УДК 004.048

РАЗРАБОТКА МЕТОДОВ ИЗВЛЕЧЕНИЯ КОНТЕКСТА СОБЫТИЙ, ПРОИСХОДЯЩИХ В ГОРОДСКОЙ СРЕДЕ

Филатова А.А. (Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)
Научный руководитель – к.т.н, инженер Мухина К.Д.
(Федеральное государственное автономное образовательное учреждение высшего образования «Национальный исследовательский университет ИТМО»)

Доклад посвящен разработке методов извлечения контекста событий, происходящих в городской среде, на основе данных, полученных из социальной сети Instagram. В докладе идет постановка научной проблемы автоматического извлечения контекста событий из данных социальных сетей, приводится описание существующих методов решения проблемы, представляются результаты анализа данных, делаются выводы о возможности использования ключевых характеристик набора данных для решения поставленной задачи и описывается разработанный алгоритм обнаружения событий, в основе которого лежит анализ графа зависимостей хештегов.

Введение. С появлением социальных сетей виртуальное общение стало ключевым аспектом человеческой жизни. Несмотря на широкое развитие традиционных медиа, социальные сети сейчас набирают высокую популярность. Они привлекают пользователей своей мобильностью и удобством, возможностью делиться своими мнениями и мыслями практически без каких-либо ограничений. Высокая скорость распространения информации в социальных сетях приводит к тому, что информация о многих событиях, начиная от глобальных мировых событий и заканчивая локальными, такими как пикет, пожар или наводнение в определенном районе города, появляется там гораздо раньше, чем в традиционных СМИ. Поэтому качественное выявление и классификация событий, происходящих в городской среде, по информации из социальных сетей поможет ускорить реакцию на такие события и выявить отношение к ним различных групп людей.

Основными целями исследования являются: оценка применимости информации, полученной из социальной сети Instagram для обнаружения событий, происходящих в городской среде; выявление ключевых характеристик имеющегося набора данных, которые можно использовать для разработки методов извлечения событий; построение и реализация алгоритмов, позволяющих извлекать информацию о событиях разного масштаба и значимости из большого объема нерелевантных данных.

При решении задачи извлечения контекста событий необходимо учитывать особенности данных, получаемых из социальных сетей: ограничение по длине публикуемых сообщений, большое количество нерелевантной информации, орфографических ошибок и опечаток, а также использование сленга, неформальной лексики и аббревиатур. В аналогичных работах, посвященных анализу данных социальных сетей, авторы использовали для разработки алгоритмов такие характерные особенности социальных сетей как хештеги, геолокацию и временные метки, а самыми популярными и результативными оказались алгоритмы, которые используют качественную предобработку данных, алгоритмы кластеризации и анализа графов, а также выявление резких пиков увеличения активности для решения задачи выявления событий, возникающих в режиме реального времени.

Основная часть. В основу разрабатываемого алгоритма лег анализ такой ключевой характеристики социальных сетей как хештеги. Хештеги являются ключевыми словами, которые кратко обобщают суть опубликованного сообщения, поэтому они являются хорошей базой для разработки методов обнаружения событий.

Набор данных для анализа, полученный из социальной сети Instagram для города Нью-Йорк за период с января 2018 года по май 2019 года, был представлен в формате JSON и содержал информацию о публикациях за указанный период с данными об отмеченной геолокации и временными метками. Публикации, представленные в наборе, также были сгруппированы по тематике и указанным координатам геолокации. В дальнейшем такие группы будут называться «тематические группы».

В результате первичной обработки данных из текста публикаций были выделены хештеги, после чего был произведен анализ информации о полученных хештегах и выявлены базовые статистические характеристики, такие как распределение количества хештегов в публикациях и распределение количества постов, в которых встречается определенный хештег. В исследуемом наборе данных содержалась информация о 173803 публикациях, в которых использовалось 157726 различных хештегов. В результате проведенного анализа было обнаружено, что более 80% выборки составляют публикации с количество хештегов, не превышающим 6, а большинство хештегов (66% от общего количества) являются полностью уникальными, то есть встречаются только в одной публикации. Такие хештеги, наряду с очень популярными, не могут являться хорошими характеристиками определенного события, поскольку с большой вероятностью они либо отражают видение конкретного автора, в случае уникальных хештегов, либо, в случае с очень популярными хештегами, могут использоваться авторами даже в нерелевантных постах с целью увеличения охватов публикации. В связи с этим возникла необходимость усечения множества используемых хештегов для увеличения точности результата и повышения производительности алгоритма. Для этого опытным путем были найдены границы, по которым отбирались хештеги для дальнейшей работы.

После сокращения множества используемых хештегов была построена матрица встречаемости хештегов, где для каждой пары было найдено количество общих тематических групп, в которых они появляются вместе. После этого полученная матрица использовалась как матрица смежности для построения графа встречаемости хештегов – графа, в котором вершины – это хештеги, а ребра между вершинами – это появление соответствующих хештегов вместе в одной тематической группе. Для дальнейшего анализа и кластеризации полученного графа использовалась библиотека graph-tool для языка Python.

Выводы. Текущим результатом исследования стал анализ ключевых характеристик набора данных социальной сети инстаграм, построение графа встречаемости хештегов и его последующая кластеризация на группы, из которых можно выделить те, которые соответствуют определенным событиям. В ходе дальнейшего исследования алгоритм будет совершенствоваться. В план ближайших работ входит улучшение качества работы существующего алгоритма, подбор оптимальных параметров для кластеризации полученного графа встречаемости хештегов и улучшение алгоритма за счет анализа других ключевых характеристик социальных сетей.

Филатова А.А. (автор) Подпись

Мухина К.Д. (научный руководитель) Подпись