УДК 004.048

УЛУЧШЕНИЕ АЛГОРИТМОВ ОБНАРУЖЕНИЯ ГОРОДСКИХ СОБЫТИЙ, РАСПРЕДЕЛЕННЫХ В ПРОСТРАНСТВЕ И ВРЕМЕНИ, НА ОСНОВЕ ДАННЫХ СОЦИАЛЬНОЙ СЕТИ INSTAGRAM

Филатова А.А., Ковальчук М.А.

Научный руководитель – к.т.н, старший научный сотрудник Насонов Д.А

(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Доклад посвящен алгоритмам обнаружения событий, происходящих в городской среде, для которых характерна распределенность во времени и в пространстве. С помощью разработанных алгоритмов, основанных на ослабленных гипотезах о временной и пространственной связности и анализе семантики публикаций социальной сети Instagram на основе мультимодальных моделей BigARTM, удалось значительно улучшить качество выделения распределенных городских событий, по сравнению с существующим методом, основанным на частотном обнаружении аномалий.

Введение. Популярность социальных сетей в мире растет с каждым днем, и чем более распространенными они становятся, тем больше информации для анализа из них можно извлечь. Людей привлекает возможность оперативно делиться своим мнением и мыслями с окружающими, постоянно быть на связи со своими близкими и единомышленниками. Широкое распространение социальных сетей вместе с их мобильностью привело к тому, что они стали полноценной заменой традиционным СМИ. Это приводит к тому, что возникающие в городской среде события различной тематики и различного масштаба быстрее освещаются именно в социальных сетях, что делает их идеальным инструментом для анализа. Все больше исследователей обращают свое внимание на анализ таких данных и разработку методов как для решения задач обнаружения событий в узких предметных областях, от природных катастроф, кибербезопасности и анализа дорожного движения, так и для разработки более универсальных методов, позволяющих определять городские события различного масштаба и различной тематики. Кроме того, такое существенное преимущество социальных медиа, как скорость распространения информации и реакции пользователей на происходящие события, позволяет эффективно использовать данные социальных сетей для анализа событий в режиме реального времени.

В основном, исследователи используют для анализа социальную сеть Twitter в силу ее популярности в англоговорящих странах и относительной легкости в получении и обработке данных. Основными минусами этой социальной сети, при этом, является малый процент геометок, которые могут значительно улучшить качество выделения событий, поскольку большинство возникающих в городской среде событий имеет сильную геопространственную привязку, а также тот факт, что в русскоговорящих странах Twitter занимает далеко не лидирующее место среди используемых социальных сетей, а значит меньше подходит для комплексного анализа русскоязычных публикаций и выявления событий, которые они описывают. Социальная сеть Instagram, в то же время, занимает лидирующие позиции в рейтингах ежемесячного количества активных пользователей в России, а также содержит достаточный процент публикаций с геометками. При этом она добавляет возможность анализа фотографий, которые являются обязательным атрибутом публикаций, и сохраняет идеологию Twitter, как платформы для обмена относительно короткими публикациями, затрагивающими то, что происходит вокруг.

Важным ограничивающим аспектом большей части существующих методов выделения событий на основе анализа данных социальных сетей также является использование сильных гипотез о временной и пространственной связности публикаций, которые к ним относятся.

Такие гипотезы позволяют выделять события, которые локализованы в определенной географической точке или ограничены строгим временным диапазоном, в течение которого происходят активные упоминания такого события. Тем не менее, такие методы не позволяют обнаруживать как одно событие, те мероприятия, для которых характерно наличие нескольких равнозначных по важности и активности центров (например, запуск праздничных салютов в разных районах города) или спад активности в упоминании события на некоторый период времени с последующим возрастанием (события, которые идут на протяжении нескольких дней с характерными спадами активности в ночной период). Для решения подобных проблем были разработаны алгоритмы, основанные на ослабленных гипотезах о временной и пространственной связности, которые позволили выделять как единые события, распределенные во времени и пространстве мероприятия.

Основная часть. Для решения задачи обнаружения распределенных в пространстве и времени городских событий, а также комплексов событий, объединенных одной тематикой, мы дополнили существующее решение, основанное на частотном поиске аномалий и использующее сверточные квадродеревья и данные из социальной сети Instagram, алгоритмами связывания аномалий, основанными на ослабленной гипотезе о временном связывании и семантическом анализе текстов и хештегов публикаций. Для определения семантического сходства выделенных базовым алгоритмом аномалий мы использовали мультимодальные модели BigARTM (в качестве модальностей использовались тексты публикаций и хештеги) с настроенными регуляризаторами, обученные отдельно на англоязычных и русскоязычных публикациях. Для определения временной связности аномалий мы использовали ослабленную гипотезу временной связности, которая учитывает общий спад пользовательской активности в социальных сетях в определенные периоды времени (например, с наступлением ночи) и допускает аналогичный спад активности в упоминании события. Совместное использование анализа семантики и ослабленной гипотезы временной связности позволило выделять группы семантически схожих аномалий, описывающих реальное событие, распределенное во времени. Для того, чтобы определять геораспределенные события, мы разработали алгоритм, в рамках которого производится проверка семантической схожести и временной связности аномалий в рамках произвольного прямоугольного географического полигона, переданного на вход. Такой алгоритм позволяет обнаруживать все аномалии в рамках переданного полигона, связанные общей тематикой и локализованные во времени, а также находить центры основной активности.

Выводы. Разработанные алгоритмы обнаружения распределенных в пространстве и времени городских событий на основе анализа семантики и ослабленных гипотез временной и пространственной связности позволили улучшить точность определения городских событий на 5% для городов Нью-Йорк и Лондон и на 14% для городов Санкт-Петербург и Москва. Алгоритмы планируются ко внедрению в Сервис Поиска Мероприятий на платформе НЦКР.

Филатова А.А. (автор) Подпись

Насонов Д.А. (научный руководитель) Подпись