

«УТВЕРЖДАЮ»

Проректор – начальник Управления научной
политики и организации научных
исследований МГУ имени М.В. Ломоносова,
доктор физико-математических наук,
профессор



Федянин Андрей Анатольевич
«16» марта 2017 г.

ОТЗЫВ ВЕДУЩЕЙ ОРГАНИЗАЦИИ

Федерального государственного бюджетного образовательного
учреждения высшего образования
«Московский государственный университет имени М.В. Ломоносова»
Факультета вычислительной математики и кибернетики
кафедры математических методов прогнозирования
на диссертационную работу
Кузьмина Арсения Александровича
«Иерархическая классификация коллекций документов»,
представленную на соискание ученой степени
кандидата физико-математических наук по специальности
05.13.17 – «Теоретические основы информатики»

Актуальность темы

Задача иерархической классификации и кластеризации коллекций документов является одной из центральных в области машинного обучения и анализа данных, а **ее актуальность не подлежит сомнению**. Представление коллекции в виде иерархической структуры упрощает анализ ее тематики, систематизацию и поиск в ней нужной информации. В работе рассматриваются частично размеченные коллекции, в которых для части документов уже задано положение в иерархической структуре тем. Данная информация позволяет оптимизировать параметры моделей классификации новых и еще неразмеченных документов. Задачи автоматической иерархической классификации текстов решаются в мире достаточно активно. Ранее были предложены как общие методы (Joachims: 1998, Nao: 2007), рассматривающие документы как вектора действительного пространства, так и методы, использующие вероятностные предположения о процессе генерации текстов (Blei: 2003, Zavitsanos: 2011).

Работа А.А. Кузьмина развивает первый тип подходов, в частности, методы оценки важности признаков по размеченным документам, весов кластеров различных уровней при иерархической классификации текстов и способы сравнения документов друг с другом. В диссертационной работе предложенные ранее методы классификации обобщаются на случай иерархических коллекций с помощью

взвешенной иерархической функции сходства, позволяющей учитывать сходство документа одновременно со всеми кластерами ветки дерева иерархии. На базе предложенной функции сходства строится вероятностная модель совместного распределения параметров данной функции и классов документов коллекции. Разрабатывается итеративный алгоритм поиска оптимальных параметров данной модели с помощью методов вариационного вывода. **Таким образом, тема данной диссертационной работы является актуальной.**

Содержание

Диссертационная работа состоит из введения, пяти разделов и заключения.

Во введении обоснована актуальность работы, установлены цели и задачи, перечислены основные выносимые на защиту положения, приведен краткий обзор методов исследования, установлена научная новизна, теоретическая и практическая значимость, обоснована достоверность результатов, перечислены основные публикации по теме диссертации.

В первой главе приведен полный обзор предметной области, поставлены задачи иерархической классификации и кластеризации, описаны известные методы их решения.

Вторая глава посвящена оптимизации весов взвешенной метрики Минковского по размеченной части текстовой коллекции. Построенная метрика используется для плоской и иерархической кластеризации тезисов конференции, анализируются различные способы векторного представления документов.

В третьей главе рассматривается задача ранжирования кластеров по релевантности новому документу, вводится понятие оператора релевантности, возвращающего ранжированную перестановку кластеров, и определяется его качество. Для вычисления сходства документа взвешенная косинусная функция сходства обобщается на иерархических случаях. Для определения весов иерархической взвешенной функции сходства предлагается иерархическая энтропийная модель важности слов. Для классификации неразмеченных документов предлагается вероятностная модель коллекции. На ее базе строится аппроксимация совместного апостериорного распределения классов неразмеченных документов и параметров модели, с помощью которой прогнозируются классы новых документов.

В четвертой главе рассматривается задача верификации тематической модели. Задается критерий качества иерархической модели и система штрафов за ее изменение. Предлагается жадный алгоритм оптимизации данного функционала. Свойства предложенного алгоритма анализируются при решении задачи верификации иерархической модели тезисов конференции.

В пятой главе описывается реализованный программный комплекс для классификации тезисов предстоящей конференции и веб-сайтов компаний на базе предложенных методов. Полученные результаты сравниваются с результатами существующих методов иерархической и плоской классификации. Предлагается способ вложенной иерархической визуализации тематической модели.

Основные результаты и их новизна

В рамках диссертационной работы А.А. Кузьмина получены следующие новые результаты:

1. Предложен алгоритм оптимизации весов взвешенной метрики Минковского с помощью частично размеченной коллекции.

2. Предложена иерархическая взвешенная функция сходства ветки дерева иерархии и документа. Предложена иерархическая энтропийная модель важности слов для оценки весов функции сходства.
3. Введен оператор релевантности, задан критерий его качества и способ оптимизации параметров иерархической функции сходства путем максимизации качества соответствующего оператора релевантности.
4. Предложена вероятностная модель коллекции на базе иерархической функции сходства, разработан ЕМ алгоритм построения оценки совместного апостериорного распределения параметров данной функции и классов неразмеченных документов.
5. Разработан программный комплекс для классификации тезисов предстоящей конференции и веб-сайтов компаний. Проведено сравнение предложенных методов с существующими аналогами.

Достоверность результатов

Математические результаты диссертационной работы А.А. Кузьмина сформулированы в виде утверждений и теорем с корректными доказательствами. Качество предложенных методов проверено на реальных задачах построения иерархических моделей тезисов крупных конференций и веб-сайтов индустриального сектора. При сравнении с существующими альтернативными методами, предложенные методы показали более высокий результат. Основные результаты по теме диссертации изложены в десяти работах в рецензируемых научных журналах и обсуждались на всероссийских и международных конференциях. Автореферат диссертации, опубликованные статьи и тезисы докладов достаточно полно отражают основное содержание работы.

Значимость результатов

Работа А.А. Кузьмина носит преимущественно теоретический характер. Основной ее ценностью является обобщение ранее предложенных методов текстовой классификации и кластеризации на иерархический случай и теоретическое обоснование полученных методов. Эффективность полученных методов подтверждается их высокими результатами на реальных задачах, превышающими результаты известных альтернативных подходов.

Замечания

1. При сравнении с другими алгоритмами в разделе 5.1. стоило бы дополнительно привести значения других метрик качества, например, average precision at K.
2. Обзор вероятностных методов тематического моделирования в разделе 1.7., не использующихся явно в работе, можно было бы сократить.
3. На стр. 27 в формуле условной независимости опечатка – вместо $p(t)$ стоит $p(w)$.
4. Не приводится иллюстрация сравнения результатов алгоритма, описанного на стр.58 с алгоритмом аппроксимации совместного апостериорного распределения, с.74-79.

Заключительная оценка

Оценивая диссертационную работу в целом, можно констатировать, что она является законченной научно-квалификационной работой, содержащей новое решение актуальной научной задачи – иерархической классификации коллекций

документов и имеющей большое значение для развития соответствующей отрасли знаний. Все результаты, вынесенные на защиту, строго научно обоснованы.

Диссертация соответствует всем критериям, установленным Положением о присуждении ученых степеней, предъявляемым к диссертациям на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 – «Теоретические основы информатики», а ее автор – Кузьмин Арсентий Александрович – заслуживает присуждения ученой степени кандидата физико-математических наук по указанной специальности.

Настоящий отзыв обсужден и одобрен на заседании кафедры «Математических методов прогнозирования» Факультета вычислительной математики и кибернетики Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М. В. Ломоносова» (протокол №1 от 15 марта 2017 г.)

Кандидат физико-математических наук



Сергей Исаевич Гуров

Почтовый адрес: 119991, Москва, ГСП-1, Ленинские горы, д.1, стр. 52, Московский государственный университет им М.В. Ломоносова, факультет вычислительной математики и кибернетики (ВМК)

Телефон: 8-495-930-51-28

Адрес электронной почты: sgur@cs.msu.ru

Организация – место работы: Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова», Факультет вычислительной математики и кибернетики, Кафедра математических методов прогнозирования

Должность: доцент кафедры

Web-сайт организации: <http://www.cmc.msu.ru>

Доктор физико-математических наук,
профессор



Ложкин Сергей Андреевич

Почтовый адрес: 119991, Москва, ГСП-1, Ленинские горы, д.1, стр. 52, Московский государственный университет им М.В. Ломоносова, факультет вычислительной математики и кибернетики (ВМК)

Телефон: 8-495-939-30-10

Адрес электронной почты: cmc@cs.msu.ru

Организация – место работы: Федеральное государственное бюджетное образовательное учреждение высшего образования «Московский государственный университет имени М.В. Ломоносова», Факультет вычислительной математики и кибернетики

Должность: Зам. декана ВМК МГУ

Web-сайт организации: <http://www.cmc.msu.ru>