

ОТЗЫВ ОФИЦИАЛЬНОГО ОППОНЕНТА
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».

Как я понимаю, данная работа выполнена в связи с необходимостью решения совершенно практической задачи автоматизации отнесения тезисов, поданных на большую международную конференцию по многоаспектной проблематике исследования операций, к релевантным им секциям. Организаторы конференции, Европейское общество по исследованию операций, за многие годы ее проведения организовали определённую иерархию тематики в соответствии не только – и, вероятно, не столько - с логикой науки, сколько с составом и интересами реально имеющихся исследовательских групп в различных странах. В годы, если так можно выразиться, «наивного» этапа развития анализа данных – лет 30-40 назад, подобного рода задача решалась просто: определи центры самых верхних кластеров иерархии, сравни с ними рассматриваемый документ и выбери ближайший кластер. Спустившись в универсум ближайшего кластера, проделай то же, и постепенно доберёшься до уровня секций – «висячих» узлов иерархии. Именно так решается данная задача, называемая «пополнением иерархии» в монографии французского специалиста М. Жамбю, переведённой мною в свое время на русский язык (М. Жамбю, Иерархический кластер-анализ и соответствия, М, Финансы и статистика, 1988). То, что для текстов проблематика определения, вычисления и сравнения центров не совсем тривиальна, преодолевалось французской школой со свойственной ей простотой и изяществом – через так называемый «аналИз де корреспондАнс», элегантную форму сингулярного разложения матриц, основанную на свойствах суммируемости таблиц частот встречаемости. К моему удивлению, эта школа не совсем вымерла до сего времени (см. недавний выпуск программного продукта Дтм-Вик Л. Лебара и М. Пирон, <https://pdfs.semanticscholar.org>, 2013). Американское сообщество анализа данных, как это бывает, проигнорировало французские подходы и изыски, и разработало свой формат агрегированного анализа текстов, так называемые латентные размещения Дирихле, прежде всего в работах Блэя в начале нового века.

Следует отметить, что сама по себе проблема не сводится к помощи организаторам конференций. Частично размеченные коллекции документов с экспертно заданной иерархической структурой тематики в виде дерева могут встречаться в анализе аннотаций к научным и иным работам, описаний патентов, биологических таксономий, системах стандартизации промышленных изделий. **В этом плане следует признать, что тема диссертационной работы А.А. Кузьмина является актуальной и имеющей важное значение для науки и практики.**

Диссертационная работа состоит из введения, пяти глав и заключения. **Во введении** обосновывается актуальность работы, определяются цели и задачи исследования. **В первой главе** вводятся основные понятия и обозначения. Формулируются задачи текстовой классификации и кластеризации, рассматриваются существующие современные методы предобработки текстовой информации, векторного представления слов, документов и кластеров, методы построения плоских и иерархических тематических моделей и методы иерархической классификации текстовых документов. **Во второй главе** рассматривается задача метрической иерархической кластеризации. Вводится понятие важности слов при классификации и кластеризации. В качестве способа сравнения документов рассматривается взвешенная метрика Минковского и предлагается алгоритм

оптимизации ее весов. Анализируются различные способы векторного представления документов и методы иерархической кластеризации с помощью построенной локально-оптимальной метрики. **Третья глава** посвящена иерархической классификации документов частично размеченной коллекции с экспертной кластерной структурой в виде дерева. На основе взвешенной косинусной функции сходства вводится иерархическая взвешенная функция сходства документа и кластера нижнего уровня иерархии, учитывающая сходство документа сразу со всей веткой дерева кластеров. Рассматривается более общая постановка задачи классификации, в которой каждому неразмеченному документу ставится в соответствие список кластеров нижнего уровня, ранжированный по их релевантности данному документу. Для этого вводится понятие оператора релевантности и предлагается способ оценки его качества. Рассматривается вероятностная постановка задачи и делаются априорные предположения о распределении параметров и гиперпараметров иерархической взвешенной функции сходства. Выводится ЕМ алгоритм для построения оценки совместного апостериорного распределения параметров модели и классов неразмеченных документов. С помощью найденного распределения оценивается вероятность неразмеченного документа принадлежать каждому из кластеров. Результаты работы данного алгоритма сравниваются с результатами оценки максимума апостериорной вероятности. **В четвертой главе** рассматривается задача аппроксимации экспертной тематической модели, почему-то названная задачей верификации. Предлагается способ оценки качества иерархической тематической модели, как функции от значений внутрикластерного и межкластерного сходства кластеров иерархии. Предлагается алгоритм оптимизации качества экспертной модели, перераспределяющий небольшую часть документов, нарушающих тематическую целостность модели. Число переносов документов регулируется предлагаемой системой штрафов за несоответствие измененной модели экспертной. **В пятой главе** исследуются свойства предложенных ранее алгоритмов и приводятся рекомендации по их внедрению и практическому применению. Описывается разработанный программный комплекс для классификации тезисов предстоящей крупной конференции на основе информации об экспертной классификации тезисов прошлых лет, а также программный комплекс для классификации веб-сайтов компаний индустриального сектора. Результаты предложенных методов сравниваются с известными современными аналогами.

В Отделе интеллектуальных систем ВЦ РАН успешно разрабатываются модификации и обобщения идей латентного размещения, связанных с выявлением и оценкой теоретических конструктов, таких как «отдельная тема» Блэя и Ко. К этой тематике примыкает и работа А.А. Кузьмина. Им предложена модель (3.44), факторизирующая распределение векторов, представляющих документы, но не через отдельные темы, а никак не названные векторные параметры «тэта», отражающие сравнительные веса уровней иерархии. А.А. Кузьмин, фактически, разработал все основные аспекты, связанные с алгоритмической оценкой параметров этой модели, ее применения к реальным данным, дополненным в качестве бонуса элегантной визуализацией (см. гл. 5), и оценкой степени адекватности получаемых результатов. В этом плане, мне кажется, результаты А.А. Кузьмина представляют собой образцовую кандидатскую диссертацию.

Предложенные подходы являются новыми, основываются на известных проверенных алгоритмах и опираются на корректное применение математического аппарата. Научные положения, выводы и рекомендации, сформулированные в диссертации, являются достоверными и в достаточной степени обоснованными математическими доказательствами и теоремами. Теоретическое обоснование подкрепляется проверкой на реальных прикладных задачах корректно проведенными вычислительными экспериментами и сравнением с другими известными решениями.

Замечания к работе.

1. Отсутствие сколь-нибудь развернутого обсуждения многих из предлагаемых конструкций, прежде всего, вышеупомянутого векторного представления уровней иерархии «тэта». Сюда же относится введенная мимоходом независимость этого параметра от гиперпараметров его распределения (3.46), являющаяся по нашему мнению одной из важнейших в работе. Без какого-либо обсуждения значение функции при вычислениях заменяется ее верхней границей – надо сказать, довольно эффективный прием, все чаще используемый в реальных задачах оптимизации.
2. Отсутствие обоснования правомерности замены эмпирического критерия АИСН (3.17) функцией правдоподобия (3.32).
3. Некоторая неаккуратность в обращении с терминами. Например, на стр. 11 один и тот же индекс используется и для обозначения числа слов в документе и числа документов в коллекции; тематической коллекцией названо разбиение, а при дальнейшем уточнении выясняется, что речь идет об иерархии (см. также стр. 35, 43, 54, 60, 85).

Еще одно замечание, фактически, являющееся пожеланием для дальнейшей работы. В работе принята в качестве основной версия равных априорных весов различных уровней иерархии в векторе «тэта». На самом деле, другие определения значений данного параметра могут быть использованы в далеко идущих обобщениях рассматриваемой модели для моделирования когнитивных систем. Например, в жизни часто встречаются человеческие когнитивные системы, ориентированные, в основном, на очень конкретные понятия и действия. Бывают, напротив, и люди, ориентированные в основном на абстрактное мышление. Очевидно, первому типу соответствуют убывающие, а второму – возрастающие веса важности уровней иерархии.

Диссертация с большим запасом отвечает требованиям, установленным Положением о порядке присуждения ученых степеней. Она написана грамотно и аккуратно оформлена. Автореферат соответствует основному содержанию диссертационной работы.

Работа полностью соответствует 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.**

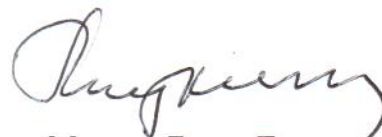
Официальный оппонент

Доктор технических наук доцент профессор департамента анализа данных и искусственного интеллекта факультета компьютерных наук Федерального государственного автономного образовательного учреждения высшего образования «Национальный исследовательский университет «Высшая школа экономики»

Адрес: 125319, Кочновский пр., д. 3, к. 329.

Тел.: +7 (499) 152-0151, +7 (963) 723-4021

E-mail: bmirkin@hse.ru



Миркин Борис Григорьевич

Подпись заверяю

Уд. Ц. Лист

Коневская Л.В.

10 марта 2014 г.

