На правах рукописи

Кузьмин Арсентий Александрович

ИЕРАРХИЧЕСКАЯ КЛАССИФИКАЦИЯ КОЛЛЕКЦИЙ ДОКУМЕНТОВ

05.13.17 — Теоретические основы информатики

АВТОРЕФЕРАТ
диссертации на соискание ученой степени
кандидата физико-математических наук

Москва — 2017

Работа выполнена на Кафедре интеллектуальных систем Федерального государственного автономного образовательного учреждения высшего образования «Московский физико-технический институт (государственный университет)».

Научный руководитель: **Стрижов Вадим Викторович**
доктор физико-математических наук, Федеральный исследовательский центр «Информатика и управление» Российской академии наук, отдел интеллектуальных систем, научный сотрудник.

Официальные оппоненты: **Миркин Борис Григорьевич**
доктор технических наук, доцент, Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет «Высшая школа экономики», Факультет компьютерных наук, Департамент анализа данных и искусственного интеллекта, профессор.

Александров Михаил Аронович
кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждения высшего профессионального образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации», Кафедра системного анализа и информатики, доцент.

Ведущая организация: Факультет вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М. В. Ломоносова»

Защита состоится “___”_____ 2017 года в ___:___ на заседании диссертационного совета Д 002.073.05 при Федеральном исследовательском центре «Информатика и управление» Российской академии наук (ФИЦ ИУ РАН) по адресу: 119333, г. Москва, ул. Вавилова, д. 40.

С диссертацией можно ознакомиться в библиотеке и на сайте ФИЦ ИУ РАН http://www.frccsc.ru/

Автореферат разослан “___”_____ 2017 года.

Ученый секретарь
диссертационного совета Д 002.073.05,
д.ф.-м.н., профессор В.В.Рязанов

Общая характеристика работы

Актуальность темы. В работе исследуются методы категоризации и классификации текстовых документов, автоматически структурирующие документы в виде иерархий тем и оптимизирующие уже существующие, выявляя в них тематические несоответствия (Hofmann: 1999, He: 2010, Blei: 2010).

Тематическая модель – модель коллекции текстовых документов, которая определяет, к каким темам относится каждый документ коллекции. В работе исследуется фундаментальная проблема тематического моделирования – классификация документов из частично размеченных коллекций с экспертно заданной иерархической структурой тем (McCallum: 1998, Лукашевич: 2008, Кузнецов: 2015). Решением задачи классификации является отображение подмножества неразмеченных документов коллекции во множество тем, наилучшим образом восстанавливающее экспертную классификацию согласно заданному критерию качества. В случае большого числа тем вместо единственного релевантного кластера предлагается ранжированный список кластеров согласно их релевантности документу. При несовпадении экспертного мнения и наиболее релевантного кластера эксперт рассматривает следующие по релевантности кластеры в качестве альтернативных вариантов.

Коллекциями документов являются аннотации к научным работам (Joachims: 1998), доклады на конференциях (Кузнецов: 2015), текстовые сообщения в социальных сетях (Лукашевич: 2016), текстовая информация веб-сайтов (Xue: 2008), описания патентов, новостные сводки (Ikonomakis: 2005, Linghui: 2011) и описания фильмов (Schedl: 2012). Предполагается, что экспертное разделение документов на темы является эталонным. В связи со значительным размером коллекций и числом тем распределение документов по темам является для экспертов трудоемкой задачей. Поэтому автоматическая классификация неразмеченных документов и поиск небольшого числа наиболее подходящих тем для каждого неразмеченного документа для дальнейшего принятия решения экспертом являются актуальными задачами.

Для текстовой классификации и кластеризации были предложены жесткие методы, в которых каждому документу ставится в соответствие единственный кластер (Hartigan: 1979, Hao: 2007), описательно вероятностные методы, в которых оценивается вероятность принадлежности документа каждому из кластеров (He: 2010), смеси моделей (Banerjee: 2003) и вероятностные методы (Blei: 2003, Воронцов: 2015) в которых темы являются распределениями над множеством слов, а документы – распределениями над множеством тем. Для коллекций с большим числом тем были предложены иерархические методы, позволяющие учитывать взаимосвязи между темами (Hao: 2007, Zavitsanos: 2011).

Важной проблемой при построении метрических алгоритмов классификации и кластеризации является выбор метрики (Leisch: 2006) как способа сравнения векторных представлений документов. В (Amorim: 2012) для учета соот-

ношения масштабов признаков рассматривается взвешенная метрика Минковского. Веса интерпретируются как важность слов. В данной работе исследуются способы оптимизации весов взвешенной метрики, а также различные способы векторного представления документов, наилучшим образом восстанавливающие экспертную классификацию. Альтернативой взвешенной функции расстояния является взвешенная функция сходства (Yih: 2009). Для уменьшения числа параметров оптимизации предлагается энтропийный метод определения важности слов во взвешенной функции сходства через их энтропию относительно экспертной кластеризации на различных уровнях иерархии. Для иерархической классификации предлагается иерархическая взвешенная функция сходства, позволяющая учитывать сходство сразу со всей веткой дерева экспертной иерархической структуры коллекции.

Для оптимизации параметров иерархической функции сходства рассматривается вероятностная постановка задачи, в которой вероятность принадлежности кластеру оценивается как нормированная экспоненциальная функция softmax от значений иерархического сходства с кластерами. Задача поиска параметров сводится к максимизации правдоподобия модели.

При наличии априорных распределений параметров аналитический байесовский вывод апостериорного распределения параметров иерархической функции сходства и совместного апостериорного распределения параметров и классов неразмеченных документов не является возможным. В работах (Gershman: 2012, Blei: 2016) рассматриваются способы приближенного вариационного вывода и аппроксимации правдоподобия. В работе данные идеи используются для аналитического вывода апостериорного распределения параметров (Bishop: 2006, Стрижов: 2016), а также для аппроксимации совместного апостериорного распределения классов неразмеченных документов и параметров.

Для размеченных коллекций возникает задача верификации. Решением этой задачи является изменение у фиксированного набора документов их тем так, чтобы качество полученной модели стало максимальным. Для этого предлагается алгоритм построения иерархической модели, схожей с существующей, для выявления значимых тематических несоответствий в модели. Предлагаются варианты устранения несоответствий путем переноса некоторых документов в другие кластеры.

Для визуализации тематической модели были предложены различные подходы (Millar: 2009, Ando: 2000). В случае, когда документы представляются в виде действительных векторов, для их визуализации используются методы понижения размерности (Lee: 2007). При этом кластеры из разных ветвей иерархической модели могут пересекаться. В данной работе предлагается метод построения плоской вложенной визуализации иерархической модели, при которой кластеры более низкого уровня остаются внутри кластеров более высокого уровня на плоскости. Предлагаемый подход опирается на методы, минимизирующие изменения относительного расстояния между документами и центрами

кластеров иерархии (Sammon: 1969).

Цели работы.

1. Исследовать метрические свойства описаний текстовых документов.
2. Предложить критерии качества модели иерархической классификации документов.
3. Построить оптимальную модель иерархической классификации.
4. Получить вариационные оценки апостериорных распределений параметров и гиперпараметров модели.
5. Разработать алгоритм построения модели и провести вычислительный эксперимент для сравнения различных подходов к решению задачи иерархической классификации документов.

Основные положения, выносимые на защиту.

1. Предложен метод иерархической классификации коллекций документов на основе оператора релевантности.
2. Разработана и исследована вероятностная модель иерархической классификации.
3. Предложены методы оптимизации параметров и гиперпараметров модели.
4. Предложен способ вычисления иерархической вероятности класса документа и построения ранжированного списка для последующей экспертной оценки.
5. Разработан программный комплекс для экспертного построения программы конференции.

Методы исследования. Для достижения поставленных целей используются методы иерархического тематического моделирования (Воронцов: 2015, Mimno: 2007, Hao: 2007, Blei: 2012, Zavitsanosh: 2011). Для метрической иерархической кластеризации применяются методы плоской кластеризации (Leisch: 2006, Kogan: 2005) совместно с агломеративным и дивизимным подходами (Ruiz: 2002, Hao: 2007). Для построения локально оптимальной взвешенной метрики используются методы отбора признаков (Воронцов: 2010) и методы условной оптимизации (Boyd: 2004, Bishop: 2006). Для сравнения документов при иерархической классификации используется взвешенная функция сходства (Yih: 2009), а для оптимизации ее параметров развивается энтропийный метод, предложенный в (Ruiz: 2002). Для оптимизации параметров иерархической взвешенной функции сходства и энтропийной модели используются методы вариационного вывода (Gershman: 2012, Blei: 2016), методы выбора моделей (Стрижов: 2014) и методы локальных вариаций (Bishop: 2006). Для построения оператора релевантности используются методы иерархической

классификации (Кузнецов: 2015, McCallum: 1998). Для построения плоской вложенной визуализации иерархической тематической модели используются методы понижения размерности (Sammon: 1969). Для учета синонимичности слов используются языковые модели (Mnih: 2009, Mikolov: 2013) и методы оптимизации параметров нейронных сетей. Кроме того, используются элементы теории вероятности и выпуклой оптимизации.

Научная новизна. Разработан новый подход иерархической классификации частично размеченных коллекций текстовых документов с экспертной иерархической структурой. Предложена иерархическая взвешенная функция сходства документа и кластера, учитывающая иерархичность экспертной кластерной структуры. Предложен метод оценки важности слов с помощью энтропийной модели. Предложена вероятностная модель текстовой коллекции и способ аппроксимации совместного апостериорного распределения параметров модели и классов неразмеченных документов. Предложен способ представления иерархической функции сходства в виде многослойной нейронной сети и способ учета синонимичности слов. Введен оператор релевантности, ранжирующий кластеры тематической модели по убыванию релевантности новому документу. Для верификации экспертной тематической модели предложен метод построения модели, схожей с экспертной, и выявления наиболее значимых несоответствий. Предложен метод вложенной визуализации экспертной иерархической тематической модели на плоскости, а также выявленных несоответствий и вариантов повышения тематической целостности модели.

Теоретическая значимость. В данной диссертационной работе предложенные ранее функции расстояния обобщаются для учета важности признаков путем введения их весов. Взвешенная функция сходства обобщается на случай иерархических моделей. Вычисляются оценки весов взвешенной функции сходства с помощью обобщения энтропийного подхода. Для вероятностной модели коллекции документов, основанной на иерархической функции сходства, предлагается способ оценки апостериорного распределения параметров, а также совместного апостериорного распределения параметров и классов неразмеченных документов. Доказываются свойства полученных оценок.

Практическая значимость. Предложенные в работе методы предназначены для иерархической классификации коллекций текстов с учетом существующих экспертных моделей; выявления тематических несоответствий в экспертных моделях и значимого повышения тематической целостности уже построенных тематических моделей с помощью небольшого числа изменений; визуализации иерархических моделей и выявленных несоответствий на плоскости.

Степень достоверности и апробация работы. Достоверность результатов подтверждена математическими доказательствами, экспериментальной

проверкой полученных методов на реальных задачах иерархической классификации коллекций тезисов конференции и коллекций сайтов индустриального сектора; 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.
2. Международная конференция “20th Conference of the International Federation of Operational Research Societies”, 2014.
3. Всероссийская конференция “Математические методы распознавания образов” ММРО-17, 2015.
4. Всероссийская конференция “58 научная конференция МФТИ”, 2015.
5. Всероссийская конференция “Ломоносов-2016”, 2016.
6. Международная конференция “28th European Conference on Operational Research”, 2016.

Работа поддержана грантами 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 диссертации. Основные результаты по теме диссертации изложены в 10 печатных изданиях, 4 из которых изданы в журналах, рекомендованных ВАК. Список публикаций приведен в конце автореферата.

Личный вклад. Все приведенные результаты, кроме отдельно оговоренных случаев, получены диссертантом лично при научном руководстве д.ф.-м.н. В. В. Стрижова.

Структура и объем работы. Диссертация состоит из оглавления, введения, пяти разделов, заключения, списка иллюстраций, списка таблиц, перечня основных обозначений и списка литературы из 123 наименований. Текст работы занимает 120 страниц.

Основное содержание работы

Во введении обоснована актуальность диссертационной работы, сформулированы цели и методы исследования, поставлены основные задачи, обоснована научная новизна, теоретическая и практическая значимость полученных результатов, приведено краткое содержание работы по главам.

В **главе 1** вводятся основные определения, рассматриваются основные этапы классификации и кластеризации коллекций документов существующими методами: предобработка коллекции текстовых документов, составление словаря коллекции, представление слов и документов в виде векторов, построение модели. Рассматриваются четыре основных подхода построения тематической модели: с помощью жестких методов, описательно-вероятностных методов, смесей моделей и вероятностных методов. Рассматриваются существующие варианты алгоритмов иерархической классификации.

Определение 1. Текстовым документом d называется множество слов $\{w_m\}$. Коллекцией документов D называется множество документов $\{d_n\}$. Размером коллекции $|D|$ называется число элементов данного множества.

Определение 2. Словарем W коллекции D называется упорядоченное подмножество неповторяющихся слов w, содержащихся в коллекции D.

Для решения задач кластеризации и классификации, каждый документ d коллекции D представляется в виде вектора $\mathbf{x} \in \mathbb{R}^{|W|}$:

$$\mathbf{x} = [\phi(w_1, d, D), \cdots, \phi(w_{|W|}, d, D)]^{\intercal}, \tag{1}$$

где $\phi(w, d, D)$ – функция, ставящая в соответствие слову w из W действительное число.

Определение 3. Кластером c называется подмножество документов коллекции D. Корневым кластером называется кластер, содержащий все документы коллекции D. Документ d имеет класс c, если $d \in c$.

Определение 4. Кластер c_1 является родительским кластером c_2 если все документы d из c_2 содержатся в c_1. При этом кластер c_2 называется дочерним кластером c_1.

Определение 5. Тематической моделью M текстовой коллекции D называется разбиение D на кластеры $\{c_1, c_2, \ldots, c_n\}$ таким образом, чтобы каждый документ $d \in D$ принадлежал хотя бы одному кластеру помимо корневого.

Тематическая модель M коллекции D называется экспертной, если для каждого документа $d \in D$ его классы задавались экспертами. Тематическая модель $\hat{M}$ называется алгоритмической, если для некоторых документов классы задавались алгоритмическим образом. Иерархическая кластерная структура задана экспертно, если изначально задан граф ее кластеров. В данной работе рассматриваются иерархические структуры кластеров в виде сбалансированных деревьев.

Каждый кластер $c_{l,k}$ индексируется двумя числами – уровнем l и порядковым номером на данном уровне k. Корневой кластер обозначается как $c_{1,1}$, число кластеров на уровне l обозначается K_l, нижний уровень иерархии обозначается h. Так как рассматриваются древовидные структуры, класс документа на нижнем уровне дерева определяет его классы на всех остальных уровнях.

Определение 6. Матрицей $\mathbf{Z}$ называется матрица экспертной кластеризации размеченных документов на нижнем уровне, $z_{nk} = [d \in c_{h,k}]$, экспертный кластер документа d на нижнем уровне иерархии обозначается $c(d)$.

В **главе 2** анализируются способы векторного представления документов коллекции, рассматривается взвешенная метрика минковского в качестве способа сравнения векторных представлений документов. Предлагается способ оптимизации весов этой метрики, а также анализируются агломеративный и дивизимный способы плоской и иерархической кластеризации документов с помощью найденной оптимальной метрики.

Для сравнения документов используется взвешенная метрика Минковского с параметром $p \geq 1$ и вектором важности слов $\boldsymbol{\lambda} \in \mathbb{R}^{|W|}$:

$$\rho(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{y}) = \sqrt[p]{\sum_{m=1}^{|W|} \lambda_m |x_m - y_m|^p}, \quad \text{где} \quad \boldsymbol{\lambda} \geq \mathbf{0}, \quad \|\boldsymbol{\lambda}\|_1 = 1. \tag{2}$$

Определение 7. Признак с индексом m называется активным, если соответствующий ему вес $\lambda_m > 0$, $\mathcal{A}$ – множество всех активных признаков.

Качество взвешенной метрики $\rho(\boldsymbol{\lambda}, \mathbf{x}, \mathbf{y})$ задается как

$$V(\rho, \boldsymbol{\lambda}, D) = \frac{1}{|D|} \sum_{\mathbf{x} \in D} \frac{\overline{r}_k(\mathbf{x}) - r_k(\mathbf{x})}{\overline{r}_k(\mathbf{x}) + r_k(\mathbf{x})}, \tag{3}$$

где $r_k(\mathbf{x})$ и $\overline{r}_k(\mathbf{x})$ – расстояние от документа $\mathbf{x}$ до k ближайших соседей из его кластера $\hat{c}(\mathbf{x})$ и остальных кластеров $\{c \neq \hat{c}(\mathbf{x})\}$ соответственно. Для поиска оптимальных весов решается задача оптимизации качества метрики

$$\hat{\boldsymbol{\lambda}} = \arg\max_{\boldsymbol{\lambda}} V(\rho, \boldsymbol{\lambda}, D).$$

Для этого используется итеративный алгоритм. Изначально веса всех признаков $\boldsymbol{\lambda} = \mathbf{0}$. На каждом шаге повторяются два действия.

1. Для каждого признака $m \notin \mathcal{A}$ найти набор весов $\hat{\boldsymbol{\lambda}}_m$ для множества признаков $\{\mathcal{A} \cup m\}$, доставляющий максимум функции качества $V(\rho, \hat{\boldsymbol{\lambda}}_m, D) = V_m$.
2. Добавить в $\mathcal{A}$ признак m^*, которому соответствует максимальное качество V_m и обновить вектор весов $\boldsymbol{\lambda} = \hat{\boldsymbol{\lambda}}_{m^*}$.

Алгоритм повторяется до тех пор, пока увеличивается значение функции качества $V(\rho, \boldsymbol{\lambda}, D_{\mathcal{T}})$ на контрольной выборке $D_{\mathcal{T}}$.

Глава 3 посвящена задаче классификации частично размеченной коллекции документов с помощью предложенной иерархической функции сходства.

Определение 8. Оператором релевантности называется оператор R, ставящий в соответствие документу $\mathbf{x} \in \mathbb{R}^{|W|}$, перестановку кластеров нижнего уровня, отсортированных по релевантности документу $\mathbf{x}$

$$R : \mathbb{R}^{|W|} \to S^{K_h}. \tag{4}$$

Решением задачи классификации для документа $\mathbf{x}$ является кластер, стоящий на первой позиции перестановки $R(\mathbf{x})$. При большом количестве кластеров, экспертное мнение в общем случае не совпадает с решением задачи классификации. В этом случае эксперт рассматривает следующие по релевантности кластеры в качестве альтернативных решений.

Определение 9. Качеством оператора релевантности R называется

$$\mathrm{AUCH}(R) = \frac{1}{K_h|D|} \sum_{j=1}^{K_h} |\{\mathbf{x} : \mathrm{pos}\big(R(\mathbf{x}), c(\mathbf{x})\big) \leq j\}|, \tag{5}$$

где $\{\mathbf{x} : \mathrm{pos}\big(R(\mathbf{x}), c(\mathbf{x})\big) \leq j\}$ – множество всех документов $\mathbf{x}$, для которых номер позиции $\mathrm{pos}\big(R(\mathbf{x}), c(\mathbf{x})\big)$ экспертного кластера $c(\mathbf{x})$ в перестановке $R(\mathbf{x})$ меньше либо равен j.

Чем больше значение AUCH, тем ближе к началу находится экспертный кластер в возвращаемых перестановках. Значение $\mathrm{AUCH}(R) = 1$ соответствует случаю, когда экспертный кластер оказывается в соответствии с R наиболее релевантным для каждого из документов выборки D.

Для определения релевантности кластера $c_{h,k}$ документу $\mathbf{x}$ вводится иерархическая функция сходства $s_{\mathsf{h}}(\mathbf{x}, c_{h,\,k})$, вычисляющая сходство документа с веткой дерева экспертной кластерной структуры.

Определение 10. Иерархическое сходство s_{h} документа $\mathbf{x}_n$ и кластера $c_{h,\,k}$ нижнего уровня h определяется как

$$s_{\mathsf{h}}(\mathbf{x}_n, c_{h,\,k}) = \mathbf{x}^{\mathsf{T}}\boldsymbol{\Lambda}\mathbf{M}_k\boldsymbol{\theta}_k = s_{n,k}, \quad \mathbf{M}_k = [\boldsymbol{\mu}_{1,k}, \ldots, \boldsymbol{\mu}_{h,k}], \tag{6}$$

где $\boldsymbol{\mu}_{l,k}$ – средний вектор родительского кластера на уровне l для кластера $c_{h,k}$, параметры $\boldsymbol{\theta}_k$ определяют с каким весом учитывается сходство с кластерами разных уровней, а диагональная матрица $\boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \ldots, \lambda_{|W|})$ определяет важности слов.

Для оценки параметров матрицы $\mathbf{\Lambda}$ предлагается энтропийная модель важности слов, ставящая в соответствие важности λ_m слова w_m значение функции, зависящей от энтропии данного слова и вектора параметров $\boldsymbol{\alpha}$. Пусть $p^l_{m,k} = p(c_{l,k}|w_m)$. Оценка $p^l_{m,k}$ через средние векторы кластеров:

$$\mathbf{p}^l_m = [\mu_{l,1,m}, \ldots, \mu_{l,K_l,m}]^\intercal, \quad \mathbf{p}^l_m \mapsto \frac{\mathbf{p}^l_m}{\|\mathbf{p}^l_m\|}, \quad m \in \{1, \ldots, |W|\}.$$

Определение 11. Энтропией слова w_m относительно экспертной кластеризации документов на уровне l называется

$$\mathsf{H}^l(w_m) = -\sum_{k=1}^{K_l} p^l_{m,k} \log(p^l_{m,k}). \tag{7}$$

Определение 12. Важность λ_m слова w_m определяется через его энтропию

$$\lambda_m = 1 + \boldsymbol{\alpha}^\intercal \boldsymbol{\iota}_m, \quad \iota_{ml} = \log\bigl(1 + \mathsf{H}^l(w_m)\bigr), \tag{8}$$

где вектор параметров $\boldsymbol{\alpha} \in \mathbb{R}^h$ определяет, с каким весом учитывается энтропия относительно различных уровней иерархии.

Для оптимизации параметров $\boldsymbol{\alpha}$ и $\boldsymbol{\theta} = \{\boldsymbol{\theta}_k\}$ иерархической функции сходства по размеченным документам $D = D_{\mathcal{V}_1} \cup D_{\mathcal{V}_2}$ предлагается итеративный алгоритм.

1. По $D_{\mathcal{V}_1}$ при фиксированных $\boldsymbol{\theta}_k$ найти оптимальные значения параметров $\boldsymbol{\alpha}$ энтропийной модели (8):

$$\boldsymbol{\alpha}^* = \arg\max_{\boldsymbol{\alpha}} \mathrm{AUCH}(R). \tag{9}$$

2. По $D_{\mathcal{V}_2}$ найти оптимальные $\boldsymbol{\theta}_k$ при фиксированных $\boldsymbol{\alpha}$, решив k задач выпуклого квадратичного программирования:

$$\boldsymbol{\theta}^*_k = \arg\max_{\boldsymbol{\theta}_k} \sum_{\mathbf{x} \in c_{h,k}} \mathbf{x}^\intercal \mathbf{\Lambda} \mathbf{M}_k \boldsymbol{\theta}_k + \psi \|\boldsymbol{\theta}_k - \mathbf{h}\|^2_2, \tag{10}$$

$$\|\boldsymbol{\theta}_k\|_1 = 1, \quad \boldsymbol{\theta}_k \geq \mathbf{0}, \quad k \in \{1 \ldots K_h\}, \quad \mathbf{h} = \left[\frac{1}{h}, \ldots, \frac{1}{h}\right]^\intercal, \tag{11}$$

где $\psi < 0$ – параметр регуляризации.

Критерий качества AUCH (5) является дискретным, что усложняет его оптимизацию по параметрам модели $\boldsymbol{\theta}$ и $\boldsymbol{\alpha}$. Так, сложность $O(ba^h|D||W|hK_h)$ описанного выше алгоритма растет экспоненциально при увеличении числа уровней, что делает данный алгоритм плохо масштабируемым. Чтобы получить вычислительно эффективный метод оптимизации параметров, вместо максимизации критерия AUCH максимизируется правдоподобие модели.

Определение 13. Вероятность документа $\mathbf{x}_n$ принадлежать кластеру нижнего уровня $c_{h,k}$ оценивается с помощью функции softmax от результата иерархической функции сходства:

$$p(z_{nk}=1|\mathbf{x}_n,\boldsymbol{\theta},\boldsymbol{\alpha})=\frac{\exp\bigl(s_{\mathsf{h}}(\mathbf{x}_n,c_{h,k})\bigr)}{\sum_{k'=1}^{K_h}\exp\bigl(s_{\mathsf{h}}(\mathbf{x}_n,c_{h,k'})\bigr)}, \tag{12}$$

где $z_{nk}=[\mathbf{x}_n\in c_{h,k}]$ – элемент матрицы экспертной классификации.

Правдоподобие модели задается как

$$L(\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\alpha})=\prod_{n=1}^{N}\prod_{k=1}^{K_h}p(z_{nk}=1|\mathbf{x}_n,\boldsymbol{\theta},\boldsymbol{\alpha})^{z_{nk}}. \tag{13}$$

В качестве дополнительных ограничений на параметры $\boldsymbol{\theta}_k$ и $\boldsymbol{\alpha}$ используются их априорные распределения

$$p(\boldsymbol{\alpha})=\mathcal{N}(\boldsymbol{\alpha}|\mathbf{0},a^{-1}\mathbf{I}),\quad p(\boldsymbol{\theta}_k)=\mathcal{N}(\boldsymbol{\theta}_k|\mathbf{m}_k,\mathbf{V}_k^{-1}). \tag{14}$$

Вектор параметров $\boldsymbol{\theta}_k$ имеет неизвестные гиперпараметры $\mathbf{m}_k$ и $\mathbf{V}_k$, поэтому на них также накладываются априорные распределения

$$p(\mathbf{m}_k|\mathbf{V}_k)=\mathcal{N}(\mathbf{m}_k|\mathbf{m}_0,(b\mathbf{V}_k)^{-1}),\quad p(\mathbf{V}_k)=\mathcal{W}(\mathbf{V}_k|\mathbf{W},\nu), \tag{15}$$

где $\mathcal{W}$ – распределение Уишарта. Общая вероятностная модель имеет вид:

$$p(\mathbf{Z},\boldsymbol{\theta},\mathbf{m},\mathbf{V},\boldsymbol{\alpha})=L(\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\alpha})p(\boldsymbol{\theta}|\mathbf{m},\mathbf{V})p(\mathbf{m}|\mathbf{V})p(\mathbf{V})p(\boldsymbol{\alpha}). \tag{16}$$

Из-за нелинейной зависимости правдоподобия L от параметров модели $\boldsymbol{\theta}$ и $\boldsymbol{\alpha}$, аналитический вывод апостериорного распределения параметров невозможен. Вместо этого предлагается искать функцию q из класса

$$q(\boldsymbol{\theta},\mathbf{m},\mathbf{V},\boldsymbol{\alpha})=q(\boldsymbol{\theta})q(\mathbf{m},\mathbf{V},\boldsymbol{\alpha}), \tag{17}$$

минимизирующую расстояние $\mathrm{KL}(q\|p)$ с истинным апостериорным распределением. Для функции плотности вероятности q, распределение наблюдаемых переменных $p(\mathbf{Z})$ представимо в виде

$$\begin{aligned}&\ln p(\mathbf{Z})=\mathcal{L}(q)+\mathrm{KL}(q\|p),\quad \text{где}\\ &\mathcal{L}(q)=\int q(\boldsymbol{\theta},\mathbf{m},\mathbf{V},\boldsymbol{\alpha})\ln\left(\frac{p(\mathbf{Z},\boldsymbol{\theta},\mathbf{m},\mathbf{V},\boldsymbol{\alpha})}{q(\boldsymbol{\theta},\mathbf{m},\mathbf{V},\boldsymbol{\alpha})}\right)\mathrm{d}\boldsymbol{\theta}\mathrm{d}\mathbf{m}\mathrm{d}\mathbf{V}\mathrm{d}\boldsymbol{\alpha}.\end{aligned} \tag{18}$$

Минимизация расстояние $\mathrm{KL}(q\|p)$ по q эквивалентна максимизации нижней границы $\mathcal{L}(q)$. При этом оптимальный вид каждого из факторов $q(\boldsymbol{\theta})$ и $q(\mathbf{m},\mathbf{V},\boldsymbol{\alpha})$, при фиксированном другом факторе задается как

$$\begin{aligned}\ln q(\mathbf{m},\mathbf{V},\boldsymbol{\alpha})&=\mathsf{E}_{\boldsymbol{\theta}}\bigl[\ln p(\mathbf{Z},\boldsymbol{\theta},\mathbf{m},\mathbf{V},\boldsymbol{\alpha})\bigr]+\mathrm{const}(\mathbf{m},\mathbf{V},\boldsymbol{\alpha}),\\ \ln q(\boldsymbol{\theta})&=\mathsf{E}_{\mathbf{m},\mathbf{V},\boldsymbol{\alpha}}\bigl[\ln p(\mathbf{Z},\boldsymbol{\theta},\mathbf{m},\mathbf{V},\boldsymbol{\alpha})\bigr]+\mathrm{const}(\boldsymbol{\theta}),\end{aligned} \tag{19}$$

где $\text{const}(\cdot)$ – функция, не зависящая от аргумента. Алгоритм оптимизации параметров распределения q, итеративно обновляющий факторы (19), сходится.

Для аналитического вывода факторов q, вместо $\mathcal{L}(q)$ используется ее верхняя оценка $\hat{\mathcal{L}}(q, \boldsymbol{\xi})$, полученная с помощью оценки знаменателя функции softmax в правдоподобии L:

$$\frac{1}{g(\mathbf{s}_n)} \leq \frac{1}{g(\boldsymbol{\xi}_n)} \exp\left(\sum_{k=1}^{K_h} \frac{\exp(\xi_{nk})}{g(\boldsymbol{\xi}_n)}(\xi_{nk} - s_{n,k})\right), \quad g(\mathbf{s}_n) = \sum_{k=1}^{K_h} \exp(s_{n,k}), \tag{20}$$

где $\boldsymbol{\xi}_n$ – вектор вариационных параметров, соответствующих $\mathbf{s}_n$.

Теорема 1. Функцией q из класса (17), заданной оптимальными оценками факторов (19), в которых правдоподобие L определяется с помощью (13) и верхней оценки функции softmax (20), а априорные распределения параметров $\boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha}$ задаются как (15) и (14), является

$$q = \mathcal{N}(\boldsymbol{\alpha}_0, a^{-1}\mathbf{I}) \prod_{k=1}^{k_h} \mathcal{N}(\mathbf{m}'_{0k}, (\nu' \mathbf{V}_k)^{-1}) \mathcal{N}(\mathbf{m}_{0k}, (b' \mathbf{V}_k)^{-1}) \mathcal{W}(\mathbf{W}_k, \nu'), \tag{21}$$

$$\mathbf{m}'_{0k} = \mathbf{m}_{0k} + \frac{1}{\nu'}(\mathbf{W}_k^{-1})^{\intercal} \mathbf{M}_k^{\intercal} \mathsf{E}_{\boldsymbol{\alpha}}[\boldsymbol{\Lambda}] \sum_{n=1}^{N} \mathbf{x}_n \left(z_{nk} - \frac{\exp(\xi_{nk})}{g(\boldsymbol{\xi}_n)}\right),$$

$$\boldsymbol{\alpha}_0 = \frac{1}{a} \sum_{n=1}^{N} \sum_{m=1}^{|W|} x_{nm} \boldsymbol{\iota}_m \sum_{k=1}^{k_h} (\mathbf{M}_k \mathsf{E}[\boldsymbol{\theta}_k])_m \left(z_{nk} - \frac{\exp(\xi_{nk})}{g(\boldsymbol{\xi}_n)}\right),$$

$$\mathbf{W}_k^{-1} = (b+1)\mathbf{m}_{0k}\mathbf{m}_{0k}^{\intercal} + b\mathbf{m}_0\mathbf{m}_0^{\intercal} + \mathsf{E}\left[\boldsymbol{\theta}_k \boldsymbol{\theta}_k^{\intercal}\right] + \mathbf{W}^{-1},$$

$$\mathbf{m}_{0k} = \frac{\mathsf{E}[\boldsymbol{\theta}_k] + b\mathbf{m}_0}{b+1}, \quad b' = b+1, \quad \nu' = \nu + 1.$$

Теорема 2. Значения вариационных параметров ξ_{nk}, минимизирующие $\hat{\mathcal{L}}(q, \boldsymbol{\xi})$ при фиксированных параметрах распределения q, совпадают со значением иерархической функции сходства в точке $\mathbf{x}_n$ для класса k, использующей в качестве параметров $\hat{\boldsymbol{\theta}}_k = \mathsf{E}\boldsymbol{\theta}_k = \mathbf{m}'_{0k}$, $\tilde{\boldsymbol{\Lambda}} = \mathsf{E}_{\boldsymbol{\alpha}}\boldsymbol{\Lambda} = \text{diag}(\{\lambda'_m\})$, $\lambda'_m = 1 + \boldsymbol{\alpha}_0^{\intercal}\boldsymbol{\iota}_m$.

Параметры распределения (21) содержат неизвестные математические ожидания $\mathsf{E}\boldsymbol{\theta}_k, \mathsf{E}[\boldsymbol{\theta}_k\boldsymbol{\theta}_k^{\intercal}], \mathsf{E}_{\boldsymbol{\alpha}}\boldsymbol{\Lambda}$ и вариационные параметры $\boldsymbol{\xi}_n$. Для поиска их значений используется следующий EM-алгоритм.

1. Инициализировать параметры $\mathbf{W}, \nu, \mathbf{m}_0, a, b, \mathbf{W}_k, \mathbf{m}_{0k}, \boldsymbol{\xi}_n$.
2. Пересчитать параметры $\mathbf{m}'_{0k}$ и найти $\mathsf{E}[\boldsymbol{\theta}_k]$, используя $q(\boldsymbol{\theta}_k)$.
3. Пересчитать параметры $\mathbf{m}_{0k}, \mathbf{W}_k, \boldsymbol{\alpha}_0$ и найти $\mathsf{E}_{\boldsymbol{\alpha}}[\boldsymbol{\Lambda}]$, используя $q(\boldsymbol{\alpha})$.
4. Уточнить вариационные параметры $\xi_{nk} = \mathbf{x}_n^{\intercal}\tilde{\boldsymbol{\Lambda}}\mathbf{M}_k\mathbf{m}'_{0k}$.
5. При значимом изменении параметров вернуться на шаг 2.

Для предсказания класса неразмеченного документа $\tilde{\mathbf{x}}_t$ с помощью найденного апостериорного распределения строятся два оператора релевантности: 1) R_1, ранжирующий кластеры по значению иерархической функции сходства с параметрами $\boldsymbol{\theta}_k^{\text{MAP}}$ и $\boldsymbol{\alpha}^{\text{MAP}}$, максимизирующими найденную аппроксимацию q апостериорного распределения, 2) R_2, ранжирующий по вероятности принадлежности кластеру

$$p(\tilde{z}_{tk}|\tilde{\mathbf{x}}_t) = \int \text{softmax}\big(\mathsf{s}_{\mathsf{h}}(\tilde{\mathbf{x}}_t|\boldsymbol{\theta},\boldsymbol{\alpha})\big)_k q(\boldsymbol{\theta},\boldsymbol{\alpha})\mathrm{d}\boldsymbol{\theta}\mathrm{d}\boldsymbol{\alpha}. \tag{22}$$

Чтобы взять интеграл (22) аналитически, используется верхняя оценка softmax с вариационными параметрами $\tilde{\boldsymbol{\xi}}_t$ и нижняя оценка экспоненты с вариационным параметром ψ_{tk}. Параметры $\tilde{\boldsymbol{\xi}}_t, \psi_{tk}$ полученной оценки вероятности находятся с помощью оптимизации:

$$\hat{p}(\tilde{z}_{tk}|\tilde{\mathbf{x}}_t)^* = \max_{\psi_{tk}} \min_{\tilde{\boldsymbol{\xi}}_t} \hat{p}(\tilde{z}_{tk}|\tilde{\mathbf{x}}_t, \tilde{\boldsymbol{\xi}}_t, \psi_{tk}). \tag{23}$$

Теорема 3. Значения качества построенных операторов релевантности $\text{AUCH}(R_1)$ и $\text{AUCH}(R_2)$ при оптимальных значениях параметров $\tilde{\boldsymbol{\xi}}_t, \psi_{tk}$ совпадают.

Для оценки интеграла (22) предлагается вместо апостериорного распределения аппроксимировать совместное апостериорное распределение параметров и классов неразмеченных документов, приближающее все подынтегральное выражение. Совместная вероятностная модель (16) принимает вид

$$p(\tilde{\mathbf{Z}}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha}) = p(\tilde{\mathbf{Z}}|\boldsymbol{\theta},\boldsymbol{\alpha})p(\mathbf{Z}, \boldsymbol{\theta}, \boldsymbol{\alpha}, \mathbf{m}, \mathbf{V}). \tag{24}$$

Аппроксимация q совместного апостериорного распределения параметров и классов неразмеченных документов $p(\tilde{\mathbf{Z}}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha}|\mathbf{Z})$ ищется в классе

$$q(\tilde{\mathbf{Z}}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha}) = q(\boldsymbol{\theta})q(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{V})q(\tilde{\mathbf{Z}}). \tag{25}$$

Оптимальным видом каждого из факторов при фиксированных остальных факторах является

$$\begin{aligned}
\ln q(\boldsymbol{\theta}) &= \mathsf{E}_{\boldsymbol{\alpha},\mathbf{m},\mathbf{V},\tilde{\mathbf{Z}}}\big[\ln p(\tilde{\mathbf{Z}}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})\big] + \text{const}(\boldsymbol{\theta}),\\
\ln q(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{V}) &= \mathsf{E}_{\boldsymbol{\theta},\tilde{\mathbf{Z}}}\big[\ln p(\tilde{\mathbf{Z}}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})\big] + \text{const}(\boldsymbol{\alpha}, \mathbf{m}, \mathbf{V}),\\
\ln q(\tilde{\mathbf{Z}}) &= \mathsf{E}_{\boldsymbol{\alpha},\mathbf{m},\mathbf{V},\boldsymbol{\theta}}\big[\ln p(\tilde{\mathbf{Z}}, \mathbf{Z}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha})\big] + \text{const}(\tilde{\mathbf{Z}}).
\end{aligned} \tag{26}$$

Для аналитического вывода q используется верхняя оценка softmax в правдоподобии $L(\mathbf{Z}|\boldsymbol{\theta},\boldsymbol{\alpha})$ с вариационными параметрами $\boldsymbol{\xi} = \{\boldsymbol{\xi}_n\}$ и в распределении классов неразмеченных документов $p(\tilde{\mathbf{Z}}|\boldsymbol{\theta},\boldsymbol{\alpha})$ с вариационными параметрами $\tilde{\boldsymbol{\xi}} = \{\tilde{\boldsymbol{\xi}}_t\}$. В результате получается оценка $\hat{\mathcal{L}}(q, \boldsymbol{\xi}, \tilde{\boldsymbol{\xi}})$ функции $\mathcal{L}(q)$.

Теорема 4. Функцией q из класса (25), заданной оптимальными оценками факторов (26), в которых правдоподобие L определяется с помощью (13) и верхней оценки функции softmax (20), а априорные распределения параметров $\boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha}$ задаются как (15) и (14), является

$$q(\tilde{\mathbf{Z}}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha}) = q(\boldsymbol{\alpha}) \prod_{k=1}^{K_h} q(\boldsymbol{\theta}_k) q(\mathbf{m}_k | \mathbf{V}_k) q(\mathbf{V}_k) \prod_{t=1}^{|D_\mathcal{T}|} q(\tilde{z}_{tk}), \quad \text{где}$$

$$q(\tilde{z}_{tk}) \sim \text{Bern}(p_{tk}), \quad p_{tk} = \exp(\zeta_{tk}) \big(\exp(\zeta_{tk}) + g(\tilde{\boldsymbol{\xi}}_t)\big)^{-1},$$

$$\zeta_{tk} = \tilde{\mathbf{x}}_t^\intercal \mathsf{E}_{\boldsymbol{\alpha}}[\boldsymbol{\Lambda}] \mathbf{M}_k \mathsf{E}[\boldsymbol{\theta}_k] + \sum_{k'=1}^{K_h} \frac{\exp(\tilde{\xi}_{tk'})}{g(\tilde{\boldsymbol{\xi}}_t)} (\tilde{\xi}_{tk'} - \tilde{\mathbf{x}}_t^\intercal \mathsf{E}_\alpha[\boldsymbol{\Lambda}] \mathbf{M}_{k'} \mathsf{E}[\boldsymbol{\theta}_{k'}]).$$

Для оптимизации параметров из теоремы 4 используется аналогичный EM-алгоритм, как и в случае приближения апостериорного распределения. Искомая вероятность $p(\tilde{z}_{tk} | \tilde{\mathbf{x}}_t)$ выражается через найденную оценку совместного апостериорного распределения параметров и классов неразмеченных документов q:

$$p(\tilde{z}_{tk} | \tilde{\mathbf{x}}_t) \approx \int q(\tilde{\mathbf{Z}}, \boldsymbol{\theta}, \mathbf{m}, \mathbf{V}, \boldsymbol{\alpha} | \tilde{\mathbf{x}}_t) \mathrm{d}\boldsymbol{\theta} \mathrm{d}\mathbf{m} \mathrm{d}\mathbf{V} \mathrm{d}\boldsymbol{\alpha} \mathrm{d}\tilde{\mathbf{Z}}_{(-tk)} = \text{Bern}(p_{tk}), \tag{27}$$

где $\mathrm{d}\tilde{\mathbf{Z}}_{(-tk)}$ означает интегрирование по всем $\tilde{z}$ кроме $\tilde{z}_{tk}$. Задача классификации нового документа сводится к

$$\hat{c}(\tilde{\mathbf{x}}_t) = \arg\max_k p_{tk},$$

а оператор релевантности R строится путем ранжирования по найденным вероятностям p_{tk}.

Теорема 5. Пусть выполняются следующие соотношения:

$$|D_\mathcal{V}| \sim |D_\mathcal{T}| \sim |D|, \quad K_h < |D|, \quad h < K_h.$$

Сложность EM-алгоритма настройки параметров распределения q из теоремы 4 равна $O(a|D||W|hK_h)$, где a – число EM шагов.

В **главе 4** рассматривается задача верификации экспертной модели путем изменения классов фиксированного числа документов.

Определение 14. Качество $\Xi(M)$ иерархической тематической модели M определяется как

$$\Xi(M) = \sum_{l=1}^{h} \left[\frac{1-\beta}{K_l} \sum_{k=1}^{K_l} |c_{l,k}| s_\mathsf{c}(c_{l,k},\ c_{l,k}) - \frac{2\beta}{K_l(K_l-1)} \sum_{k=1}^{K_l} \sum_{k'=k+1}^{K_l} s_\mathsf{c}(c_{l,k},\ c_{l,k'}) \right],$$

где s_c – среднее сходство между документами из указанных кластеров, а структурный параметр β отвечает за приоритет межкластерного сходства.

Каждому документу $\mathbf{x}$ ставится в соответствие вектор $\boldsymbol{\zeta}(\mathbf{x}) \in \mathbb{R}^h$, в котором элемент ζ_l равняется единице, если на уровне l экспертный кластер этого документа совпадает с его алгоритмическим кластером. В силу древовидности рассматриваемых иерархических кластерных структур, всего возможно h вариантов вектора $\boldsymbol{\zeta}(\mathbf{x})$, а сумма элементов $\|\boldsymbol{\zeta}(\mathbf{x})\|_1$ показывает на скольких уровнях экспертная кластеризация $\mathbf{x}$ совпадает с алгоритмической. Каждой операции переноса документа $\mathbf{x}$ из кластера $c_{h,k}$ в кластер $c_{h,k'}$ ставится в соответствие пара векторов $\boldsymbol{\zeta}$ вида

$$\boldsymbol{\zeta}\big(\mathbf{x}, c(\mathbf{x}), c_{h,k}\big) \mapsto \boldsymbol{\zeta}\big(\mathbf{x}, c(\mathbf{x}), c_{h,k'}\big),$$

а каждому уникальному варианту переноса $\boldsymbol{\zeta} \mapsto \boldsymbol{\zeta}'$ ставится в соответствие штраф $\delta(\boldsymbol{\zeta} \mapsto \boldsymbol{\zeta}')$ за его осуществление. Всего возможно h^2 различных штрафов, для их определения используется матрица размером $h \times h$, на которую накладывается набор ограничений.

Оптимизационный алгоритм на каждом шаге ищет для всех документов $\mathbf{x}$ наиболее подходящий кластер и осуществляет перенос документа, дающий наибольший прирост качества модели за вычетом штрафа за данный перенос. Обозначим Ξ_1 значение качества модели до переноса $\boldsymbol{\zeta} \mapsto \boldsymbol{\zeta}'$ документа $\mathbf{x}$, и Ξ_2 – значение после переноса. Перенос осуществляется только при выполнении условия

$$\Xi_2 - \Xi_1 \geq \gamma\delta(\boldsymbol{\zeta} \mapsto \boldsymbol{\zeta}'),$$

где $\gamma \geq 0$ – весовой множитель штрафов, регулирующий допустимую степень несоответствия построенной кластеризации и экспертной. Свойства предложенного алгоритма анализировались при решении задачи верификации экспертной модели конференции.

В **главе 5** приводится описание разработанного программного комплекса и результаты анализа предложенных методов, а также сравнение их с существующими аналогами. Приводится алгоритм вложеннной визуализации иерархической кластерной структуры на плоскости, основанный на методе проекций Саммона, позволяющий отображать выявленные тематические несоответствия алгоритмом, описанным в главе 4, и способы их устранения.

В качестве прикладной задачи рассматривалась задача иерархической классификации тезисов предстоящей крупной конференции European Conference on Operational Research (EURO) по имеющимся экспертным моделям этой конференции с 2006 по 2016 год. Программа данной конференции имеет структуру дерева, в котором кластерами второго уровня являются научные областей (Area), а кластерами третьего уровня являются научные направления (Stream).

Для проверки предложенных методов – иерархической взвешенной функции сходства hSim (6) и ее аналога, построенного с помощью обученной языковой модели и векторного представления слов hSimWV, их результаты сравнивались с результатами других алгоритмов: иерархического наивного байеса hNB,

вероятностной модели с адаптивной иерархической регуляризацией SuhiPLSA и иерархического мультиклассового svm.

Коллекция документов D делилась на две части: обучающую $D_{\mathcal{V}}$ и тестовую $D_{\mathcal{T}}$. Для анализа работы алгоритмов при различном размере обучающей выборки $|D_{\mathcal{V}}|$, ее размер менялся от 500 до 10000 документов. С помощью каждого из алгоритмов строился оператор релевантности R (4), после чего его качество оценивалось на тестовой выборке $D_{\mathcal{T}}$ как площадь под огибающей кумулятивной гистограммой AUCH (5). В таблице 1 приведены результаты сравнения перечисленных алгоритмов, жирным шрифтом выделены лучшие статистически эквивалентные результаты для каждого размера выборки.

Таблица 1. Значения функционала качества AUCH (5) для операторов релевантности, построенных с помощью сравниваемых алгоритмов.

Размер выборки $\lvert D_{\mathcal{L}}\rvert$ / Алгоритм	500	1000	1500	3000	5000	7000	10000
svm	0.76	0.80	0.81	0.84	0.85	0.86	0.87
hNB	0.77	0.82	0.84	0.87	0.90	0.91	0.92
suhiPLSA	0.75	0.79	0.80	0.81	0.82	0.84	0.84
hSim	0.80	**0.86**	**0.88**	**0.90**	**0.91**	**0.92**	**0.93**
hSimWV	**0.82**	**0.86**	0.87	**0.89**	**0.92**	**0.92**	0.92

В качестве второй прикладной задачи рассматривалась задача классификации веб-сайтов индустриального сектора по индустриям и отраслям. Для этого специальные символы и тэги веб-сайтов отбрасывались, и текст всех страниц рассматривался как единый документ. Для случая иерархической классификации предложенный алгоритм показал наилучший результат.

В заключении представлены основные результаты диссертационной работы.

1. Поставлены задачи иерархической кластеризации и классификации, рассмотрены основные этапы построения иерархических тематических моделей коллекции документов: методы представления слов и документов в виде векторов и методы иерархической кластеризации и классификации, включающие в себя алгоритмы построения жестких, вероятностных, описательно-вероятностных моделей и смесей моделей.
2. Проанализированы способы векторного представления документа с помощью булевых, целочисленных и частотных признаков слов. Предложен способ оценки качества взвешенной метрики и алгоритм построения локально оптимального набора весов. Проведено сравнение агломеративного и дивизимного подходов иерархической кластеризации с помощью полученной метрики.

3. Предложена взвешенная функция сходства документа и кластера, учитывающая информативность слов в задачах кластеризации и классификации. Предложен алгоритм оптимизации параметров взвешенной функции сходства, использующий энтропию слов относительно экспертной кластеризации на различных уровнях иерархии. Предложен иерархический вариант взвешенной функции сходства, позволяющий вычислять сходство документа с веткой экспертной иерархической кластерной структуры. Для классификации нового документа предложен оператор релевантности, возвращающий ранжированный список кластеров нижнего уровня по убыванию их релевантности этому документу. Предложен метод иерархической классификации на основе этого оператора. Введен критерий качества AUCH оператора релевантности. Предложена вероятностная модель иерархической классификации, построена вероятностная модель коллекции, разработан способ оценки вероятности принадлежности документа кластеру нижнего уровня с помощью иерархической функции сходства. Предложен алгоритм оптимизации параметров и гиперпараметров вероятностной модели. Для случая, когда на параметры модели накладываются априорные распределения, получена аппроксимация апостериорного распределения параметров, а также совместного апостериорного распределения параметров и классов неразмеченных документов. Получены аналитические оценки вероятности принадлежности неразмеченных документов кластерам нижнего уровня экспертной иерархической структуры. Предложен способ учета синонимичности слов с помощью векторных представлений слов.
4. Рассмотрена задача верификации экспертной тематической модели. Введен функционал качества экспертной модели. Предложен неметрический алгоритм построения иерархической тематической модели, схожей с экспертной, изменяющий класс документа в экспертной модели, если при этом прирост качества больше заданного штрафа за такое изменение. Предложен критерий выбора штрафов за различные виды переносов.
5. Проведен анализ свойств предложенных методов. Описан реализованный программный комплекс, классифицирующий аннотации докладов крупной конференции EURO с помощью экспертных тематических моделей прошедших конференций. Построена экспертная система, позволяющая классифицировать веб-сайты компаний индустриального сектора. Проведено сравнение предложенных алгоритмов с известными решениями. Предложенные алгоритмы показали более высокие результаты. Для визуализации результатов верификации экспертной модели предложен метод вложенной визуализации иерархической модели на плоскости.

Публикации соискателя по теме диссертации

Публикации в журналах из списка ВАК.

1. Кузьмин А. А., Адуенко А. А., Стрижов В. В. Выбор признаков и оптимизация метрики при кластеризации коллекции документов // Известия ТулГУ, 3 (2012). С. 119-131.
2. Кузьмин А. А., Стрижов В. В. Проверка адекватности тематических моделей коллекции документов. // Программная инженерия, 4 (2013). С. 16-20.
3. Кузьмин А. А., Адуенко А. А., Стрижов В. В. Тематическая классификация тезисов крупной конференции с использованием экспертной модели // Информационные технологии, 6 (2014). С. 22-26.
4. Златов А. С., Кузьмин А. А. Построение иерархической тематической модели крупной конференции // Искусственный интеллект и принятие решений, 3 (2016). С. 77-86.

Остальные публикации.

5. Кузьмин А. А. Многоуровневая классификация при обнаружении движения цен // Машинное обучение и анализ данных, 3 (2012). С. 318-327.
6. Kuzmin A. A., Aduenko A. A., Strijov V. V. Hierarchical thematic model visualizing algorithm // 26th European Conference on Operational Research, Rome, (2013). P. 155.
7. Kuzmin A. A., Aduenko A. A., Strijov V. V. Thematic Classification for EURO/IFORS Conference Using Expert Model // 20th Conference of the International Federation of Operational Research Societies, Barcelona, (2014). P. 173.
8. Кузьмин А. А., Стрижов В. В. Построение иерархических тематических моделей крупных конференций // Математические методы распознавания образов ММРО-17. Тезисы докладов 17-й Всероссийской конференции с международным участием, г. Светлогорск: Торус пресс., (2015). С. 224–225.
9. Кузьмин А. А., Адуенко А. А. Построение иерархических тематических моделей крупных конференций // Сборник тезисов 23 международной научной конференции студентов, аспирантов и молодых ученных “Ломоносов-2016” секция “Вычислительная математика и кибернетика”, г. Москва: МАКС Пресс., (2016). С. 73–75.
10. Kuzmin A. A., Aduenko A. A., Strijov V. V. Thematic Classification for EURO/IFORS Conference Using Expert Model // 28th European Conference on Operational Research, Poznan, (2016). P. 206.