УДК 004.042

Ковальчук М. А., Вишератин А. А., Мухина К. Д.

Методы и подходы определения аномалий и событий разного масштаба в городской среде

Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики

Научный руководитель: к.т.н Вишератин А. А.

В современном мире социальные сети занимают важную роль в жизни человека. Люди используют их для общения, делятся в них своим мнениями и публикуют разнообразную информацию. Информация представлена различными видами, такими как: изображения, геолокация, текст, видео, аудио, - которые могут быть использованы для извлечения информации об окружающей среды. Одним из видом полезных данных, которые можно извлечь из соцсетей является информация о событиях разного масштаба происходящих в городской среде в реальном времени. Это может быть использовано как для своевременного обнаружения чрезвычайных событий, пожаров, забастовок, так и для поиска культурно развлекательных мероприятий. Различные исследования продемонстрировали, что информация о событиях через Instagram, Twitter и другие сети появляется быстрее чем в традиционных СМИ. Другим их преимуществом является возможность обнаружения не только крупных событий, но и небольших локальных, например, выступления уличных музыкантов.

Для автоматического обнаружения событий в реальном времени исследователи преимущественно используют данные из Instagram и Twitter. Они обладают необходимыми свойствами, делающими эти соцсети одними из лучших вариантов источников данных для данной задачи. Это в первую очередь высокая популярность, публичность информации и возможность прикреплять к постам геолокацию. Каждая из них обладает определенной спецификой, которую необходимо учитывать при создании алгоритмов. Для Twitter основным источником информации является текст, зачастую содержащий сокращения и орфографические ошибки, также он обладает высоким уровнем зашумленности, что усложняет нахождения релевантных постов, и всего лишь 1.5% постов с геометками. Для Instagram же основным источником информации являются метаданные, такие как: теги, геометки и время публикации - , однако задача усложняется закрытым API.

Для решения данной задачи нужен быстрый алгоритм из-за необходимости обрабатывать большие объемы данных и при этом работать в режиме онлайн. Для обнаружения события нужно найти всплески активности в соцсети и отфильтровать те, которые не относятся к локальным событиям, например увеличение активности из-за глобальных событий: выбором, выхода нового сериала; или часа пика в городе. Следующей подзадачей является связывание постов или твитов, относящихся к одному и тому же событию, определение координат и тематики события и отслеживание его последующего перемещения и изменения со временем. Данная комплексная задача привлекла внимание многих исследователей и ученых.

Для решения данной задачи зачастую применяются алгоритмы пространственно тематической кластеризации, но они как правило не достигаюты высокой точности из-за того плохо отличают глобальные всплески активности от локальных и не учитывают пространственное распределение постов и изменение активности пользователей в течении дня.

Данная проблема решается путем разбиения пространства на ячейки и использования исторических данных для предсказания количества постов в данной ячейки. При

предсказании необходимо учитывать изменение активности в соцсети как в течении дня, так и в течении года, в зависимости от месяца и сезона. Изменение активности в течении дня или даже недели можно предсказывать используя LSTM (Long Short Term Memory) нейронные сети, которые неплохо продемонстрировали себя для данной задачи, а рассмотрение активности соседних ячеек позволяет выявить глобальные события, таким образом увеличить итоговую точность обнаружения событий.

Другим способом учесть изменения активности является построение исторических сеток. Их построение на годичном интервале позволяет учесть сезонное изменение количества постов, а также изменение активности в зависимости от дня недели и часа. Квадродеревья хорошо проявили себя в этой задаче, но они обладают существенным недостатком - они не учитывают распределение постов по карте. Было продемонстрировано, что использование сверточных квадродеревьев позволило увеличить скорость и точность.

Другой полезной методологией является объединение данных с нескольких различных источников. Это могут быть не только Twitter и Instagram, но датчики заполненности парковок, концентрация такси, IoT и многое другое. Это позволяет уменьшить пороговые значения обнаружения аномалии и собрать больше информации о потенциальном событии.

Подводя итог, комбинирование данных с разных источников позволяет собрать больше информации о событии и увеличить точность обнаружения, использование исторических данных важно для учитывания сезонного и ежедневного изменения активности в соцсетях, использование сверточных квадродеревьев не только оптимизирует скорость алгоритма, но и повышает точность обнаружения аномалий.