Правительство Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Санкт-Петербургский государственный университет»

Кафедра системного программирования

Гуликов Антон Александрович

Определение мест для размещения рекламы оператора мобильной связи

Курсовая работа

Научный руководитель:
ведущий разработчик ООО ”НМТ” Невоструев К. Н.

Санкт-Петербург
2016

Оглавление

Введение

В современном бизнесе многие компании принимают те или иные решения, опираясь на данные, накопленные за продолжительные годы существования организации. Такие решения существовали и раньше, но в последние годы наблюдается бурный рост количества накопленной информации в руках компаний, что послужило причиной развития направлений исследования программной инженерии, нацеленных на структурированные подходы к данным.

Ежедневно компании мобильных операторов получают терабайты различных данных, исследуя которые методами машинного обучения и применяя алгоритмы Big Data, операторы имеют возможность решать задачи по увеличению прибыли своей компании.

Одними из таких источников информации являются биллинговые данные. Имея их, можно предугадать, собирается ли определенный абонент отказаться от услуг оператора в ближайшее время. Кроме вопроса об оттоке абонентов, можно задаться вопросом привлечения новых клиентов компании.

Одним из эффективных способов привлечения новых пользователей является реклама. Спортивные мероприятия – главные приоритеты для мобильных операторов при размещении рекламы. Имея информацию о пользователях во время мероприятий, проводимых вблизи спортивных сооружений, можно попытаться определить популярные пути абонентов вокруг них.

В данной работе будут проведены исследование полученных данных и поиск возможных мест для расположения рекламы возле стадиона "Петровский" с целью повышения GRP[5] – маркетингового показателя, отображающего масштаб рекламного воздействия.

1. Цель и постановка задач

Целью данной курсовой работы является определение возможных мест расположения рекламы компании мобильного оператора.

Для достижения поставленной цели были выделены следующие задачи:

1. получение биллинговых данных у мобильного оператора;
2. рассмотрение задачи с точки зрения теории графов;
3. разработка методов поиска мест для рекламы;
4. визуализация полученных результатов;
5. оценка качества полученных результатов.

2. Данные

Все данные в ходе этой работы были получены в результате нескольких выгрузок IT-специалистами компании мобильного оператора. Информация была представлена файлами в формате XLSM.

По содержайщейся в файлах информации, данные разделены на два вида:

- биллинговые данные;
- данные расположения базовых станций.

2.1. Базовые станции

Каждая базовая станция ответственна за покрытие небольшого участка территории, который является частью определенной зоны местоположения. У каждой зоны есть уникальный код (LAC[6]), а у станций в зонах есть уникальные номера (Cell Id[4]). Таким образом, использовалась пара $(LAC, Cell\ Id)$ для идентификации одной станции – ее расположение на карте, адрес.

В итоге информация о базовых станциях получилась в виде списка со следующей информацией: координаты станции (широта; долгота), CID, LAC и адрес. На рисунке отображено расположение базовых станций в центре города (Рис. 1).

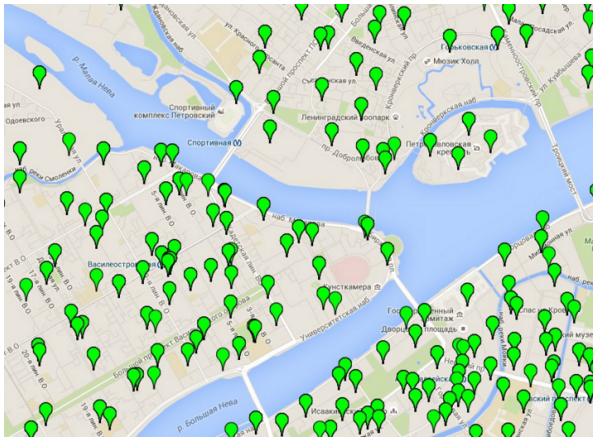

Рис. 1: Расположение базовых станций

2.2. Биллинговые данные

Биллинговые данные, которые предоставила компания, – множество таблиц, имеющих общую структуру (Таблица 1).

Время	Тип передачи информации	Cell Id	LAC	User id
10.09.02	Входящие SMS	53119	4708	8875
10.09.04	Входящие SMS	53119	4708	8875
10.24.18	Входящие SMS	33751	4708	8875

Таблица 1: Пример данных из таблицы

Главной информацией из таких данных является расположение абонента в различные моменты времени. Таким образом, получена информация о передвижении каждого абонента в течение дня (Рис. 2).

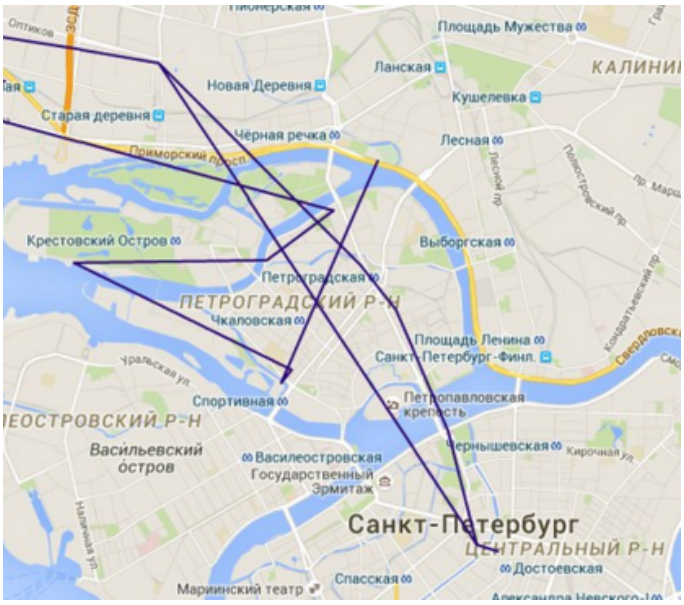

Рис. 2: Пример передвижения абонента в течение дня

3. Терминология

В данной работе постоянно использовались термины теории графов, поэтому напомним основные понятия.

Граф G – упорядоченная пара (V, E), где V – непустое множество **вершин**, а E – мультимножество пар вершин. Элементы множества E принято называть **ребрами** графа G.

В данной работе вершинами выступали координаты базовых станций, либо пара значений – (время; координаты). В качестве ребер – наличие перехода абонента от одной базовой станции к другой. Следовательно, ребра графа некоторым образом описывают движения абонентов по карте. На рисунке приведен пример графа, но вместо координат пара из LAC и $Cell\ Id$ (Рис. 3).

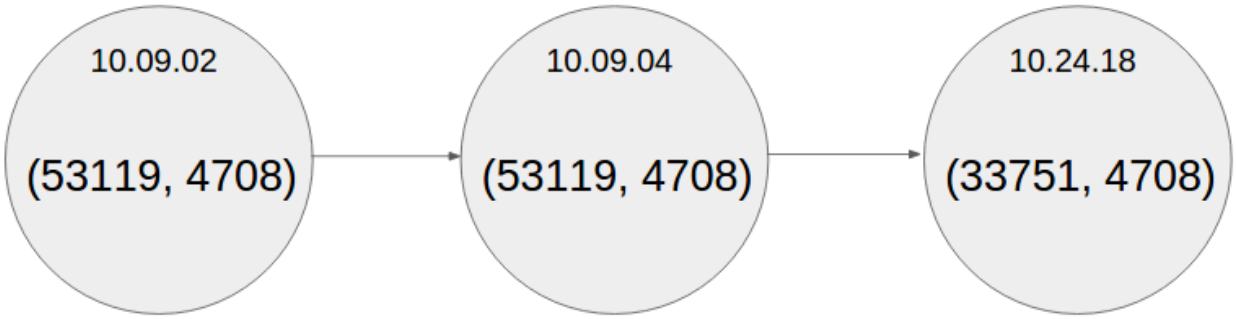

Рис. 3: Пример графа

Введем еще одно понятие, которое поможет в достижении поставленной цели. **Важное ребро** – потенциально возможный участок для размещения рекламы.

4. Инструменты

Основным языком разработки был выбран Python3 в силу его простоты и популярности. Для него написано множество удобных в использовании библиотек. Например, в работе использовались следующие библиотеки:

- openpyxl[1] – библиотека для чтения/записи Excel файлов;
- requests[3] – библиотека для удобного использования HTTP.

Времязатратные участки были реализованы на C++.

Для визуализации полученных результатов и получения дополнительной информации о маршрутах между точками на карте были использованы следующие Google API:

- Google Maps Directions API;
- Google Maps JavaScript API.

5. Реализация

На языке введенной терминологии целью данной работы является поиск важных ребер в графе. Главное предположение – важными ребрами являются наиболее популярные ребра, т.е. те участки карты, через которые проходит как можно больше людей за сутки.

5.1. Построение графов перехода

Для каждого пользователя известно только то, какая базовая станция поймала его сигнал в некоторые моменты времени. Поэтому было сделано несколько допущений:

1. Положение абонента в определенный момент времени равняется координатам базовой станции, последней принявшей его сигнал.

2. Знание координат абонента каждую секунду не требуется. Достаточно понимать, где пользователь находится через определенный фиксированный интервал времени.

3. Позиции на карте, которые расположены далеко от стадиона, малоинформативны. Они не рассматривались в данной работе.

На основании выдвинутых предположений выполнен следующий порядок действий. Во-первых, удалены из рассмотрения все базовые станции, находящиеся на расстоянии больше чем 5 км от стадиона. Во-вторых, построен не один граф, а сразу множество графов (для каждого интервала).

Теперь, когда положение пользователя в каждую секунду времени стало не столь важной информацией, значение точного времени было изменено следующим образом:

$$\textbf{новое время} = [\frac{\textbf{старое время}}{60 \cdot \textbf{промежуток}}]$$

где $\textbf{промежуток} \in \{10, 20, 30, 40, 60, 80, 100, 120, 1440\}$.

Эта формула весьма естественна. Действительно, если расположить все временные моменты на прямой и сделать засечки через одинаковые промежутки, то каждый момент времени будет находиться между какими-то двумя засечками. Тогда номер левой будет равен значению **нового времени**.

Затем рассматривался граф, вершины в котором были пары: (новое время; координаты), а ребра между парой вершин s и f существовали тогда и только тогда, когда время вершины s строго меньше времени вершины f. Если же времена всех вершин графа равны 0 (такое возможно, когда **промежуток** = 1440, т.е никак не учитывалось время), то ребро между двумя вершинами существовало, когда их координаты не совпадали.

В зависимости от длины **промежутка** изменялись размеры графа (Таблица 2).

Промежуток	Количество вершин	Количество ребер
10	20263	257981
20	11289	213463
30	7812	185332
40	6087	166784
60	4181	139595
80	3167	111420
100	2671	96592
120	2145	80919
1440	196	154173

Таблица 2: Количество вершин и ребер в полученных графах

5.2. Частотный метод

Самое главное предположение гласило, что важные ребра – ребра, через которых проходит наибольшее количество людей за сутки. На этой идее основывается частотный метод. Для каждого ребра посчитано, сколько людей прошло по нему за день. Результатом данного метода

являлся список ребер в порядке убывания количества людей, прошедших по ребру.

Главным плюсом данного метода является его простота: идейная и реализационная. Минусом же является малоинформативность полученных результатов. Ведь можно было и без запуска данного метода понять, что одним из концом важных ребер будет базовая станция, находящаяся на стадионе, а другим – станция, располагающаяся где-то поблизости от первой. Минусом является также сильная разряженность базовых станций вокруг стадиона, что не позволяет показывать качественные результаты (Рис. 4).

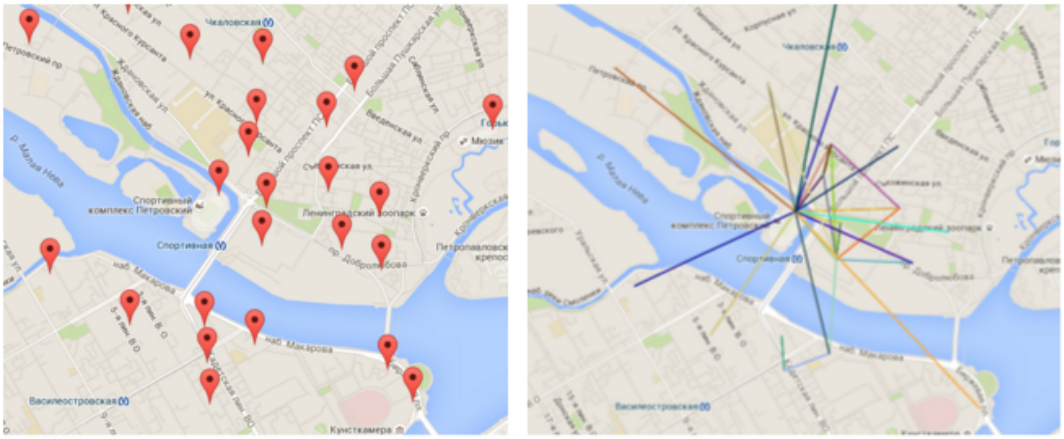

Рис. 4: Результаты частотного метода

5.3. Добавление промежуточных вершин

Для начала была предпринята попытка исправить ситуацию с разряженностью базовых станций вокруг стадиона. Учитывая прямолинейность улиц Санкт-Петербурга и обширную базу данных компании Google, выдвинуто предположение, что абоненты перемещаются между соседними на карте вершинами так как предлагает Google Maps Direction API.

Таким образом были проложены пути между соседними парами вершин на карте, и добавлены на карту новые вершины – промежуточные

точки маршрута, предложенного Google Maps (Рис. 5).

Рис. 5: Вершины графа после добавления промежуточных вершин

5.4. Метод поиска максимального потока

Данный метод является модификацией частотного метода.

Граф рассматривался, как поточная сеть с пропускной способностью ребер, равной результату частотного метода. Т.е. по ребру не мог пройти поток величиной большей, чем количество людей, которое прошло по ребру за день.

В основу метода положена аналогия между перемещением людей по карте и движением потока в сети. В соответствии с этой аналогией важными ребрами были названы наиболее насыщенные потоком ребра.

Сам метод состоял из 100-200 итераций поиска максимального потока в сети[2] со случайными истоком и стоком. Пропущенный поток суммировался по всем итерациям. Результатом являлся список ребер в порядке убывания пропущенного потока за все итерации.

В качестве результатов приведено наглядное предложение по размещению рекламы. Чем ярче (краснее) ребро или вершина, тем больше людей проходили мимо них.

Таким образом, установлено, что самыми выгодными для размещения рекламы являются мосты и самые ближайшие окрестности стади-

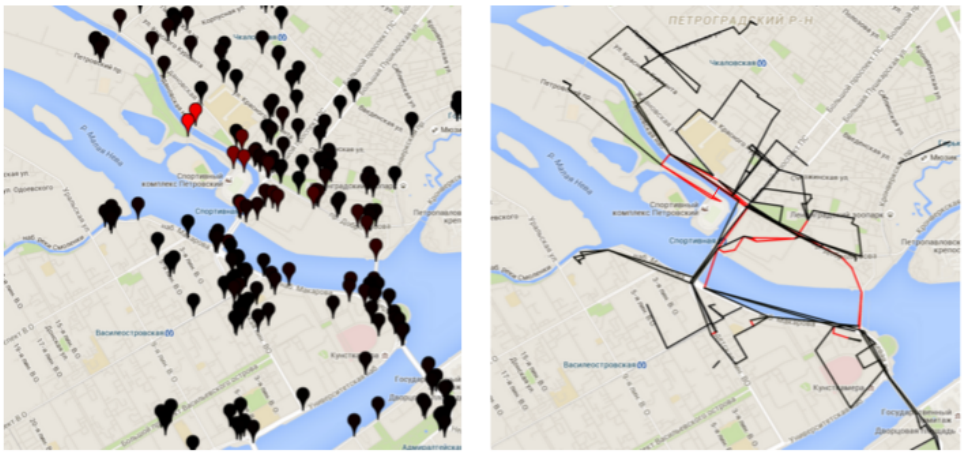

Рис. 6: Результаты метода поиска максимального потока

она (Рис. 6).

5.5. Кластеризация популярных вершин

Одним из конструктивных предложений для компании может служить предложение по размещению рекламы не только вблизи стадиона и возле мостов. Для реализации этого предложения граф был разбит на несколько подграфов, а затем в каждом из подграфов был применен метод поиска максимального потока (Рис. 7).

Оставалось понять как именно разбивать на подграфы, и сколько их должно быть. Для этого применялся алгоритм кластеризации графов MCL[7]. Основываясь на случайном блуждании, этот алгоритм предназначен для кластеризация именно графов, в отличие от многих метрических алгоритмов кластеризации.

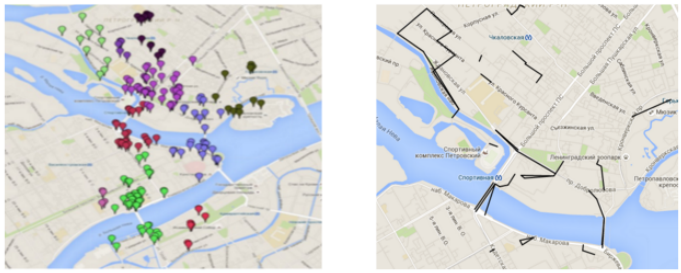

Рис. 7: Результаты работы MCL-кластеризации и последующего выделения важных ребер

6. Результаты

Оценка качества полученного результата проводилась на основе первоначальных данных (Таблица 3). В столбце со значением N отражается процент пользователей, которые прошли хотя бы по N предложенным маршрутам.

День	1	5	10	20
03.10.2015	26%	13%	6%	2%
20.10.2015	30%	15%	6.5%	2%
24.10.2015	27%	13.6%	6.3%	1.9%
31.10.2015	24%	11%	5%	1.6%
21.11.2015	22.7%	10.2%	4.9%	1.5%
24.11.2015	30.2%	15.5%	7.2%	2.3%

Таблица 3: Процент пользователей, прошедших по важным дорогам определенное количество раз

Таким образом, четверть абонентов хотя бы раз увидят новую рекламу компании.

Заключение

В ходе работы над проектом для достижения поставленной цели были реализованы следующие задачи:

- обработаны биллинговые данные сотового оператора;
- разработаны методы поиска важных ребер в графе;
- изучена работа с Google Maps API;
- реализован алгоритм MCL кластеризации графов;
- найдены потенциальные места для размещения рекламы.

Список литературы

[1] Charlie Clark Eric Gazoni. OpenPyXL // MIT/Expat. — 2016. — URL: https://openpyxl.readthedocs.io/en/default/index.html (online; accessed: 19.05.2016).

[2] James B. Orlin Ravindra K. Ahuja Thomas L. Magnanti. Network Flows: Theory, Algorithms, and Applications. Network Flows. — Prentice Hall, 1993. — Google Books : https://books.google.ru/books/about/Network$_F lows.html?id = WnZRAAAAMAAJredir_esc = y$.

[3] Reitz Kenneth. Requests: HTTP for Humans // MIT/Expat. — 2016. — URL: http://docs.python-requests.org/en/master/ (online; accessed: 19.05.2016).

[4] Wikipedia. Cell ID // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Cell_ID (online; accessed: 19.05.2016).

[5] Wikipedia. Gross rating point // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Gross_rating_point (online; accessed: 19.05.2016).

[6] Wikipedia. Location area identity // Wikipedia, the free encyclopedia. — 2016. — URL: https://en.wikipedia.org/wiki/Location_area_identity (online; accessed: 19.05.2016).

[7] van Dongen Stijn. MCL // CWI. — 2000. — URL: http://micans.org/mcl/ (online; accessed: 19.05.2016).