Заключение по содержанию диссертации

Райгородский Андрей Михайлович
(Ф.И.О. члена диссертационного совета)

Булатов Виктор Геннадьевич
(Ф.И.О соискателя ученой степени)

диссертация «Методы оценивания качества и многокритериальной оптимизации тематических моделей в библиотеке TopicNet», представленная на соискание степени кандидата технических наук по специальности 05.13.18 — «Математическое моделирование, численные методы и комплексы программ»
(Название диссертации, ученая степень, на которую представлена диссертация, специальность)

Дата защиты 28.12.2020

Оценка соответствия диссертации требованиям Положения о присуждении ученых степеней кандидата наук, доктора наук в МФТИ (далее - Положение):

1. **Актуальность тематики диссертации:**

   В диссертации рассматривается проблема выбора оптимальных гиперпараметров в тематическом моделировании. Эта проблема является актуальной, поскольку выбор подходящих параметров существенно влияет на ряд желаемых свойств полученной тематической модели.

2. **Научная новизна выносимых на защиту результатов:**

   Выносимые на защиту результаты являются оригинальными. Их перечень приведен далее.

   В третьей главе диссертации изучается распространённый подход к оценке тематических моделей посредством анализа списков верхних слов. Автор приводит оригинальный аргумент об его несостоятельности, основанный на информации о сочетаемости слов внутри документов: из того, что верхние слова темы являются когерентными, не следует, что все остальные слова данной темы являются когерентными. Предлагается альтернативный подход к интерпретируемости, показывается его жизнеспособность.

   В четвёртой главе диссертации получены существенные продвижения в технике относительных коэффициентов регуляризации, которые существенно увеличивают их практическую применимость. Также в четвёртой главе изучаются свойства тематических моделей, построенных с учётом ограничения-равенства, выражающего требование функциональной зависимости \Theta = f(\Phi). Показано, что учёт этого требования улучшает качество моделей по ряду критериев качества и описано, как эффективно использовать такие модели на практике.

   Основной результат пятой главы состоит в разработке архитектуры и методологии библиотеки TopicNet, упрощающей многие сложные аспекты работы с тематическими моделями и узаконивающей принятые в сообществе практики (такие как журналирование экспериментов).

Шестая глава демонстрирует применение библиотеки TopicNet на практике в прикладной задаче кластеризации обращений пользователей в контакт-центр. Введённые автором относительные веса модальностей позволяют построить сложную двухуровневую иерархию, принцип построения которой переносится на другие коллекции похожей природы.

3. **Теоретическая и практическая значимость диссертационной работы:**

В целом, работа носит прикладной характер. В диссертации описано много результатов, имеющих практическую ценность. Большая их часть выложена в открытый доступ в виде модулей библиотеки TopicNet, уже используемой в ряде проектов как в России, так и за рубежом.

4. **Полнота опубликования основных результатов диссертации в рецензируемых научных изданиях в соответствии с требованиями Положения:**

По теме диссертации опубликовано 3 работы (ещё одна принята в печать). На момент написания данного отзыва проиндексированы две из них.

Результаты диссертации в полном объёме представлены в рецензируемых профильных изданиях. Диссертация прошла достаточную апробацию как по уровню и числу публикаций, так и по уровню профильных семинаров и конференций, на которых автор делал доклады.

5. **Вопросы и замечания (в соответствии с п. 4.13 Положения соискатель отвечает на сформулированные здесь вопросы и замечания на заседании по защите диссертации):**

1) Глава 5 описывает две серии экспериментов, измеряющих качество тематических моделей, построенных "с настройками по умолчанию". В первой серии сравнивается ряд моделей из библиотеки STTM, одна вариация LDA из библиотеки GenSim и "рецепт моделирования" из библиотеки TopicNet. TopicNet показывает наилучшую различность тем, а одна из моделей STTM оказывается лидером по когерентности.

Во второй серии рассматриваются некоторые модификации этого рецепта, которые сравниваются только с LDA из GenSim (но с шестью его вариантами). Не совсем понятно, почему производится сравнение именно между TopicNet и GenSim, а не между TopicNet и STTM.

Вероятно, этот выбор связан с высокими требованиями библиотеки STTM к вычислительным ресурсам и большей распространённостью библиотеки GenSim. Тем не менее, автору следовало бы более явно объяснить этот выбор.

2) Уже упомянутое сравнение стоило бы провести на большем числе текстовых (и не только текстовых) коллекций. Это представляется перспективным направлением для дальнейшей работы.

3) Рассматриваемая в работе тематическая модель с быстрой векторизацией обозначается как TARTM в тексте и на графиках. При этом уже какое-то время известна другая модель с похожим обозначением: транзакционная (гиперграфовая)

модель T-ARTM. Во избежание дальнейшей путаницы хочется порекомендовать автору изменить название данной модели в дальнейших публикациях.

6. **Общая характеристика диссертации (не включает резолютивную часть):**

Диссертационная работа Булатова В. Г. является оригинальным завершенным научным исследованием и удовлетворяет критериям Положения МФТИ о присуждении ученых степеней.

Дата _____

Подпись _____ / Райгородский Андрей Михайлович