УДК 004.042

Вишератин А. А.
(Университет ИТМО, Санкт-Петербург)
Мухина К. Д.
(Университет ИТМО, Санкт-Петербург)
Струков А. М.
(Университет ИТМО, Санкт-Петербург)
Научный руководитель –Насонов Д. А.
(Университет ИТМО, Санкт-Петербург)

Разработка интеллектуальной платформы мониторинга состояния городской среды на основе разнородных данных

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

Социальные сети играют важную роль в современном обществе - люди делятся своими мнениями, эмоциями и опытом связанными со всеми областями жизни. Информация во всем своем многообразии - текст, фото, видео, геолокация - может быть использована для предоставления пользователям полезной информации о происходящем вокруг. Широкий спектр событий различного масштаба, например, концерты под открытым небом, политические демонстрации, автокатастрофы отражаются в социальных сетях. Вот почему анализ социальных сетей для обнаружения событий набирает обороты в последнее десятилетие. Было продемонстрировано, что во многих случаях, информация о событиях может быть найдена в социальных сетях быстрее чем в СМИ.

Тем не менее, обнаружение событий на основе данных из социальных сетей является непростой задачей, из-за большого количества шума. Посты, сделанные в близких местах могут быть не связаны с каким-либо значимым событием. Извлечение информации, которая может быть полезна для идентификации действительно важных событий, требует тщательного анализа пространственно-временных особенностей сообщений, их содержания и дополнительной информации. Еще одной проблемой обнаружения событий в социальных сетях является правильное определение и анализ аномалий.

Обычно эти проблемы решаются с использованием сложных методов кластеризации, которые находят близкие посты с точки зрения пространственно-временного распределения и / или контекста. Однако способ разделения целевой области на части предопределен и не зависит от самих данных, что приводит к чрезмерному или недостаточно подробному разделению и, следовательно, к неэффективному обнаружению событий.

Для решения данной проблемы, был предложен метод обнаружения событий, при котором пространственное и временное распределение данных тщательно учитывается. Разработанный подход основан на построении нормальных состояний целевой области для всех комбинаций месяца, типа дня (рабочий день или выходные) и часа, чтобы исключить возможное временное смещение во время пространственного анализа. Нормальное состояние представлено адаптивной географической сеткой - деревом квадрантов, где каждый лист определяет часть целевой области, в которой соблюдаются определенные критерии, такие как среднее число постов или размер части. Тем не менее, стандартные quadtree не учитывают распределение данных, и поэтому была разработана улучшенная версия quadtree - сверточное quadtree (ConvTree), которое использует расположение точек в пространстве при

разделении целевой области. Использование ConvTree позволяет точно различать области высокой и низкой активности публикации и, таким образом, повысить чувствительность алгоритма обнаружения событий. Алгоритм обнаружения событий состоит из поиска аномалий, в котором текущая пост-активность в ячейке геосетки сравнивается с базовой линией ячейки, и идентификации события, которая основана на построении связанных компонентов ключевых слов в аномальной ячейке.

Для тщательной оценки разработанного подхода был собран большой набор данных, в виде постов Instagram из города Нью-Йорк с 2010 по 2018 год. Полученный набор данных содержит более 62 миллионов постов в 114 000 местоположений. Это позволило создать базисные адаптивные геосетки на целый год и исследовать эффективность данного решения в двух экспериментах. Экспериментальные результаты показывают, что данный метод достигает точности 73%, что близко или даже выше, чем у самых передовых методов в данной области.

Автор: _____/Струков А. М.

Научный руководитель: _____/Насонов Д. А.

Директор мегафакультета ТиНТ: _____/Бухановский А. В.