УТВЕРЖДАЮ
Декан факультета ВМК
МГУ им. М.В. Ломоносова,
академик РАН
_____Соколов И.А.
"4" сентября 2020 г.

ЗАКЛЮЧЕНИЕ

Федерального государственного бюджетного образовательного учреждения высшего образования «Московский государственный университет имени М.В. Ломоносова»

Диссертация "Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией" Апишева Мурата Азаматовича выполнена на кафедре Математических методов прогнозирования факультета Вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

В период подготовки диссертации с 01.10.2017 по 01.10.2020 Апишев М.А. — очный аспирант кафедры ММП ВМК МГУ.

В 2017 году Апишев М.А. с отличием окончил магистратуру факультета Вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова.

Научный руководитель — доктор физико-математических наук, профессор Воронцов Константин Вячеславович. Основное место работы — Федеральное государственное автономное образовательное учреждение высшего образования "Московский физико-технический институт (государственный университет)", заведующий лабораторией машинного интеллекта, профессор.

По итогам обсуждения диссертации "Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией" **принято следующее заключение**:

Тема диссертации является актуальной, поскольку, с одной стороны, аддитивно регуляризованные тематические модели применяются при решении различных задач анализа текстов и их автоматической обработки, с другой стороны, на практике подобные модели часто нужно строить для больших корпусов текстов. Примерами такого рода задач являются построение интерпретируемых векторных представлений слов и документов, кластеризация и классификация текстов, выявление и отслеживание во времени трендов, тематический разведочный поиск и т.д. Качество решения указанных задач существенно зависит и от объёма обучающих данных, и от размера модели, и от адекватности подбора и настройки регуляризаторов.

По этим причинам важными направлениями исследований являются разработка более эффективных методов обучения аддитивно регуляризованных тематических моделей и стратегий комбинирования регуляризаторов для решения практических задач.

Цель диссертационной работы — разработка и реализация параллельных алгоритмов обучения аддитивно регуляризованных тематических моделей на больших данных. Разработка и исследование комбинаций регуляризаторов для задачи поиска тем специфической направленности. Адаптация реализованных алгоритмов к обработке данных, имеющих транзакционную природу.

Основные положения, выносимые на защиту:

1. Алгоритм параллельного асинхронного онлайнового обучения регуляризованных мультимодальных тематических моделей;
2. Алгоритм обучения регуляризованных мультимодальных тематических моделей с разреженным хранением параметров;
3. Модификация EM-алгоритма с ускоренным E-шагом без нормировки и стратегии её применения для оффлайнового и онлайновых алгоритмов;
4. Стратегия комбинирования регуляризаторов для выделения специфических тем по заданному словарю с приложением к анализу этно-релевантных тем в текстах социальной сети;
5. Реализация алгоритма обучения гиперграфовых тематических моделей транзакционных данных.

Все реализованные модели и использованные регуляризаторы доступны в открытой программной реализации BigARTM: https://github.com/bigartm/bigartm.

Научная новизна данной работы состоит в разработке нового параллельного асинхронного варианта EM-алгоритма для обучения аддитивно регуляризованных тематических моделей, учитывающиего разреженную структуру матриц параметров и счётчиков модели. Научной новизной обладают экспериментальные результаты применения ускоренного E-шага в оффлайновом и онлайновых EM-алгоритмах, а также результаты комбинирования различных регуляризаторов в одной модели для выделения специфических тем.

Теоретическая значимость заключается в исследовании свойств различных регуляризаторов в задаче выделения специфических тем, границ их применимости и условий, допускающих их одновременное использование.

Практическая значимость работы определяется тем, что разработанные алгоритмы позволяют обучать аддитивно регуляризованные тематические модели быстрее, чем их предшественники, что даёт возможность за меньшее время добиваться более высокого качества моделирования. Алгоритмы обучения гиперграфовых тематических моделей улучшают качество моделирования транзакционных данных.

Результаты данной работы излагались Апишевым М.А. в 2017-2019 гг. на курсе "Машинное обучение и большие данные" на факультете инноваций и высоких технологий (ФИВТ) МФТИ.

Достоверность полученных результатов обеспечивается большим объемом исследуемого материала, подробными описаниями проведённых экспериментов, допускающими их воспроизводимость, а также доступностью исходных кодов программ и обучающих данных.

Апробация работы. Результаты работы докладывались автором на конференциях:

1. 26-я международная конференция ассоциации FRUCT, Ярославль, 2020.
2. 5-я международная конференция по анализу изображений, социальных сетей и текстов (AIST), Екатеринбург, 2016.
3. 23-я международная научная конференция студентов, аспирантов и молодых учёных "Ломоносов-2016", Москва, 2016.

Список работ, опубликованных автором по теме диссертации:

1. K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Suvorova, A. Yanina, Non-bayesian additive regularization for multimodal topic modeling of large collections // Proceedings of the CIKM 2015 Workshop on Topic Models: Post-Processing and Applications, ACM, pp. 29-37, 2015.
2. K. Vorontsov, O. Frei, M. Apishev, P. Romov, M. Dudarenko, BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections // AIST'2015, Analysis of Images, Social networks and Texts. Communications in Computer and Information Science (CCIS), Springer International Publishing, pp. 370-384, 2015.
3. М. Апишев, Реализация мультимодальных регуляризованных тематических моделей в библиотеке с открытым кодом BigARTM // Сборник тезисов XXII Международной научной конференции студентов, аспирантов и молодых учёных "Ломоносов-2015" секция "Вычислительная математика и кибернетика", МАКС Пресс, с. 91-92, 2015.
4. O. Frei, M. Apishev, Parallel non-blocking deterministic algorithm for online topic modeling // AIST'2016, Analysis of Images, Social networks and Texts. Communications in Computer and Information Science (CCIS), Springer International Publishing, vol. 661, pp. 132-144, 2016.
5. M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko K. Vorontsov, Mining Ethnic Content Online with Additively Regularized Topic Models // Computación y Sistemas, vol. 20, №3, pp. 387-403, 2016.
6. M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko K. Vorontsov, Additive Regularization for Topic Modeling in Sociological Studies of User-Generated Texts // Advances in Computational Intelligence, 15th Mexican International Conference on Artificial Intelligence, MICAI 2016. Proceedings, Part I. Lecture Notes in Artificial Intelligence, vol. 10061, pp. 166-181, 2016.
7. М. Апишев, Аддитивная регуляризация тематических моделей в задаче анализа этносоциального дискурса // Сборник тезисов XXIII Международной научной конференции студентов, аспирантов и молодых учёных "Ломоносов-2016", секция "Вычислительная математика и кибернетика", МАКС Пресс, с. 117-119, 2016.
8. D. Kochedykov, M. Apishev, L. Golitsyn, K. Vorontsov, Fast and modular regularized topic modelling // Proceeding of the 21st Conference of FRUCT Association, pp. 182-193, 2017.
9. И. Жариков, М. Апишев, К. Воронцов, Гиперграфовые многомодальные вероятностные тематические модели транзакционных данных // Интеллектуализация обработки информации (ИОИ-2018): Тезисы докл., Торус Пресс, с. 148-149, 2018.
10. М. Апишев, Эффективные реализации алгоритмов тематического моделирования // Труды ИСП РАН, т. 32, №1, с. 137-152, 2020.
11. M. Apishev, K. Vorontsov, Learning Topic Models With Arbitrary Loss // Proceedings of the 26th Conference of FRUCT Association, pp. 30-37, 2020.

Личный вклад диссертанта является решающим во всех результатах, выносимых на защиту.

Содержание диссертации соответствует паспорту специальности **05.13.17 — теоретические основы информатики**.

Диссертация "Эффективная реализация алгоритмов тематического моделирования с аддитивной регуляризацией" рекомендуется к защите на соискание учёной степени кандидата технических наук по специальности **05.13.17 — теоретические основы информатики**.

Заключение принято на заседании кафедры Математических методов прогнозирования факультета Вычислительной математики и кибернетики Московского государственного университета имени М.В. Ломоносова. Присутствовало на заседании 12 человек. Результаты голосования “за” — 12 человек, “против” — нет, “воздержалось” — нет.

Протокол №_7__ от “_9_” __сентября_____2020 г.

Председатель: _____ профессор Дьяконов А.Г.

Секретарь: _____ научный сотрудник Кропотов Д.А.