УДК 004.048

УЛУЧШЕНИЕ ПОИСКА ЛОКАЛЬНЫХ СОБЫТИЙ НА ОСНОВЕ СВЕРТОЧНЫХ КВАДРОДЕРЕВЬЕВ С ПОМОЩЬЮ СЕМАНТИЧЕСКОЙ ФИЛЬТРАЦИИ НА ОСНОВЕ ДАННЫХ INSTAGRAM.

Ковальчук М.А., Филатова А.А.
Научный руководитель – к.т.н., ординарный доцент Насонов Д.А.
Санкт-Петербургский национальный исследовательский университет
информационных технологий, механики и оптики

Доклад посвящен обнаружению разномасштабных городских событий с помощью потоковых данных из социальных сетей. С помощью разработанной системы фильтрации шума на основе ансамбля моделей из BigARTM, SentenceBERT и тематического моделирования на основе BERT удалось значительно улучшить существующий алгоритм обнаружения событий, основанный на частотном обнаружении аномалий добавив в него анализ.

Введение.

В настоящее время социальные сети имеют жизненно важное значение для многих людей. Более половины населения планеты использует социальные сети для выражения эмоций, обмена мыслями и поддержания социальных отношений. Обычные пользователи публикуют информацию о своей повседневной жизни, организаторы массовых мероприятий транслируют публичный контент через официальные страницы. Эта тенденция делает социальные сети полезным источником данных для различных задач, посвященных анализу городских процессов. Такие данные позволяют создавать, например, рекомендательные системы или системы мониторинга преступности, основанные на выявлении разномасштабных событий. Более того, предыдущие исследования показали, что информация о значимых событиях, таких как ураганы, землетрясения и наводнения, появляется в социальных сетях быстрее, чем в традиционных СМИ.

Среди всего многообразия социальных сетей, Instagram и Twitter наиболее подходят для задачи обнаружения событий. Обе они чрезвычайно широко распространены и продолжают набирать популярность. Публикации могут содержать не только текстовые данные, изображения или видео, некоторые из них привязаны к определенным местам и временным меткам, что упрощает идентификацию событий. Однако данные из этих социальных сетей содержат большое количество шума: публикации с едой, одеждой, спамом или рекламой не отражают информацию о каком-либо событии и приводят к плохим результатам.

Большая часть существующих решений направлена на обнаружение крупных событий, которым посвящено большое количество публикаций с помощью данных из социальной сети Twitter, концентрация на крупных событий позволяет легко фильтровать шум, так как для детектирования события нужно находить сразу много публикаций связанных по теме и близких по локации и времени. Первым недостатком этого подхода является сам по себе Twitter, у которого лишь незначительное количество публикаций обладает геометками, что значительно усложняет локализацию событий и его низовая популярность на территории РФ. Другая проблема данного подхода заключается в том, что остаются незамеченными практически все небольшие локальные событий, например выступление уличного музыканта, открытие нового магазина, живая музыка в ресторане, и теряется часть событий районного масштаба, такие как небольшие ярмарки и фестивали, выступления малоизвестных музыкальных групп. При этом при повышении чувствительности алгоритмов поиска событии значительно снижается точность событий, так как растет количество ложных срабатываний из-за большого количества шума и рекламы в социальных сетях.

Основная часть. Для решения задачи обнаружения разномасштабных городских событий мы дополнили существующее решение, основанное на частотном поиске аномалий и использующее сверхточные квадродеревья, и данные из социальной сети Инстаграм, семантическим анализом текстов публикаций. Для повышения чувствительности алгоритма и обнаружения локальных событий мы разработали модуль фильтрации публикаций на основе ансамбля из трех моделей: BigARTM, zero shot классификация с помощью SentenceBERT, BERTopic. BERTopic является развитием идеи и тематического моделирования, но в отличие от классических алгоритмов использует текстовые трансформаторы для векторизации текстов. После этого он использует алгоритм UMAP для снижения размерности и HDBSCAN для кластеризации документов. После этого он выбирает ключевые слова для каждой темы на основе TF-IDF. BigARTM является развитием моделей на основе LDA и позволяет комбинировать регуляризаторы для создания моделей с заданными свойствами. Кроме этого, данная модель позволяет нам использовать различные дополнительные модальности, такие как временная метка или геопространственная координата. Каждая модель определяла связан ли пост с событием или нет, после чего выполнялось голосование по большинству и отфильтровывались все посты, за которые две модели проголосовали как за неинформативную публикацию для данной задачи. При этом полнота событийных публикаций снизилась незначительно, но зато удалось сильно снизить количество рекламы и шума, поступающей к алгоритму поиска событий, что позволило увеличить его чувствительность и увеличить количество обнаруживаемых событий на порядок, например для Нью-Йорка с 10757 до 177315 событий за 2019 и первые 4 месяца 2020.

Выводы. Разработанный алгоритм позволил увеличить качество обнаружения разномасштабных городских событий более чем на 40%, позволяя обнаруживать события, с которыми связанны хотя бы 3–4 публикации. Данный алгоритм планируется к внедрению в Сервис Поиска Мероприятий на платформе НЦКР.

Ковальчук М.А. (автор) Подпись

Насонов Д.А. (научный руководитель) Подпись