Отзыв научного руководителя на ВКР Лии Рауфовны Халиуллиной по теме: «Вероятностное моделирование в классификации коллекции документов»

Актуальность темы ВКР следует из необходимости систематизации и последующего анализа больших быстро растущих информационных потоков. Одним из наиболее эффективных методов систематизации документов является тематическое моделирование. Вероятностная тематическая модель коллекции документов представляет каждый документ в виде дискретного вероятностного распределения. Важной особенностью такого моделирования является осуществление «нечеткой кластеризации».

Целью работы является построение вероятностной тематической модели для многозначной классификации коллекции документов при отсутствии качественной обучающей выборки. В качестве примера коллекции документов рассматривается совокупность выпускных квалификационных работ студентов.

Решение задачи сводится к поиску приближения к заданной матрице частот слов в документах в виде произведения двух неизвестных стохастических матриц. Поиск этих матриц производится методом максимизации логарифма правдоподобия коллекции.

В работе рассматриваются две основные модели: Модель PLSA (probabilistic latent semantic analysis), или вероятностный латентно-семантический анализ, и Модель LDA (latent Dirichlet allocation), или латентное размещение Дирихле, проводится аддитивная регуляризация тематических моделей (ARTM). Дивергенция Кульбака-Лейблера является одним из важнейших инструментов конструирования регуляризаторов. В работе построен алгоритм построения вероятностной тематической модели классификации с использованием предложенного подхода. Алгоритм был реализован на коллекции из 1208 документов. Каждый документ представляет собой выпускную квалификационную работу студента, написанную на русском языке. Коллекция собрана из открытых репозиториев университетов РФ и не имеет маркировок принадлежности к готовым классам. Была создана обучающая выборка путем мягкой кластеризации через вероятностно тематическую модель.

В результате проведенного исследования получены интересные научные результаты, представляющие несомненный интерес. Считаю, что ВКР заслуживает оценки отлично. Результаты ВКР можно рекомендовать к опубликованию в открытой научной печати.

Профессор, д.т.н.  В. М. Буре