ОТЗЫВ

члена диссертационного совета Махортова Сергея Дмитриевича на диссертацию Али Ноамана Мухаммада Абоалязида Мухаммада на тему «Аналитика больших текстовых данных», представленную на соискание ученой степени кандидата технических наук по специальности 2.3.8. Информатика и информационные процессы

Аналитика больших данных представляет одно из перспективных и быстро развивающихся направлений современных информационных технологий. При этом обработка естественного текста по-прежнему остается весьма сложной задачей, как с теоретической точки зрения, так и вычислительной. До сих пор не существует надежных стандартных методов для эффективного исследования семантики текстов, тем более представляющих собой «большие данные».

В работе рассматривается и решается несколько интересных и актуальных задач исследования больших объемов информации в глобальной сети, которые в некотором роде могут быть отнесены к области «семантический web».

В частности, весьма актуальным направлением исследований в настоящее время стал анализ тональности текстов в социальных сетях, информационных каналах, электронных рынках. Подготовка и преобразование необработанных, неструктурированных текстовых данных в формат, доступный для применения структурного анализа, представляет одну из наиболее теоретически сложных и вычислительно нагруженных стадий при решении задач анализа данных. В представленной работе вводится модель предварительной обработки текстовых данных на основе комбинации методов обработки естественного языка. Предложенная автором методика направлена на повышение качества получаемых данных, с сохранением специфики исходного текста.

При изучении поведения пользователей в сети интересную задачу представляет также гендерная идентификация. С целью ее решения в работе предлагается динамическая модель N-грамм для обработки имен пользователей. Она опирается на доступность данных, содержащих отзывы пользователей на веб-сайтах, и извлекает их имена. Задача определения пола клиента по имени его профиля содержит ряд проблем, обозначенных и анализируемых автором диссертации. Предлагаемая в работе модель направлена на решение выделенных проблем. Она включает несколько взаимодействующих подзадач, таких как сегментация, числовая подстановка, нечеткое сопоставление.

При поиске информации в Интернете пользователи сети сталкиваются с большими объемами данных, обработать которые визуально невозможно. Такая ситуация известна как «проблема информационной перегрузки». В ее преодолении могут помочь так называемые рекомендательные системы. В диссертации представлена система рекомендации товаров, названная "SmartTips". Используемая автором методика основана на анализе тональности текстов с использованием отзывов потребителей и специальной модели Извлечения Аспектных Терминов. В результате производятся оценки различных продуктов и извлекаются предпочтения пользователей. Рассмотрен и учитывается ряд

влияющих факторов, включая голоса читателей, частоту терминов аспекта (то есть некоторой характеристики продукта), частоту слов в отзыве.

Подводя итог, можно сказать, что в работе предложены методы и алгоритмы, способствующие извлечению как можно большего объема смысловой информации в глобальной сети с учетом различных факторов, влияющих на качество этой информации. На основе предложенных моделей разработано программное обеспечение, экспериментально демонстрирующее их практическую значимость и эффективность.

Результаты исследований опубликованы в авторитетных международных сборниках, включая издания с поддержкой IEEE.

К недостаткам работы можно отнести отсутствие теоретических (например, вероятностных) обоснований качества и эффективности моделей, предложенных в диссертации, в том числе в сравнении с результатами других исследователей в области аналитики текстовых данных. Дело в том, что под «моделями» здесь подразумеваются не строго математические модели, поэтому трудно было бы получить метрики для таких обоснований. Также следует отметить отсутствие оценок вычислительной сложности представленных алгоритмов, что было бы особенно желательно при обработке больших данных.

Однако указанные недостатки во многом являются следствием трудности поставленных и решаемых в работе задач. И достигнутые положительные результаты значительно перевешивают, подтверждая актуальность работы, ее значимость и сохраняя итоговую положительную оценку.

Диссертация Али Ноамана Мухаммада Абоалязида Мухаммада на тему: «Аналитика больших текстовых данных» соответствует основным требованиям, установленным Приказом от 19.11.2021 № 11181/1 «О порядке присуждения ученых степеней в Санкт-Петербургском государственном университете», соискатель Али Ноаман Мухаммад Абоалязид Мухаммад заслуживает присуждения ученой степени кандидата технических наук по специальности 2.3.8. Информатика и информационные процессы. Нарушения пунктов 9 и 11 указанного Порядка в диссертации не обнаружены.

Член диссертационного совета

Доктор физ.-мат. наук, доцент,
зав. кафедрой программирования и
информационных технологий

ФГБОУ ВО ВГУ  Махортов С.Д.

25.07.2022