УДК 004.891, 004.021

РАЗРАБОТКА МЕТОДОВ ИДЕНТИФИКАЦИИ АНОМАЛЬНОГО ПОВЕДЕНИЯ ПОЛЬЗОВАТЕЛЕЙ СОЦИАЛЬНЫХ СЕТЕЙ НА ОСНОВЕ АНАЛИЗА ПРОСТРАНСТВЕННО-ВРЕМЕННОЙ АКТИВНОСТИ

Мухина К.Д.
(Университет ИТМО, Санкт-Петербург)
Вишератин А.А.
(Университет ИТМО, Санкт-Петербург)

Научный руководитель – Бухановский А.В.
(Университет ИТМО, Санкт-Петербург)

Для муниципальных органов власти так же, как и для организаций, работающих на уровне города или определенной городской области, представляет интерес решение задачи выявления аномального поведения пользователей, связанного с перемещениями людей. Для городских систем необходимы данные, максимально приближенные к реальным, о скоплениях людей и их распределению в пространстве. Достижение этого может быть обеспечено с помощью методов, позволяющих выделить аномалии в поведении пользователей, у которых данные о перемещениях не соответствуют действительности. В данной работе представлены методы идентификации аномального поведения пользователей на основе анализа их активности в социальной сети Instagram: метод определения фиктивных перемещений на основе изохрон и метод идентификации фиктивных отметок о местоположении на основе анализа хэштегов.

Под фиктивным перемещением в контексте исследования данных социальных сетей понимается такой набор последовательных постов пользователя, для которых указанная пользователем локация не является истинной хотя бы для одного поста, следовательно, данный маршрут в реальности не был проделан пользователем. Таким образом, задача по выявлению фиктивных перемещений может быть сведена к определению отметки о локации, не соответствующей действительности, для каждого отдельного поста.

В качестве данных с геолокацией были выбраны полные профили пользователей из социальной сети Instagram. Исходный набор данных социальной сети Instagram содержит посты за период с 1 января 2016 года по 1 июля 2017 года на территории Санкт-Петербурга в 6672 локациях. Пост в Instagram представляет собой фотографию, подпись к ней и метку соответствующей локации. Подпись, как и метка о локации, может присутствовать не во всех постах, поэтому для эффективного выявления информации о фиктивных перемещениях нельзя полагаться только на один тип данных.

Определение *фиктивного перемещения* основано на использовании изохрон – областей транспортной доступности за определенную единицу времени. Для расчета изохрон используется реальная дорожная сеть и данные о транспортных узлах, что позволяет снизить ошибку в определении реальных постов как фиктивных. Сервис открытых данных – Open Street Maps (OSM) - содержит данные о самих дорогах, их типе (пешеходные или автомобильные), возможности перемещения по ним (закрытые дороги, дороги с односторонним движением) и железнодорожном сообщении. Исходя из этого было решено выделить дополнительные области доступности: пешеходная доступность в области 5 км и область доступности в радиусе 500 км с использованием личного транспорта. В дополнение к этому, если посты пользователя размещены с интервалом в семь дней и более, оба поста считаются размещенными с верной локацией и перемещение между считается истинным, поскольку за такой промежуток времени пользователь мог сменить несколько видов транспорта. В качестве времени, необходимого на дальнее перемещение между двумя городами, были использованы данные о кратчайшем перелете из сервиса AirTickets. Результаты исследования показали, что пространственное

распределение для рекламного профиля с высокой долей постов с фиктивной локацией (более 52%) и обычного пользователя (доля постов с фиктивной локацией менее 20%) имеет существенные различия. В отличии от рекламных аккаунтов, чьей целью является максимальное привлечение аудитории и, как следствие, равномерное распределение своих постов по различным локациям, обычные люди склонны размещать фотографии в тех местах, где они действительно побывали, что приводит к сосредоточению постов в некоторой географической области. Большинство пользователей склонны верно отмечать свои локации (95% пользователей имеют менее 6% фиктивных отметок), та же тенденция наблюдается для распределения доли фиктивных постов в локациях. Однако для локаций можно наблюдать некоторой количество мест с очень высокой долей фиктивных постов, такое происходит из-за того, что в социальной сети Instagram не только пользователи сами указывают свое местоположение, но и используется список мест, созданных пользователями в Facebook. Это приводит к наличию дубликатов у самых популярных локаций, отличающихся названием и географическими координатами. В некоторых случаях указанные координаты могут довольно сильно отличаться от реального места, что приводит к определению таких отметок как фиктивных.

Метод идентификации фиктивных отметок о местоположении на основе анализа хэштегов основан на использовании априорных знаний о локациях в городе, в которых может отметиться пользователь. Поскольку каждая локация соответствует определенному месту, для нее существует характерный набор особенностей, который отличает ее от других мест в данной городской области. В рамках социальных сетей особенности места выражаются с помощью набора ключевых слов – хэштегов, которые активно используются пользователями для конкретного места. Хэштеги могут соответствовать как названию конкретного места, например, #невскийпроспект или #эрмитаж, так и отражать особенности природного ландшафта или перечень активностей, доступных на определенной территории, например, #море или #учеба. Таким образом для каждой локации можно построить распределение типичных хэштегов. Отклонение от такого списка используемых хэштегов в рамках отдельного поста может означать, что конкретный пользователь неверно указал свое местоположение. Локации, у которых доля фиктивных постов близка к единице, представляют собой места невысокой популярности с небольшим набором типичных тегов, что приводит к определению отметок о локациях в этих местах, как фиктивных. В более популярных местах доля постов с фиктивной отметкой не превышает 40%, а количество постов в местах более высокой долей фиктивных отметок не превышает 1000. При этом высокая доля свойственна местам, в которых часто проводятся разнообразные мероприятия, например, отелям Hampton (0.6), Crowne Plaza (0.36), ресторанам сеть Токио Сити (0.35, 0.31), Del Mar (0.38), Чайный Дом (0.40), либо местам, соответствующим улицам (Яхтенная – 0.46, Большеохтинский проспект – 0.33) или станциям метро (Маяковская – 0.58; Сенная площадь – 0.39) – такие места привлекают большое количество людей по разным причинам и в подобных местах преобладает тенденция описывать хэштегами действие, а не место. Таким образом представленные методы позволяют успешно выделить профили, которые содержат данные, не соответствующие действительности.