ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА

на диссертационную работу Кузьмина Арсентия Александровича «Иерархическая классификация коллекций документов», представленную на соискание учёной степени кандидата физико-математических наук по специальности 05.13.17 - «Теоретические основы информатики»

Диссертационная работа посвящена проблемам анализа коллекций текстовых документов, в частности, проблеме иерархической классификации коллекций документов с экспертно-заданной структурой тем. Такими коллекциями являются сборники научных статей и тезисов докладов, базы патентных документов, новостные сводки. Для определения тем новых неразмеченных документов привлекается большое число экспертов, что является затратным в связи с большим размером коллекций. Разрабатываемые в данной диссертации подходы позволяют для каждого нового документа определять наилучшую тему в рамках принятой иерархии, а также построить ранжированный список наиболее релевантных тем. Очевидно, что такое решение значительно упрощает задачу экспертам.

Текстовая классификация предполагает представление документов в векторном пространстве уникальных слов коллекции. В этом же пространстве задается взвешенная функция сходства, позволяющая определить степень близости между документами. Новый документ относится к наиболее близкому классу, определяемому данной функцией. Размер вектора весов при этом совпадает с размерностью пространства и числом уникальных слов в коллекции, что усложняет задачу поиска его оптимального значения. В связи с этим разработка методов оптимизации весов данной функции, а также обобщения ее для задач иерархической классификации и кластеризации, представляет научный и практический интерес. **На основании изложенного тема диссертационной работы А.А. Кузьмина является весьма актуальной.**

Диссертационная работа состоит из введения, пяти глав и заключения.

Во введении обосновывается актуальность темы диссертации, формулируются основные цели и задачи исследования.

Первая глава работы носит обзорный характер. Автор формулирует задачи тематического моделирования, иерархической классификации и кластеризации и последовательно описывает существующие подходы для их решения.

Вторая глава посвящена различным способам векторного представления документов и взвешенному расстоянию Минковского в качестве способа сравнения документов. Предлагается алгоритм оптимизации весов метрики с помощью размеченной части коллекции. С помощью полученной локально-оптимальной метрики проводится плоская и иерархическая кластеризация коллекции текстов тезисов научных докладов.

В третьей главе рассматривается иерархическое обобщение взвешенной косинусной меры близости для сравнения документов и кластеров. Для уменьшения числа весовых параметров предлагается энтропийная модель важности слов. Вводится понятие оператора релевантности, возвращающего перестановку кластеров в порядке убывания их сходства с заданным документом, задается его критерий качества. Предлагается алгоритм оптимизации параметров

иерархической функции сходства, максимизирующий качество оператора релевантности. Для уменьшения вычислительной сложности рассматривается вероятностная постановка задачи иерархической классификации с помощью иерархической функции сходства. Делаются априорные предположения о распределениях параметров и гиперпараметров и с помощью методов вариационного байесовского вывода ищется апостериорное распределение и совместное апостериорное распределение параметров и классов неразмеченных документов. С помощью найденных распределений предлагается способ оценки вероятности документа принадлежать кластеру. Предложенные подходы сравниваются между собой и с базовым методом при решении задачи иерархической классификации тезисов конференции.

В четвертой главе рассматривается задача верификации построенной модели. Вводится критерий качества модели, зависящий от усредненного внутрикластерного и межкластерного сходства документов. Предлагается жадный алгоритм оптимизации данного функционала и система штрафов за изменение классов документов, определяющая допустимый уровень изменения исходной кластерной структуры.

В пятой главе описывается разработанный программный комплекс для классификации тезисов предстоящей крупной конференции EURO и сайтов компаний индустриального сектора на основе предложенных методов. Полученные результаты сравниваются с результатами известных современных методов. Рассматривается способ вложенной визуализации иерархической структуры коллекции и результатов ее верификации.

В диссертационной работе получены следующие **основные результаты**:

1. Предложена энтропийная модель оценки весов взвешенной функции сходства.
2. Предложена иерархическая функция сходства и способ оптимизации ее параметров, максимизирующий качество оператора релевантности.
3. Предложена вероятностная модель иерархической классификации, основанная на иерархической функции сходства, разработан алгоритм оптимизации ее параметров и гиперпараметров,
4. Разработана экспертная система для классификации неразмеченных тезисов крупной конференции и сайтов компаний, проведено сравнение результатов предложенных методов с существующими аналогами.

Предложенные в работе методы являются новыми, проверяются на реальных задачах иерархической классификации и кластеризации, сравниваются с существующими признанными методами и показывают значимые результаты. Все утверждения и теоремы подтверждаются корректными математическими доказательствами и выводами. Это дает возможность считать **полученные результаты, научные положения, выводы и рекомендации достоверными и достаточно обоснованными.**

Замечания к работе

1. Библиография, приведенная в работе вполне покрывает область исследований. Однако, не указаны работы зарубежных исследователей, занимающихся иерархической классификацией текстов. В частности, я мог бы указать работы мексиканских ученых (Адольфо Гузмана и др.), отраженные во многих публикациях.
2. Не приводятся значения экспертно задаваемых параметров распределений (3.42) и (3.43), использованных в вычислительном эксперименте.
3. Отсутствует сравнение результатов, полученных с помощью прямой оптимизации AUCH алгоритмом, описанным в разделе 3.5., с результатами вероятностного алгоритма из раздела 3.7.
4. На рис. 5.8- 5.9. у некоторых кластеров отсутствует указанная метка центра.
5. Из-за применения метода главных компонент при визуализации в разделе 2.4. нумерация кластеров на рис. 2.1. а-г отличается. Из-за этого, сравнивать их между собой несколько сложно.

Указанные недостатки, однако, не снижают ценности полученных результатов.

Заключение

Диссертационная работа А.А. Кузьмина выполнена на высоком научном уровне. В ней предложены новые методы иерархической классификации, значимо улучшающие результаты существующих методов. Предложенные алгоритмы являются теоретически обоснованными, доказавшими свою применимость и прикладную значимость в экспериментах на реальных данных. Автореферат соответствует основному содержанию диссертации. Хочу особенно отметить высокую математическую культуру автора

Работа полностью отвечает требованиям, установленным Положением о порядке присуждения ученых степеней, и требованиям ВАК РФ, предъявляемым к диссертациям на соискание ученой степени кандидата физико-математических наук по специальности 05.13.17 – Теоретические основы информатики – а ее автор, А.А. Кузьмин, заслуживает присуждения ему ученой степени кандидата наук по данной специальности.

Официальный оппонент
Кандидат физико-математических наук, доцент, Федеральное государственное бюджетное образовательное учреждение высшего профессионального образования «Российская академия народного хозяйства и государственной службы при Президенте Российской Федерации», Кафедра системного анализа и информатики, доцент
Адрес: 119571, г. Москва, пр-т Вернадского, д. 82, стр. 1
Тел.: +7 (495) 933-80-04
E-mail: MAlexandrov@mail.ru

ЗАВЕРЯЮ
УЧЁНЫЙ СЕКРЕТАРЬ
РОССИЙСКОЙ АКАДЕМИИ НАРОДНОГО ХОЗЯЙСТВА И ГОСУДАРСТВЕННОЙ СЛУЖБЫ ПРИ ПРЕЗИДЕНТЕ РОССИЙСКОЙ ФЕДЕРАЦИИ
К.Э.Н. К.К. БОНДАРЕВ
27.03.2017г.

Александров Михаил Аронович