deepika Sharma, Dr. Deepak Garg. Improved stemming approach used for text processing in infirmation retrieval. 2012.

[4] Ee-Peng Lim Arindam Banerjee Qi He, Kuiyu Chang. Keep it simple with time: A re-examination of probabilistic topic detection models. 2009.

[5] Documents of conference euro-2012. https://sourceforge.net/p/mlalgorithms/code/head/tree/group274/zlatov2015conferencemodel/

[6] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[7] Nikhil R. Pal and James C. Bezdek. On cluster validity for the fuzzy c-means model. *IEEE T. Fuzzy Systems*, 3(3):370–379, 1995.

[8] Strijov V.V. Kuzmin, A. A. Thematic modeling. 2015.

[9] Richa Loohach and Kanwal Garg. Effect of distance functions on simple K-means clustering algorithm. (6/), 2012.

[10] Vorontsov K.V. Probabilistic Thematic Modeling. 2013.

[11] G. N. Lance W. T. Williams *Hierarchical Clustering Method* Nauka, 1986.

[12] G. S. Lbov N. G. Zagoruyko V. N. Elkina *Алгоритмы обнаружения эмпирических закономерностей.* Novosibirsk: Nauka, 1985.

[13] N. G. Zagoruyko *Прикладные методы анализа данных и знаний*. Novosibirsk, 1999.

[28] Guy Lebanon, Learning Riemannian Metrics, In UAI'03.

[29] https://sourceforge.net/p/mlalgorithms/code/HEAD/tree/EURO_data/Data%20preparation/

[30] Puurula A., Read J., Bifet A. Kaggle LSHTC4 winning solution // arXiv preprint arXiv:1405.0546. – 2014.

[31] 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. – Т. 2. – С. 1589-1596.

[32] Eric Gaussier, Cyril Goutte, Kris Popat and Francine Chen, A Hierarchical Model for Clustering and Categorising Documents, in "Advances in Information Retrieval -- Proceedings of the 24th BCS-IRSG European Colloquium on IR Research (ECIR-02)", 2002.

[33] 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

[34]

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.

[35] Adaptive thematic forecasting of major conference proceedings A. A. Aduenko, A. A. Kuzmin, V. V Strijov May 4, 2014.

[36] 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 : 119-131. Article