Министерство образования и науки Российской Федерации
Московский физико-технический институт (государственный университет)
Факультет управления и прикладной математики
Вычислительный центр им. А. А. Дородницына РАН
Кафедра «Интеллектуальные системы»

Кузьмин Арсентий Александрович

Построение иерархических тематических моделей крупных конференций

010990 — Интеллектуальный анализ данных

Выпускная квалификационная работа магистра

Научный руководитель:
д. ф.-м. н. Стрижов Вадим Викторович

Москва
2015

Содержание

Аннотация

В данной работе предлагается метод построения иерархических тематических моделей крупных конференций. Для этого предлагается иерархическая взвешенная функция сходства документа и кластера нижнего уровня иерархии, учитывающая важность терминов в словаре коллекции. Вводится оператор релевантности, ранжирующий кластеры нижнего уровня иерархии по убыванию сходства с неразмеченным документом. С помощью предложенных методов строится тематическая модель коллекции аннотаций к докладам на конференции EURO 2010. Для обучения и настройки алгоритмов используются коллекции аннотаций к докладам EURO 2012 и EURO 2013 с экспертными моделями.

Ключевые слова: косинусная функция сходства, энтропия, тематические модели, иерархические модели.

Введение

Актуальность темы. В связи с быстрым увеличением объемов текстовой информации, в последнее время становятся все более востребованными методы тематического моделирования, позволяющие автоматически структурировать текстовые данные в виде иерархий тем и оптимизировать уже существующие, выявляя в них тематические несоответствия.

В работе исследуется фундаментальная проблема тематического моделирования – построение иерархических тематических моделей коллекций коротких документов. Данными коллекциями могут быть краткие описания фильмов, аннотации к научным работам, литературным произведениям или докладам на конференциях, новостные сводки или текстовые сообщения в социальных сетях. Присваивание темы каждому документу коллекции и объединение полученных тем в иерархическую структуру позволяет экспертам эффективнее анализировать выбранную коллекцию, понимать ее основные темы и упрощает поиск нужной информации в ней.

В некоторых прикладных задачах уже имеется накопленная экспертами информация об иерархической структуре коллекции и о принадлежности части документов конкретной теме в данной структуре. В работе предлагается развить существующие методы кластеризации новых документов методами частичного обучения. В частности, для классификации неразмеченных документов предлагается ввести оператор релевантности, ставящий в соответствие каждому документу перестановку тем нижнего уровня иерархии в порядке убывания сходства документа с темой. Для вычисления сходства документа и темы предлагается ввести взвешенную функцию сходства, учитывающую иерархическую структуру коллекции, и методы настройки ее весов по историческим данным.

Цель работы. Целью данной работы является решение проблемы построения иерархических тематических моделей коллекции коротких документов.

Методы исследования. Для достижения поставленной цели используются методы иерархического тематического моделирования [1–5] и дивизимные алгоритмы жесткой неметрической кластеризации [3,6]. Для поиска оптимальных значений ввo-

димых гиперпараметров используются элементы теории оптимизации [7].

Основные положения, выносимые на защиту.

1. Иерархическая взвешенная мера сходства документа и кластера.
2. Алгоритм построения иерархической тематической модели частично размеченной коллекции коротких текстов.

Научная новизна. Разработан новый подход построения иерархических тематических моделей для частично размеченных коллекций, состоящих из коротких текстов. Предложена взвешенная мера сходства документа и кластера, учитывающая иерархичность структуры тематической модели, метод оптимизации параметров данной меры с помощью энтропийного подхода. Введен оператор релевантности, ранжирующий кластеры тематической модели по убыванию сходства для классификации нового документа.

Практическая значимость. Предложенные в работе методы позволяют строить иерархические модели крупных коллекций коротких текстов, учитывая существующие экспертные модели.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной проверкой полученных методов на реальных задачах построения тематических моделей конференции; публикациями результатов исследования в рецензируемых научных изданиях, в том числе рекомендованных ВАК. Результаты работы докладывались и обсуждались на следующих научных конференциях:

1. Международная конференция “26th European Conference on Operational Research”, 2013 г.
2. Международная конференция “20th Conference of the International Federation of Operational Research Societies”, 2014 г.

Работа поддержана грантом Российского фонда фундаментальных исследований и Министерства образования и науки РФ:

1. 14-07-31264, Российский фонд фундаментальных исследований в рамках гранта “Развитие методов визуализации иерархических тематических моделей”,

2. 07.524.11.4002, Министерство образования и науки РФ в рамках Государственного контракта “Система агрегирования и публикации научных документов ВебСервис: построение тематических моделей коллекции документов”.

Публикации по теме дипломной работы. Основные результаты по теме диплома изложены в трех изданиях из списка ВАК:

1. А. А. Кузьмин, А. А. Адуенко, В. В. Стрижов Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ.. — 2012. — № 3. — С. 119-131. — ISSN 2071-6141.

2. А. А. Кузьмин, В. В. Стрижов Проверка адекватности тематических моделей коллекции документов. // Программная инженерия. — 2013. — № 4. — С. 16-20. — ISSN 2220-3397.

3. А. А. Кузьмин, А. А. Адуенко, В. В. Стрижов Тематическая классификация тезисов крупной конференции с использованием экспертной модели // Информационные технологии. — 2014. — № 6. — С. 22-26.

Обзор литературы.

Тематическое моделирование активно развивается [1, 2, 8–13] последние 15 лет вследствие появления коллекций документов большого объема. В области построения иерархических тематических моделей было предложено множество методов [3–6].

В общем случае при построении тематических моделей задано:

1. коллекция документов – текстов произвольной длинны,

2. информация о структуре требуемой тематической модели – общий тип структуры, например, направленный ациклический граф DAG [14, 15] или древовидная структура [3]), параметры структуры, например, число узлов на каждом уровне дерева,

3. документы коллекции, для которых априори известно экспертное положение в структуре тематической модели,

4. способ отнесения документа к элементу структуры тематической модели, например, при жесткой модели, каждый документ относится только к одной теме, а при вероятностном подходе каждый документ принадлежит одной или нескольким темам с некоторой вероятностью.

Требуется определить положение каждого документа коллекции в структуре тематической модели.

Ранее был предложен ряд методов [3–6] для решения подобных задач. Эти методы используют различные гипотезы и предположения о структуре искомой модели. В большинстве из них можно выделить следующие этапы построения модели: 1) предобработка документов коллекции, 2) построение словаря коллекции, 3) представление документов в виде числовых векторов и 4) применение алгоритма построения тематической модели к полученному набору векторов.

На этапе предобработки документов проводится удаление стоп-слов и нормализация оставшихся слов в документах. Для этого применяются алгоритмы стемминга и лемматизации [16], например, метод удаления аффиксов [17], метод разнообразия продолжений [18], N-граммный метод [19].

Из нормализованных слов документов составляется терминологический словарь. Предполагается, что тема документа не зависит от порядка слов в документе. Тем самым делается предположение, что документ является “мешком слов” [20].

Каждый документ представляется в виде числового вектора [21]. Для отсева неинформативных слов могут быть использованы либо взвешенные метрики [22], либо векторные представления документов с учетом частоты встречаемости данного термина не только в документе, но и в коллекции [23].

Построенная таким образом матрица плана используется для построения тематической модели коллекции. Методы построения тематических моделей можно разделить на четыре раздела по тому, каким способом они описывают документ и тему документа в коллекции (см. Таб. 1).

Таблица 1: Основные типы алгоритмов текстовой кластеризации

Тип моделей	Документ	Тема	Пример алгоритма
Жесткие	вектор	вектор	k-means [24], SVM [3], Нейронные сети [6]
Описательно-вероятностные	вектор	вероятность	DPM [12]
Смеси	вектор	распределение	mixture of Gaussian, vMF [13]
Вероятностные	распределение	распределение	LDA [8], PAM [14], HDP [10]

Алгоритмы построения жестких моделей сводятся к кластеризации произвольных объектов в метрическом [24] или неметрическом [25] пространстве. Документы представляются в виде векторов, а темы – в виде векторов-центров кластеров.

Важным этапом построения одномерной кластеризации является выбор функции расстояния (сходства) векторов документов, при помощи которой можно было бы попарно сравнивать документы и объединять их в кластеры.

Для этого используются взвешенная метрика Минковского [22], взвешенная косинусная мера сходства [26], либо ее частный случай – косинусная мера сходства [27, 28]. Для построения иерархических моделей используются дивизимные алгоритмы [3, 6].

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 одновременно к нескольким темам. Одним из первых алгоритмов построения вероятностных тематических моделей был PLSA [8]. Документы и темы в данной модели представляются распределениями. Для уменьшения числа настраиваемых параметров и вероятно-

сти переобучения используются различные варианты регуляризации, как, например, LDA [9] или ARTM [29].

Для построения иерархических вероятностных моделей был предложен алгоритм hLDA [1] – обобщение алгоритма LDA. Он строит древовидную структуру тем, поэтому темы из разных ветвей не могут одновременно присутствовать в одном документе. Это может приводить к дублированию тем в разных ветвях. Чтобы этого избежать, были предложены алгоритмы PAM [30] и HPAM [2], в которых вместо древовидной структуры тем используется направленный ацикличный граф (DAG).

Если число тем на каждом уровне иерархии не фиксировано, то используются непараметрические методы [5,10], в частности, основанные на процессе Дирихле [31].

Вероятностные методы дают возможность строить гибкие модели, в которых каждый документ может содержать все возможные темы в различных пропорциях, и каждый термин являться в какой-то мере термином для каждой темы. Однако ценой данной свободы и мягких, вероятностных, предположений является большое число параметров [8], а Баесовский вывод сильно усложняет [5,12] адаптацию данных методов под конкретную специфику задачи.

Компромиссом между жесткими и вероятностными подходами являются методы, наследующие от вероятностных необходимые свойства (например, возможность документа принадлежать нескольким темам), простоту описания и вывода жестких моделей. Так, алгоритм [13] рассматривает документ как числовой вектор, в то время тему – как смесь распределений vMF [32,33], а алгоритм DPM [12], используя подход жестких моделей, позволяет вычислять вероятности принадлежать выбранной теме для каждого документа.

1 Постановка задачи

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 доку-

менте d_s k раз, то $x_{s,j} = k$, $k \geq 0$. Получим матрицу $\mathbf{X}$ плана, где каждая строка $\mathbf{x}_s = [x_{s,1}, \ldots, x_{s,n}]$ – признаковое описание документа d_s:

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

Представим иерархическую тематическую модель в виде дерева $T = (V, E)$, где V – множество вершин дерева, а E – множество ребер дерева (см. рис. 1). При этом вершины дерева в силу его иерархической структуры можно разбить на уровни. Таким образом, ребра дерева соединяют только вершины соседних уровней. 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 $h+1$. Кластером c будем называть подмножество коллекции документов D. Сопоставим каждому узлу i уровня l дерева кластер $c_{l,i}$, состоящий из документов d_s, путь до которых от вершины $c_{1,1}$ проходит через узел (l, i). Введем оператор предшествования B, ставящий в соответствие кластеру $c_{l,\,i}$ его предшественника в пути от корня дерева к $c_{l,\,i}$ (при этом $B(c_{1,\,1}) = c_{1,\,1}$ по определению):

$$B(c_{l,\,i}) = c_{l-1,\,j}, \text{ где ребро } (c_{l-1,\,j},\, c_{l,\,}) \in E. \tag{2}$$

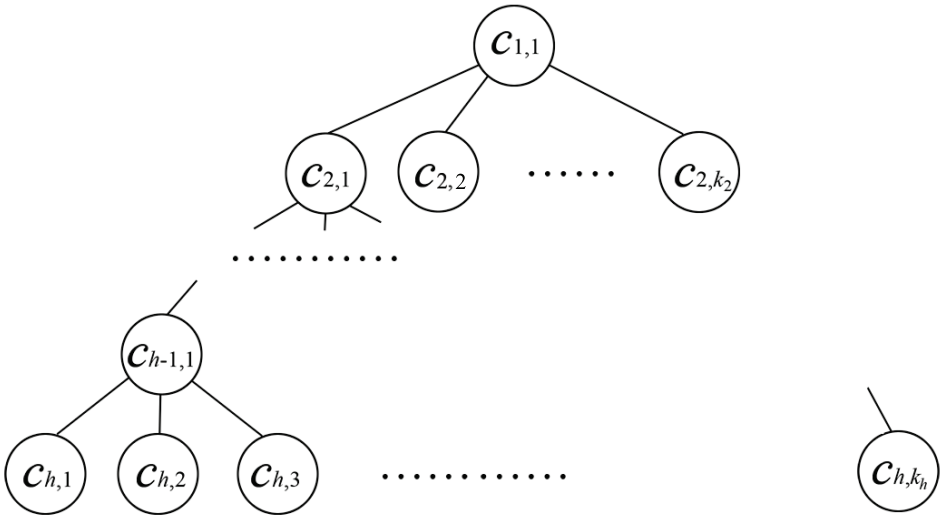


Рис. 1: Иерархическое представление тематической модели.

2 Функция сходства для иерархической структуры

Для использования жестких методов кластеризации, зададим способ сравнения документов.

2.1 Функция сходства документов

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

$$s(\mathbf{x}_i,\ \mathbf{x}_j) = \frac{\mathbf{x}_i^\mathsf{T} \mathbf{\Lambda} \mathbf{x}_j}{\sqrt{\mathbf{x}_i^\mathsf{T} \mathbf{\Lambda} \mathbf{x}_i} \sqrt{\mathbf{x}_j^\mathsf{T} \mathbf{\Lambda} \mathbf{x}_j}}. \tag{3}$$

В случае равенства знаменателя нулю значение функции сходства считается равной нулю. Симметричная неотрицательно определенная матрица $\mathbf{\Lambda} = \mathbf{\Lambda}^\mathsf{T}$ введена для учета важности признаков и зависимости между ними. Эту матрицу мы будем считаем диагональной, то есть $\mathbf{\Lambda} = diag(\lambda_1,\ \ldots,\ \lambda_n),\ \lambda_i \geq 0$, так как каждый объект имеет лишь небольшую долю ненулевых признаков и оптимизация всех элементов матрицы $\mathbf{\Lambda}$ размера $n \times n$ представляется нежелательной. Для удобства дальнейшего изложения нормируем все ненулевые строки матрицы документ – признак $\mathbf{X}$ следующим образом:

$$\mathbf{x}_s \mapsto \frac{\mathbf{x}_s}{\sqrt{\mathbf{x}_s^\mathsf{T} \mathbf{\Lambda} \mathbf{x}_s}}. \tag{4}$$

С учетом этой нормировки функция сходства (3) приобретает вид

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

Косинусная функция сходства (3) позволяет документам $\mathbf{x}_i$ и $\mathbf{x}_j$, имеющим разную длину в словах, но одинаковый словарный состав, быть близкими. 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.

Расширим понятие функции сходства для кластеров и иерархической структуры кластеров.

2.2 Функция сходства кластеров

Под функцией 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 (5). Среднее сходство $S(\cdot, \cdot)$ внутри одного кластера для каждого документа d_s определяется как среднее сходство $s(\cdot, \cdot)$ с документами данного кластера. При этом считаем, что учитывается и сходство документа с самим собой.

$$S(c_{l,i},\ c_{l,j}) = \frac{1}{|A|} \sum_{(\mathbf{x},\ \mathbf{y}) \in A} s(\mathbf{x},\ \mathbf{y}), \tag{5}$$

где A – множество всех пар документов из кластеров $c_{l,i}$ и $c_{l,j}$ таких, что $\mathbf{x} \in c_{l,i}$, $\mathbf{y} \in c_{l,j}$. Определим $\bar{\mathbf{x}}_i$ как средний вектор в кластере $c_{l,i}$

$$\bar{\mathbf{x}}_i = \frac{1}{|c_{l,i}|} \sum_{\mathbf{x} \in c_{l,i}} \mathbf{x}. \tag{6}$$

Тогда в соответствии с (5) и (3) получим

$$S(c_{l,i},\ c_{l,j}) = \frac{1}{|c_{l,i}||c_{l,j}|} \sum_{\mathbf{x} \in c_{l,i}} \sum_{\mathbf{y} \in c_{l,j}} \mathbf{x}^\intercal \mathbf{\Lambda} \mathbf{y} =$$

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

Аналогично, для внутрикластерного сходства:

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

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

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

В последнем выражении учтена нормировка $\mathbf{x}^\intercal \mathbf{\Lambda} \mathbf{x} = 1$. Таким образом, сходство между парой кластеров определяется только средними векторами кластеров, что позволяет его эффективно считать и пересчитывать при изменении состава кластеров.

2.3 Сходство документа с кластером

Пусть у нас есть кластерная структура документов, созданная экспертами, а нашей задачей является классифицировать новый документ $\mathbf{x}$ с неизвестной меткой класса. Будем рассматривать $\mathbf{x}$ как одноэлементный кластер. Определим сходство документа $\mathbf{x}$ с кластером c_i как

$$s(\mathbf{x},\ c_i) = \mathbf{x}^{\mathsf{T}} \mathbf{\Lambda} \overline{\mathbf{x}}_i \tag{7}$$

с учетом введенной нормировки (4) и введенной функции сходства кластера к кластеру (5).

В таком случае, поиск нужного кластера для нового документа $\mathbf{x}$ сводится к задаче:

$$\hat{c} = \arg\max_{c} s(\mathbf{x}, c). \tag{8}$$

Сходство документа с кластером иерархической структуры. В случае с древовидной структурой иерархии (частный случай DAG [30]) для определения принадлежности документа $\mathbf{x}$ к кластерам $c_{l,i}$ на каждом уровне l иерархии достаточно решить задачу (8) только для кластеров нижнего уровня $c_{h,i}$, так как это определит кластеризацию данного документа на всех остальных уровнях $c_{l,i}, l \in \{1 \dots h-1\}$. Однако, данный подход является неустойчивым, потому что кластеры нижнего уровня могут содержать небольшое количество документов (в случае с иерархической структурой тезисов конференции, кластеры нижнего уровня содержат по 4 документа). Добавление или удаление одного документа из данного кластера приведет к значительному изменению его среднего вектора $\overline{\mathbf{x}}$, и, как следствие, сходство нового документа $\mathbf{x}$ с данным кластером.

Другим способом решения данной задачи является аналог дивизимного подхода. На первом шаге определим кластер c_{2,i_2} документа $\mathbf{x}$ на втором уровне иерархии (первым уровнем считаем $c_{1,1}$ – кластер, содержащий всю коллекцию), решив задачу (8) для кластеров второго уровня. Для определения кластера третьего уровня c_{3,i_3}, решим задачу (8) для кластеров третьего уровня, лежащих внутри кластера c_{2,i_2}. Будем продолжать данный процесс, пока не дойдем до уровня h. Данный подход является более стабильным, так как при изменении состава кластера последнего

уровня c_{h,i_h}, будет выполняться условие:

$$\| \overline{\mathbf{x}}_{h,i_h} - \overline{\mathbf{x}}'_{h,i_h} \| \geq \| \overline{\mathbf{x}}_{h-1,i_{h-1}} - \overline{\mathbf{x}}'_{h-1,i_{h-1}} \| \geq \ \ldots \ \geq \| \overline{\mathbf{x}}_{1,i} - \overline{\mathbf{x}}'_{1,i} \|,$$

поэтому изменения в кластеризацию документа $\mathbf{x}$ будут, скорее всего, только на нижних уровнях иерархии. Однако это одновременно делает невозможным для документа $\mathbf{x}$ попасть в другой (даже очень схожий с ним) кластер c_{h,i'_h}, если он лежит в другом кластере более высокого уровня, $c_{h,i'_h} \in c_{l,j} \neq c_{l,i_l}, \ \ \mathbf{x} \in c_{l,i_l}$.

Обозначим h–индекс максимального уровня иерархии, рассматриваемого в модели. Введем весовой вектор $\boldsymbol{\theta} \in \mathbb{R}^h_+$ для каждого из уровней иерархии, начиная с верхнего. При этом полагаем, что $\theta_1 = 0$, так как этот уровень объединяет все документы коллекции. Определим эффективное сходство документа $\mathbf{x}$ с кластером $c_{h,\,i}$ нижнего уровня h как

$$s(\mathbf{x},\ c_{h,\,i}) = \sum_{j=0}^{h-1} \theta_{h-j} s(\mathbf{x},\ B^j(c_{L,\,i})), \tag{9}$$

где $B()$ – оператор, возвращающий родительский кластер кластера c. При этом $B^0(c)$ возвращает c, а $B^j(c_{h,i})$ возвращает кластер уровня $h-j$, содержащий кластер $c_{h,i}$.

Задав таким образом функцию сходства, мы можем решать задачу:

$$\hat{c} = \underset{c_{h,\,i}, i \in \{1 \ldots k_h\}}{\arg\max} \ s(\mathbf{x}, c_{h,i}). \tag{10}$$

При этом мы будем учитывать факт, что документ, схожий с кластером c_{h,i_h} нижнего уровня h должен быть схожим со всеми его родительскими кластерами $c_{h-1,i_{h-1}}, \ldots c_{2,i_2}$.

3 Адаптивная модель

Рассмотрим способы настройки диагональных элементов матрицы Λ предложенной функции сходства (10) по коллекции с экспертной иерархической тематической моделью.

3.1 Оператор релевантности

Будем обозначать S^k множество перестановок порядка k. Определим оператор релевантности

$$R:\ \mathbb{R}^n \to S^{k_h}, \tag{11}$$

ставящий в соответствие документу $\mathbf{x} \in \mathbb{R}^n$, перестановку кластеров нижнего из рассматриваемых уровней h. При этом кластеры отсортированы по убыванию сходства документа $\mathbf{x}$ с кластерами в соответствии с (10).

Будем называть кластер $c_{h,i}$ наиболее релевантным для документа $\mathbf{x}_i$ относительно оператора релевантности R, если номер i данного кластера стоит на первом месте в перестановке, возвращаемой R.

Оценка качества оператора релевантности. Пусть имеется выборка $D = (\mathbf{X},\ \mathbf{Z})$, где $\mathbf{X}$ – матрица размера $|D| \times n$, содержащая признаковое описание документов, а $\mathbf{Z}$ – матрица размера $|D| \times h$, содержащая информацию об экспертной кластеризации каждого из документов на каждом уровне иерархии. Элемент $z_{j,\,l}$ – номер кластера l-го уровня иерархии, к которому отнесен документ $\mathbf{x}_j$. Введем функцию

$$\mathrm{pos}(s,\ j) : S^q \times \{1,\ 2,\ \ldots,\ q\} \mapsto \{1,\ 2,\ \ldots,\ q\}.$$

Данная функция ставит в соответствие перестановке $s \in S^q$ и некоторому числу $j \in \{1,\ 2,\ \ldots,\ q\}$ номер позиции в перестановке, на котором стоит число j.

Определим качество оператора релевантности R (11) как среднюю позицию экспертного кластера $z_{j,\,h}$ нижнего уровня h для документа $\mathbf{x}_j$ в перестановке $R(\mathbf{x}_j)$

$$Q(R) = \frac{1}{|D|} \sum_{j=1}^{|D|} \mathrm{pos}(R(\mathbf{x}_j),\ z_{j,\,h}). \tag{12}$$

Чем меньше значение $Q(R)$, тем больше сходство документов с их экспертными кластерами, в сравнении с не экспертными и тем меньше номер их позиции в перестановке, которую возвращает предложенный оператор релевантности R.

Построим кумулятивную гистограмму следующим образом: пусть столбец с номером i принимает значение

$$\#\{\mathrm{pos}(R(\mathbf{x}_j),\ z_{j,\,h}) \le i\},$$

где $\{\text{pos}(R(\mathbf{x}_j),\, z_{j,h})\}$ – множество всех документов, для которых номер позиции экспертного кластер меньше либо равен i в перестановке, возвращаемой оператором R, а $\#\{\cdot\}$ есть число элементов во множестве $\{\cdot\}$.

Сходным критерием качества с $Q(R)$ будет служить $AUC(R)$ – площадь под верхней огибающей кумулятивной гистограммы (3.1):

$$\text{AUC}(R) = \frac{1}{k_h|D|}\sum_{i=1}^{k_h}\#\{\text{pos}(R(\mathbf{x}_j),\, z_{j,h}) \le i\}. \tag{13}$$

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

3.2 Метод оптимизации матрицы весов признаков

Введение матрицы $\mathbf{\Lambda}$ в функции сходства позволяет учесть важность слов. Слова, отделяющие одни кластеры от других в тематической модели, являются наиболее важными. Предположим, что все документы из кластера $c_{l,i}$ содержат термин w, а документы из всех остальных кластеров $c_{l,j}, j \neq i$ не содержат термин w. Предположим, нам нужно классифицировать новый документ, содержащий термин w. Справедливо предположить, что данный документ стоит отнести к кластеру $c_{l,i}$. Данный пример показывает, как термин w может хорошо отделять одни кластеры от других.

Формализуем данную идею, сопоставив каждому слову w_j его энтропию относительно уровня иерархии l. Энтропийный подход оценки важности признаков предлагался в [6]. Расширим данный подход для иерархического случая. Будем считать, что у нас имеются две выборки $D^1 = \{\mathbf{X}^1,\, \mathbf{Z}^1\}$ и $D^2 = \{\mathbf{X}^2,\, \mathbf{Z}^2\}$. Пусть мы знаем экспертную кластеризацию для коллекции D^1 и не знаем ее для D^2.

Определим, пользуясь (6), средние векторы всех кластеров уровня l: $\overline{\mathbf{x}}_{l,1}, \ldots, \overline{\mathbf{x}}^{l,k_l}$ по выборке D^1. Рассмотрим их компоненту, соответствующую слову w_j:

$$\mathbf{p}_l^j = [\overline{\mathbf{x}}_{l,1}^j,\, \ldots,\, \mathbf{x}_{l,k_l}^l]^\intercal \quad \text{и нормируем полученный вектор:} \quad \mathbf{p}_l^j \mapsto \frac{\mathbf{p}_l^j}{\sum_{i=1}^{k_l} p_l^{ji}},$$

где p_l^{ji} – компонента i вектора $\mathbf{p}_l^j$.

Энтропией слова w_j относительно уровня l назовем

$$I_l(w_j) = \sum_{i=1}^{k_l} -p_l^{ji}\log(p_l^{ji}). \tag{14}$$

Минимальное значение энтропии $I_l(w_j) = 0$ соответствует случаю, когда слово w_j встречается в документах только одного кластера уровня l, выделяя тем самым его из остальных. Случай, когда $p_l^{ji} = \text{const}$ для всех $i = 1 \dots kl$ соответствует максимальному значению энтропии и случаю, когда слово w_j никак не выделяет ни один из кластеров на фоне остальных.

Для того, чтобы избежать переобучения, воспользуемся следующей моделью для определения $\lambda_1, \lambda_2, \dots, \lambda_n$, которая уменьшит число переменных в матрице Λ:

$$\lambda_j = 1 + \sum_{l=1}^{L} \alpha_l \log(1 + I_l(w_j)). \tag{15}$$

При этом с учетом высокой корреляции энтропии на разных уровнях (для данных конференций EURO 2012-2013 годов корреляция энтропии слов для уровней областей и направлений превышает 0.9) возможно использование упрощения модели (15), в которой оставлен только один признак – энтропия на некотором фиксированном уровне иерархии l, то есть

$$\lambda_j = 1 + \alpha_l \log(1 + I_l(w_j)). \tag{16}$$

Чтобы оценить параметр(ы) модели (15) или (16), определим средние векторы кластеров каждого уровня по выборке D_1, то есть для кластера $c_{l,i}$ уровня l имеем

$$\overline{\mathbf{x}}_{l,i} = \frac{1}{|D^1_{l,i}|} \sum_{k \in D^1_{l,i}} \mathbf{x}_k, \text{ где}$$

$$D^1_{l,i} = \{k : z_{k,l} = c_{l,i}\}.$$

При этом считаем, что $\mathbf{x}_k^\intercal \mathbf{\Lambda} \mathbf{x}_k = 1$, поэтому при изменении $\mathbf{\Lambda}$ требуется производить перенормировку векторов $\mathbf{x}_k$ из выборки D_1 и пересчет $\mathbf{x}_{l,i}$. Располагая выборкой $D_2 = \{\mathbf{X}^2, \mathbf{Z}^2\}$, ставим задачу оптимизации $Q(R)$ (12) по $\alpha_1, \dots, \alpha_L$ для (15) или по α_l для (16):

$$[\alpha_1^*, \dots, \alpha_L^*(\alpha_l^*)] = \underset{\alpha_1, \dots, \alpha_L(\alpha_l)}{\arg\min} Q(R). \tag{17}$$

Заметим, что при расширении выборки D^2 адаптация модели может производиться двумя способами:

1. Добавление части объектов из выборки D^2 в выборку D^1 и пересчет средних векторов и $\mathbf{\Lambda}$;

2. Расширение выборки D^2 и пересчет $\mathbf{\Lambda}$ согласно (15) или (16) с помощью (17).

4 Построение тематической модели конференции

Для проверки работы предложенных алгоритмов автоматизируем с их помощью процесс построения тематической модели крупной конференции EURO.

Построение тематической модели конференции экспертами. Перед программным комитетом конференции с большим числом участников ежегодно встает задача построения 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 прикрепляется группа экспертов в соответствующей области. Для подачи заявки авторы отправляют аннотацию (далее документ) к своей работе, состоящую из не более чем 600 символов. 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 их содержании. Затем эксперты распределяют отобранные ими документы по иерархической структуре их главной темы, исходя из своей стратегии организации докладов конференции.

Структура конференции (набор областей и направлений) практически не изменяется из года в год. Предлагается автоматизировать процесс построения тематической модели, имея информацию об экспертной кластеризации тезисов за предыдущие года, используя предложенную иерархическую функцию сходства.

В качестве данных использовались выборки D^1 и D^2, содержащие тезисы конференции EURO за 2010, 2012 и 2013 годы, названия тезисов были включены в их текст. Выборка D^1 содержала 1342 тезиса из 2012 года и 2313 тезисов из 2013 года. Объем словаря $n = 1675$ слов.

В качестве выборки D^2 использовалась выборка тезисов за 2010 год размером в 1663 документа. Всего после сопоставления областей и направлений разных лет было выявлено $k_2 = 24$ области и 178 направлений, 15 из которых были представлены только в 2010 году. В связи с этим активно в выборке D_2 использовались только 1480

документов из направлений, которые были в 2012 и/или 2013 году. Таким образом, использовались только $k_3 = 163$ направлений.

4.1 Оператор релевантности на основании тривиального метода

Заметим, что можно предложить тривиальный оператор релевантности $R_0(\mathbf{x})$, который генерирует случайную перестановку из группы перестановок S^{k_3}. Для такого метода

$$\mathbb{E}Q(R_0) = k_3/2 = 81.5,$$

то есть средний ранг равен 81.5. Кроме большой величины среднего ранга этот метод ранжирования по релевантности не детерминирован, то есть результат не точно определен входным вектором $\mathbf{x}$. Сравнение предложенного метода ранжирования по релевантности будем проводить с другим тривиальным методом.

Соответствующий оператор релевантности $R_1(\mathbf{x})$ определим следующим образом. Упорядочим направления по убыванию их размера в выборке D^1. Пусть $c_{3,i_1}, \ldots, c_{3,i_{k_3}}$ – соответствующий порядок, то есть

$$|c_{3,i_1}| \geq |c_{3,i_2}| \geq \ldots \geq |c_{3,i_{k_3}}|.$$

При этом направления одинакового размера располагаем в произвольном, но фиксированном порядке. Независимо от документа $\mathbf{x}$ будем выдавать одну и ту же перестановку номеров направлений $(i_1, i_2, \ldots, i_{k_3})$, то есть $R_1(\mathbf{x}) = (i_1, i_2, \ldots, i_{k_3})$. С таким оператором релевантности будем сравнивать оператор, предлагаемый в работе, основанный на введенной функции сходства.

4.2 Сравнение тривиального оператора релевантности с предложенным

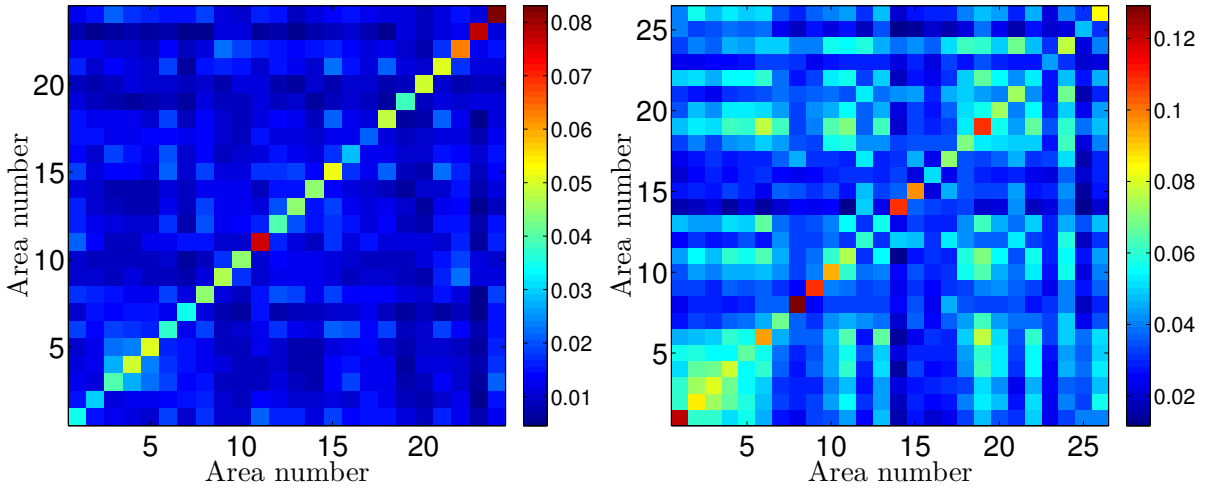
Построение предложенного оператора релевантности. Сосчитав энтропию каждого из 1675 слов словаря для областей и направлений, заметим, что корреляция Пирсона энтропии для областей и направлений равна 0.9478, а корреляция Кендалла равна 0.7999. Поэтому воспользуемся моделью (16) для $\mathbf{\Lambda}$ при $l = 2$. Положим

в (9) $\theta_3 = 1$, $\theta_2 = 1.5$. Выполняя оптимизацию $Q(R)$ по параметру α_2 в соответствии с (17), получим $\alpha_2^* = -0.7063$. Таким образом, так как $\alpha_2^* < 0$, менее информативные слова (то есть обладающие большей энтропией) получили меньший вес, чем более информативные. Примерами слов, вес которых снизился до менее чем 0.1 являются слова generate, integrate, limit, depend, exist. Оптимизация $Q(R)$ при фиксированном α_2 дает значение $\theta_2^* = 1.8$. Однако значение среднего ранга имеет небольшое изменение в области $\theta_2 \in [1, 2]$ (см. рис. 3б), а потому полную оптимизацию по θ_2 можно не производить.

Сравним результаты применения операторов релевантности $R_1(\mathbf{x})$ и предложенного $R(\mathbf{x})$ для выборки D^2.

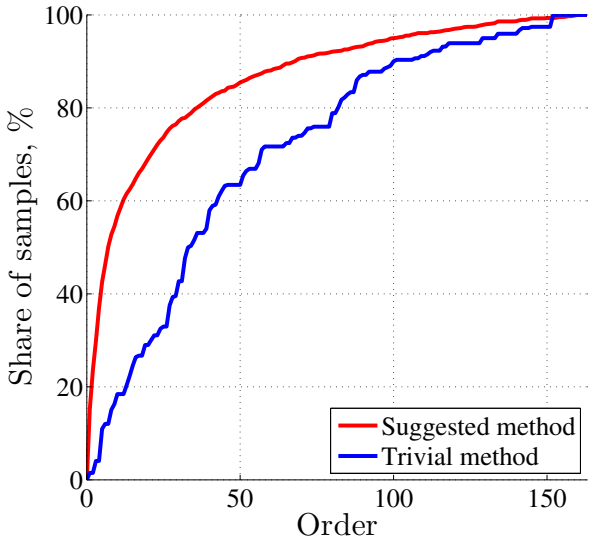
Таблица 2: Значения функционалов качества для сравниваемых операторов релевантности Q (12) и AUC (13)

Функционал качества \ Оператор релевантности	$R_1(\cdot)$	$R(\cdot)$
$Q(\cdot)$	46.86	22.54
AUC$(\cdot)$	0.719	0.868

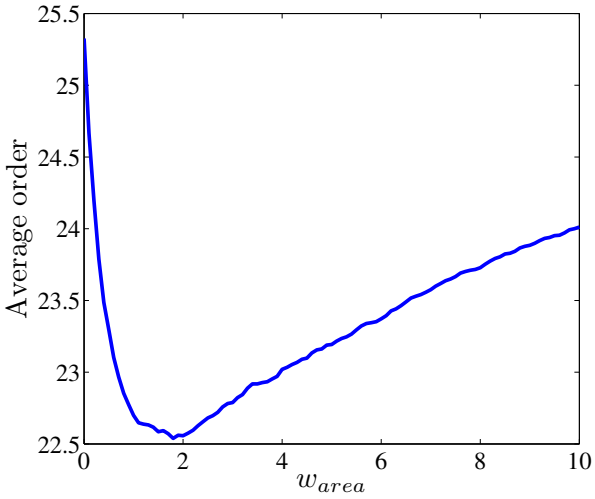


a: $\mathbf{\Lambda} = \mathbf{\Lambda}^*$ b: Введенная функция сходства для $\mathbf{\Lambda} = \mathbf{I}$

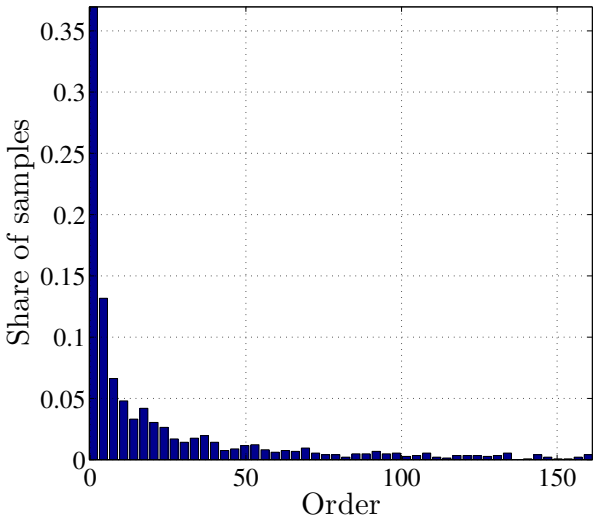
Рис. 2: Средние сходства кластеров.



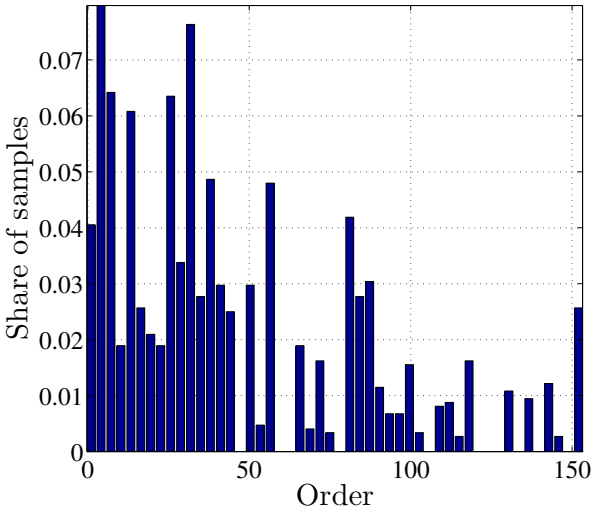
a: Сравнение операторов релевантности $R(\cdot)$ и $R_1(\cdot)$ по AUC



b: Зависимость $Q(R)$ от θ_2.



c: Гистограмма распределения объектов по релевантности их экспертного направления по $R(\cdot)$.



d: Гистограмма распределения объектов по релевантности их экспертного направления по $R_1(\cdot)$.

Рис. 3: Иллюстрация свойств операторов релевантности $R_1(\cdot)$ и $R(\cdot)$.

Из табл. 2 видно, что предложенный оператор релевантности $R(\cdot)$ значительно превосходит оператор релевантности $R_1(\cdot)$. Приведем далее кривые – верхние огибающие кумулятивных гистограмм (3.1) $\#\{\text{pos}(R(\mathbf{x}_j),\, z_{j,3}) \leq i\}$ для $i \in [1,\, k_3]$ (см. рис. 3а и выражение (13)). На рис. 3в, г приведем гистограммы распределения $\#\{\text{pos}(R(\mathbf{x}_j),\, z_{j,3}) = i\}$ для $i \in [1,\, k_3]$, показывающую долю объектов, для которых

их экспертное направление имеет i-ое место по релевантности по оператору релевантности $R(\cdot)$ и $R_1(\cdot)$.

Сравним средние сходства документов из разных областей и для выборки D^1 при найденном значении $\mathbf{\Lambda} = \mathbf{\Lambda}^*$ (см. рис. 2а) со средними сходствами до оптимизации (см. рис. 2б). Можно заметить, что после оптимизации значений $\mathbf{\Lambda}$, внутрикластерные сходства (элементы на диагонали) стали заметно больше не диагональных элементов (межкластерных сходств). Так, среднее значение сходства документов из одной области равно 0.0468, в то время как из разных областях 0.0121. Таким образом, введенная функция сходства с найденной $\mathbf{\Lambda}$ позволяет выделить сходство документов внутри одного/одной направления/области и указать на малое сходство документов из разных кластеров.

4.3 Реализация предложенных методов

Предложенный метод построения тематической модели крупной конференции был реализован в ввиде интернет ресурса, доступный по сайту http://europrogramadvisor.com. Рис. (4).

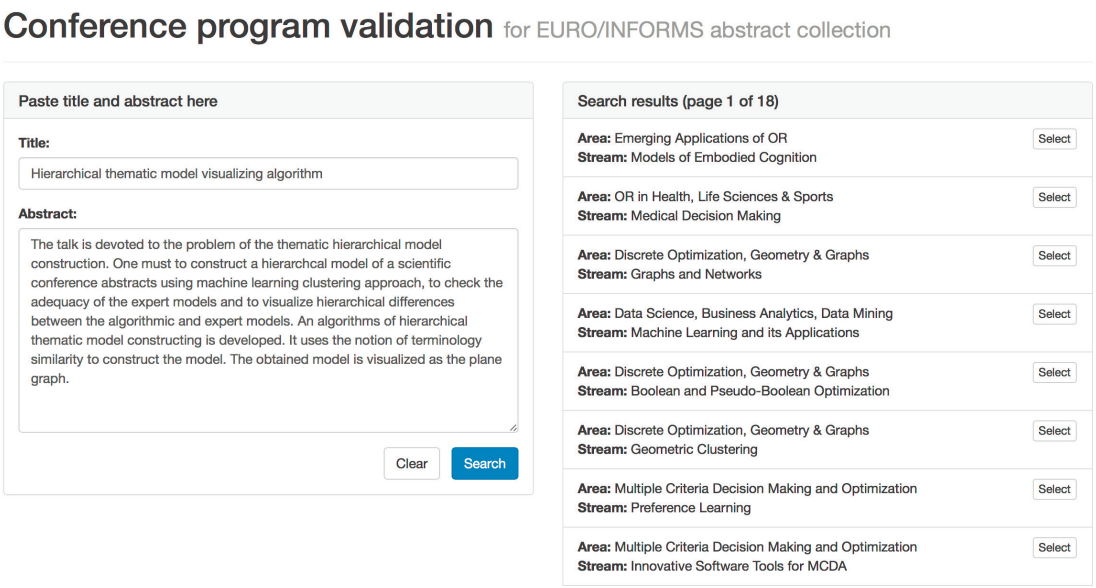


Рис. 4: Реализация предложенных методов.

Для использования сервиса необходимо в поле “Abstract” вставить текст анно-

тации к докладу и нажать кнопку “Search”. Поле “Search results” содержит список подходящих областей и направлений для доклада, отсортированные в порядке убывания сходства.

Заключение

В работе предлагается метод построения иерархической модели по частично размеченной коллекции коротких документов. Предполагается, что иерархическая структура коллекции содержит фиксированное число уровней и кластеров на каждом уровне. Данные параметры структуры определяются экспертами. Предлагается иерархическая взвешенная функция сходства документа и кластера, и оператор релевантности документа, который ранжирует с помощью функции сходства кластеры нижнего уровня иерархии по убыванию сходства с выбранным документом. Предлагается два критерия качества оператора релевантности – средняя позиция экспертного кластера в отранжированном списке кластеров и площадь под кумулятивной гистограммой. Предлагается метод оценки параметров матрицы весов функции сходства, основанный на энтропийном подходе, определяющий важность терминов в коллекции.

Работа предложенных алгоритмов демонстрируется построением иерархической тематической модели крупной конференции European Conference on Operational Research – 2010(EURO – 2010). Коллекции EURO – 2012 и EURO – 2013 использовались для определения структуры конференции и настройки параметров алгоритма. Реализация предложенных алгоритмов доступна на http://europrogramadvisor.com и может использоваться для классификации новых аннотаций к докладам.

Список литературы

[1] *Blei David M., Griffiths Thomas L., Jordan Michael I.* The Nested Chinese Restaurant Process and Bayesian Nonparametric Inference of Topic Hierarchies // *J. ACM.* — 2010. — Февраль. — Vol. 57, no. 2. — Pp. 7:1–7:30.

[2] *Mimno David, Li Wei, McCallum Andrew.* Mixtures of Hierarchical Topics with Pachinko Allocation // Proceedings of the 24th International Conference on Machine Learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 633–640.

[3] *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.

[4] *Li Wei, Blei David M., McCallum Andrew.* Nonparametric Bayes Pachinko Allocation // *CoRR.* — 2012. — Vol. abs/1206.5270.

[5] *Zavitsanos Elias, Paliouras Georgios, Vouros George A.* Non-Parametric Estimation of Topic Hierarchies from Texts with Hierarchical Dirichlet Processes // *J. Mach. Learn. Res.* — 2011. — Ноябрь. — Vol. 12. — Pp. 2749–2775.

[6] *Ruiz Miguel E., Srinivasan Padmini.* Hierarchical Text Categorization Using Neural Networks // *Information Retrieval.* — 2002. — Vol. 5, no. 1. — Pp. 87–118.

[7] *Bishop C.M.* Pattern Recognition and Machine Learning. Information Science and Statistics. — Springer, 2006.

[8] *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.

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

[10] Hierarchical Dirichlet Processes / Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, David M. Blei // *Journal of the American Statistical Association.* — 2006. — Vol. 101, no. 476. — Pp. 1566–1581.

[11] *Blei D., Lafferty J.* Correlated Topic Models // *Advaces in neural information processing systems.* — 2006. — Vol. 18. — P. 147.

[12] Keep It Simple with Time: A Reexamination of Probabilistic Topic Detection Models / Qi He, Kuiyu Chang, Ee-Peng Lim, A. Banerjee // *Pattern Analysis and Machine Intelligence, IEEE Transactions on.* — 2010. — Oct. — Vol. 32, no. 10. — Pp. 1795–1808.

[13] Generative Model-based Clustering of Directional Data / Arindam Banerjee, Inderjit Dhillon, Joydeep Ghosh, Suvrit Sra // Proceedings of the Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. — KDD '03. — New York, NY, USA: ACM, 2003. — Pp. 19–28.

[14] *Li Wei, McCallum Andrew.* Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations // Proceedings of the 23rd International Conference on Machine Learning. — ICML '06. — New York, NY, USA: ACM, 2006. — Pp. 577–584.

[15] *Mimno David, Li Wei, McCallum Andrew.* Mixtures of Hierarchical Topics with Pachinko Allocation // Proceedings of the 24th International Conference on Machine Learning. — ICML '07. — New York, NY, USA: ACM, 2007. — Pp. 633–640.

[16] *Frakes W. B.* Stemming Algorithms // Information Retrieval / edited by-William B. Frakes, Ricardo Baeza-Yates. — Upper Saddle River, NJ, USA: Prentice-Hall, Inc., 1992. — Pp. 131–160.

[17] *Lovins J. B.* Development of a stemming algorithm // *Mechanical Translation and Computational Linguistics.* — 1968. — Vol. 11. — Pp. 22–31.

[18] *Hafer M., Weiss S.* Word Segmentation by Letter Successor Varieties // *Information Storage and Retrieval.* — 1974. — Vol. 10. — Pp. 371–385.

[19] *Adamson George W., Boreham Jillian.* The use of an association measure based on character structure to identify semantically related pairs of words and document titles // *Information Storage and Retrieval.* — 1974. — Vol. 10, no. 7–8. — Pp. 253–260.

[20] *Salton Gerard, McGill Michael J.* Introduction to Modern Information Retrieval. — New York, NY, USA: McGraw-Hill, Inc., 1986.

[21] *Кузьмин А. А., Адуенко А. А., Стрижов В. В.* Выбор признаков и оптимизация метрики при кластеризации коллекции документов // *Известия ТулГУ.* — 2012. — Т. 3. — С. 119–131.

[22] *Cordeiro de Amorim Renato, Mirkin Boris.* Minkowski Metric, Feature Weighting and Anomalous Cluster Initializing in K-Means Clustering // *Pattern Recogn.* — 2012. — Март. — Vol. 45, no. 3. — Pp. 1061–1075.

[23] *Srivastava Asho, Sahami Mehran.* Text mining : classification, clustering, and applications. — Boca Raton, FL: CRC Press, 2009. — Pp. –.

[24] *Hartigan J. A., Wong M. A.* Algorithm AS 136: A k-means clustering algorithm // *Applied Statistics.* — 1979. — Vol. 28, no. 1. — Pp. 100–108.

[25] *Ackermann Marcel R., Blömer Johannes, Sohler Christian.* Clustering for Metric and Nonmetric Distance Measures // *ACM Trans. Algorithms.* — 2010. — Сентябрь. — Vol. 6, no. 4. — Pp. 59:1–59:26. http://doi.acm.org/10.1145/1824777.1824779.

[26] *Yih Wen-tau.* Learning Term-weighting Functions for Similarity Measures // Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2 - Volume 2. — EMNLP '09. — Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. — Pp. 793–802. http://dl.acm.org/citation.cfm?id=1699571.1699616.

[27] *Schedl Markus.* #Nowplaying Madonna: A Large-scale Evaluation on Estimating Similarities Between Music Artists and Between Movies from Microblogs // *Inf. Retr.* — 2012. — Июнь. — Vol. 15, no. 3-4. — Pp. 183–217.

[28] *Кузьмин А. А., Адуенко А. А., Стрижов В. В.* Тематическая классификация тезисов крупной конференции с использованием экспертной модели // *Информационные технологи.* — 2014. — Т. 6. — С. 22–26.

[29] *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.

[30] *Li Wei, McCallum Andrew*. Pachinko Allocation: DAG-structured Mixture Models of Topic Correlations // Proceedings of the 23rd International Conference on Machine Learning. — ICML '06. — New York, NY, USA: ACM, 2006. — Pp. 577–584.

[31] *Ferguson Thomas S.* A Bayesian Analysis of Some Nonparametric Problems // *The Annals of Statistics*. — 1973. — Vol. 1, no. 2. — Pp. 209–230.

[32] *Mardia K. V., Jupp. P.* Directional Statistics (2nd edition). — John Wiley and Sons Ltd., 2000.

[33] *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.