Санкт-Петербургский государственный университет

Матмематико-механический факультет
Кафедра системного программирования

Федоров Роман Дмитриевич

Кластеризация путей абонентов мобильного оператора в евклидовом пространстве

Курсовая работа

Научный руководитель:
Ведущий разработчик ООО "НМТ" Невоструев К.Н.

Санкт-Петербург
2016

SAINT-PETERSBURG STATE UNIVERSITY

Mathematics and mechanics faculty
Software Engineering Department

Roman Fedorov

Clustering of mobile network operator's users' routes in euclidean space

Graduation Thesis

Scientific supervisor:
Senior developer OOO "HMT" Konstantin Nevostruev

Saint-Petersburg
2016

Оглавление

Введение в предметную область, постановка задачи

Кластеризация – процесс исследования объектов и объединения их в группы (кластеры) в соответствии с какой-либо мерой расстояния между объектами. Цель кластеризации состоит в том, чтобы объекты, принадлежащие к одному и тому же кластеру, имели сравнительно маленькие расстояния друг до друга, в то время как объекты из разных кластеров имели в среднем бóльшие расстояния между собой. Особый интерес для исследования представляют задачи кластеризации в многомерных пространствах и задачи кластеризации, оперирующие большими объемами данных.

Операторы мобильной связи хранят (в том числе и из-за законодательных ограничений) множество данных о своих клиентах, например, обо всех фактах использования абонентами мобильной связи с привязкой к месту использования. Операторы связи заинтересованы в использовании этих данных для получения дополнительной прибыли. Одним из возможных вариантов использования этих данных является исследование перемещений абонентов в пространстве и времени, выявление популярных маршрутов для отдельных абонентов и для групп абонентов.

В данной работе исследованы пути абонентов и проведена кластеризация путей с помощью различных алгоритмов, использующих евклидову метрику для сравнения объектов. Евклидово расстояние для объектов p и q, имеющих размерность n, определяется по формуле:

$$dist(p, q) = \sqrt{\sum_{i=1}^{n}(p_i - q_i)^2}, \tag{1}$$

Евклидова метрика является стандартной и естественной для большого количества задач, поэтому с использованием этой метрики разработано большое количество алгоритмов кластеризации. Чтобы иметь возможность использовать многие стандартные методы и алгоритмы,

было принято решение в данной работе использовать именно эту метрику.

Для анализа решено было выбрать крупное спортивное мероприятие и провести исследование о перемещении абонентов после его окончания.

Практическая значимость исследования заключается в возможном использовании информации о популярных путях для размещения рекламы и мест досуга в тех районах, через которые проходит большое количество людей. Объединение абонентов в группы может быть использовано для создания таргетированной рекламы, поскольку можно отделить людей, живущие в разных районах города, и имеющих соответственно разное материальное положение и разные интересы.

1. Обзор литературы и существующих решений

Исследование последовательностей активностей абонентов операторов связи является актуальной темой для исследований. Так в работе [2] были исследованы закономерности временных рядов активностей абонента для предсказания направления его движения, а именно для определения следующей базовой станции сотовой связи, которую абонент посетит. В работе авторы также пытались объединять схожие пути в группы, для сравнения путей использовалась мера, сходная с мерой Жаккара, которая для путей p и q определяется по формуле:

$$sim(p, q) = \frac{|p \cap q|}{|p \cup q|}, \tag{2}$$

но учитывающая порядок следования элементов (базовых станций). В данной работе исследователи не обладали информацией о пространственном расположении базовых станций, поэтому работали лишь с точным совпадением элементов пути абонента.

Тема сравнения путей и выделения кластеров путей рассмотрена в работе [5]. В данной работе исследователи обладали информацией о расположении базовых станций. Для сравнения путей абонентов использовался метод поэлементного сравнения базовых станций с определенным временным промежутком. Для каждой активости первого абонента в заданном временном интервале искалась активность второго абонента вблизи этого же места, результаты суммировались. Для кластеризации путей использовался алгоритм QT, изначально разработанный для кластеризации геномных последовательностей. Данный алгоритм имеет временную сложность $O(n^3)$, где n – число путей для кластеризации, что не позволяет использовать его для больших объемов данных.

2. Описание предлагаемого решения

2.1. Подготовка данных

Данные были представлены мобильным оператором в виде таблиц активностей абонентов. Абоненты были обезличены, для каждого абонента был указан только некоторый идентификатор пользователя. Обработка данных производилась с помощью библиотеки pandas [3] для языка программирования Python. Таблицы активностей абонентов содержали идентификатор абонента, тип активности (входящие/исходящие SMS, входящие/исходящие телефонные звонки, пакетная передача данных и др.), время активности (с точностью до секунд) и номер базовой станции в двух различных столбцах (cellid и lacid). Сырые данные состояли из более чем 527 тысяч строк (активностей).

Данные содержали пропуски, так, например, не для всех активностей были указаны номера базовых станций. Такие данные были отфильтрованы. Данные о пространственном расположении станций находились в другой таблице, содержащей столбцы номеров базовых станций (cellid и lacid), географические координаты базовых станций (с разной точностью – от десятитысячных до десятимилионных), а также их адреса. Для таблицы активностей абонентов была сделана подстановка географических координат базовых станций вместо их номеров.

Данные в таблице содержали активности за 24 ноября 2015 года. Для анализа решено было выбрать место массового скопления людей – стадион Петровский, на котором в 20:00 начался футбольный матч, в связи с этим был выбран временной промежуток после окончания матча между 21:30 и 23:59 часами. В этом временном промежутке были выбраны абоненты, имеющие больше двух активностей. Это было сделано для того, чтобы сравниваемые пути были более осмысленными, содержали больше информации. После данной обработки выборка составила 2458 путей абонентов. Среднее количество уникальных базовых станций в пути каждого абонента составило 3.93.

Накладываемые ограничения на евклидовость пространства путей

породили идею сделать все пути равными по длине, для удобства сравнения путей и нахождения среднего. Для выравнивания был выбран временной интервал в 15 минут. Через каждые 15 минут, начиная с 21:30, в пути абонента была выбрана ближайшая по времени активность и соответствующая ей базовая станция. Таким образом было получено представление пути абонента в пространстве фиксированной размерности.

Выровненные пути были обработаны для фильтрации возможных некорректных значений, так как в ходе визуализации было обнаружено некоторое количество неестественно резких перемещений, что, скорее всего, было связано с неточностью в таблице с информацией о базовых станциях. В выборке были оставлены только пути, между 15 минутными отрезками которых пользователи перемещались не более, чем на 15 километров. Размер выборки после данного шага составил 2201 маршрут.

На следующем шаге обработки данных из таблицы были выбраны только абоненты, присутствовавшие (имевшие активность) во время матча на расстоянии 200 метров от стадиона, так как именно они и составляют главный объект исследования. Размер итоговой выборки для анализа составил 1430 маршрутов.

2.2. Оценка качества кластеризации

Анализ результатов кластеризации проводился с помощью количественных оценок:

- **Диаметра кластера**

$$diameter(cluster) = \max_{r,s \in cluster} dist(r, s) \tag{3}$$

- **Коэффициента силуэта**

$$s = \frac{b - a}{\max(a, b)}, \tag{4}$$

где: a – среднее расстояние от данного объекта до объектов того же кластера, b – среднее расстояние от данного объекта до объектов следующего ближайшего кластера.

Коэффициент силуэта для всех объектов определяется как среднее среди коэффициентов силуэта для каждого объекта. Как нетрудно заметить, данный коэффициент может принимать значения от -1 (что соответствует абсолютно неправильному определению метки кластера, т.е. $b = 0$) до 1 (этот случай соответствует идеально хорошему случаю, т.е. $a = 0$)

Также проводился визуальный анализ получившихся кластеров, для этого был разработан инструмент визуализации единичных путей и кластеров с помощью Google Maps API [1].

2.3. Кластеризация данных

Вышеописанным способом данные были подготовлены к виду, пригодному для кластеризации: путь абонента состоял из фиксированного количества базовых станций с указанием географических координат их расположения.

Были рассмотрены различные представления путей. От представления путей как последовательности географических координат было решено отказаться в силу того, что изменение координат по широте и долготе на 1 градус дает разное изменение в расстоянии. Пути были приведены к виду последовательностей относительных координат абонентов относительно стадиона Петровский. Далее был произведен перевод из географических координат в метрическую систему (километры).

2.3.1. Алгоритмы кластеризации K-Means и Mini-Batch K-Means

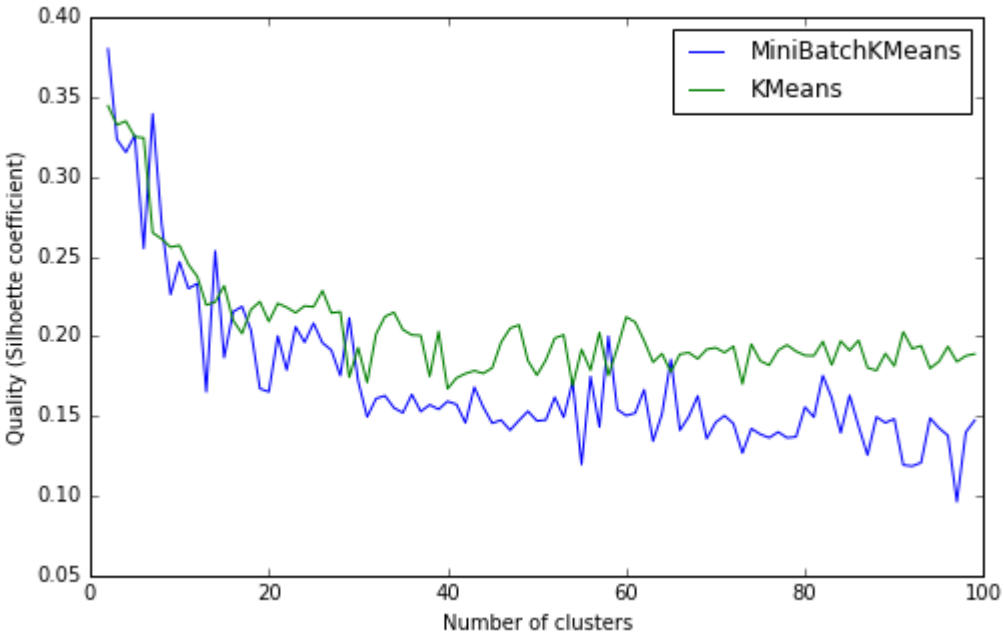
Для кластеризации данных был использован алгоритм K-Means из библиотеки scikit-learn [6], предполагающий евклидово расстояние между объектами: 1. Алгоритм инициализирует центры кластеров, добав-

ляет объекты к самому близкому центру для каждого из них и пересчитывает центры кластеров, затем следуют итерации алгоритма: объекты из одного кластера могут перейти в другой кластер, так как изменяются центры кластеров. Алгоритм имеет настраиваемые параметры: $n\_clusters$ – количество кластеров, max_iter – максимальное количество итераций алгоритма, tol – порог, при котором происходит остановка алгоритма.

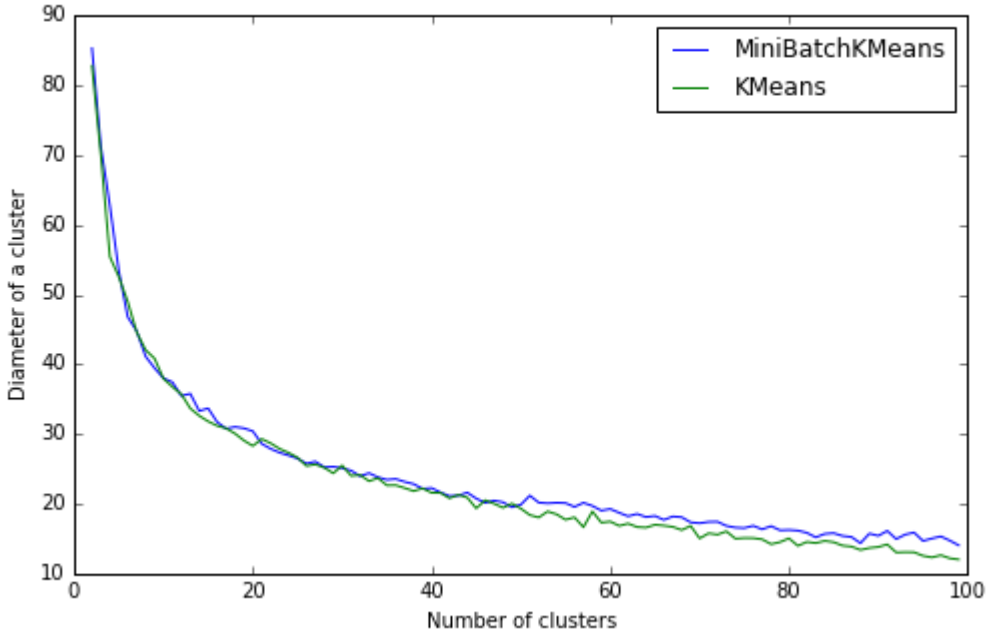
Был произведен перебор параметров, построены графики оценки качества кластеризации 1.

Для сравнения был также использован алгоритм Mini-Batch K-Means из той же библиотеки. Данный алгоритм является вариацией алгоритма K-Means, обеспечивающей более быстрое разделение объектов на кластеры.

В данной работе производительность K-Means не являлась проблемой, также этот алгоритм давал лучшие оценки качества разделения на кластеры, поэтому для дальнейшего сравнения и визуализации был выбран именно он.



(a) Зависимость коэффициента силуэта от количества кластеров



(b) Зависимость диаметров кластеров от количества кластеров

Рис. 1: Перебор количества кластеров в алгоритмах K-Means и Mini-Batch K-Means

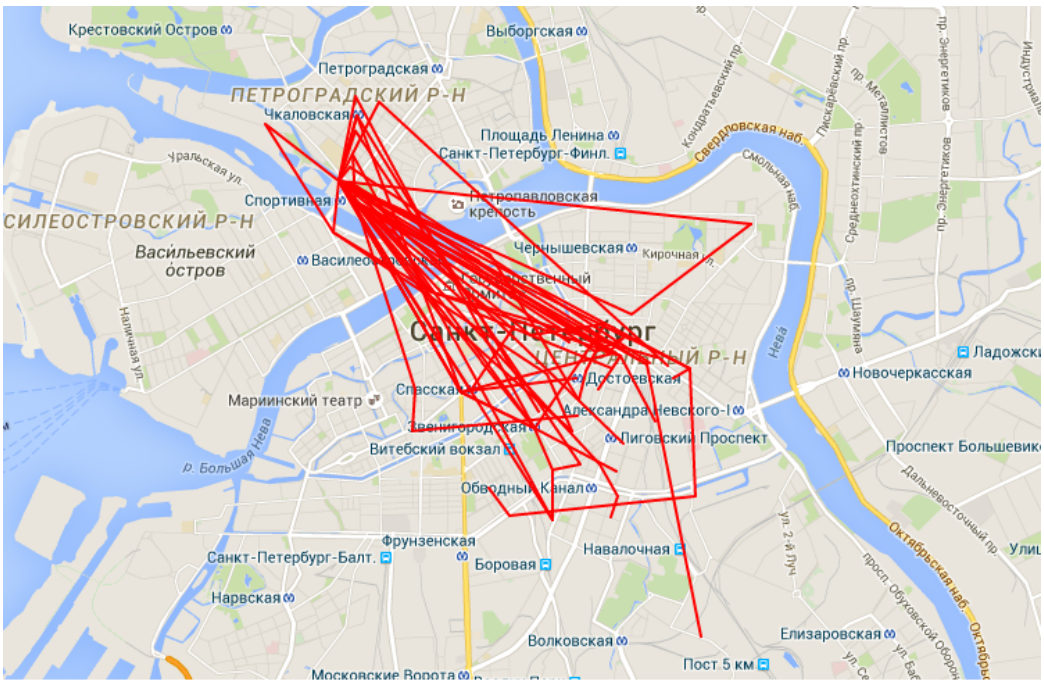
Исходя из графиков оценки качества решено было производить кластеризацию с 25 кластерами, так как это количество кластеров является

локальным минимумом для оценки диаметров кластеров 3 и локальным максимумом для коэффициента силуэта 4.

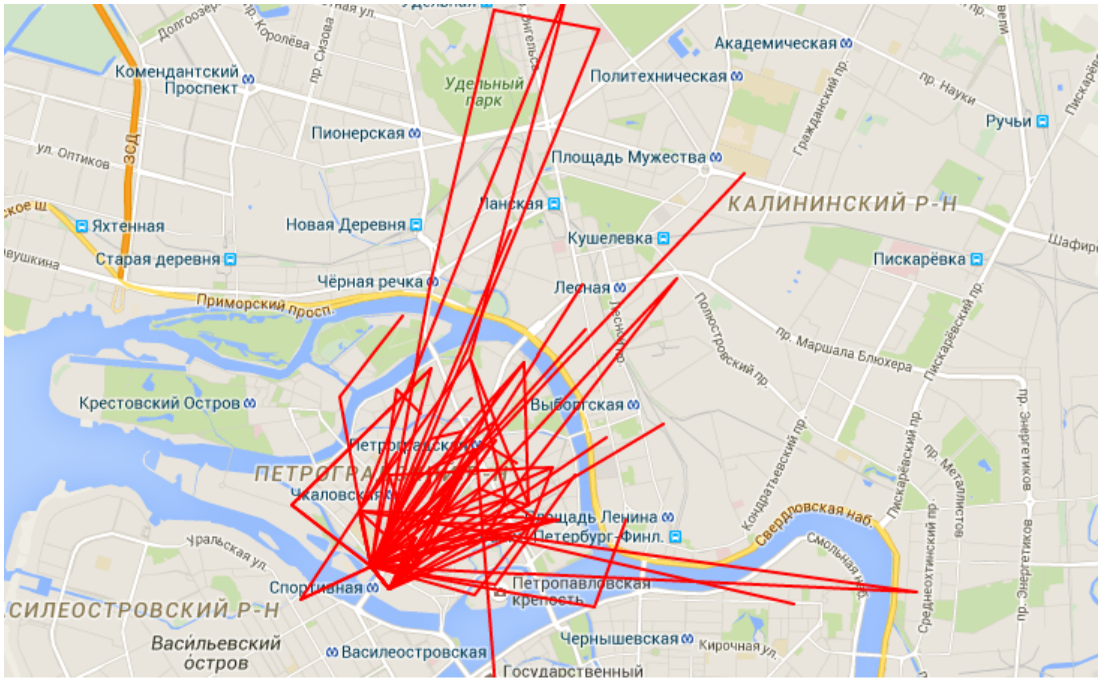
Однако дальнейший визуальный анализ разделения путей по кластерам показал, что относительно малые колебания значений коэффициентов качества кластеризации не сказываются на визуальном восприятии компактности и отделимости кластеров.

Также было выявлено, что для кластеризации с числом кластеров от 20 до 50, в среднем 15% кластеров состояли из малого количества путей (менее 0.25% от объема выборки). Такие кластеры были исключены из дальнейшего анализа.

Была проведена визуализация кластеров 2.



(a)



(b)

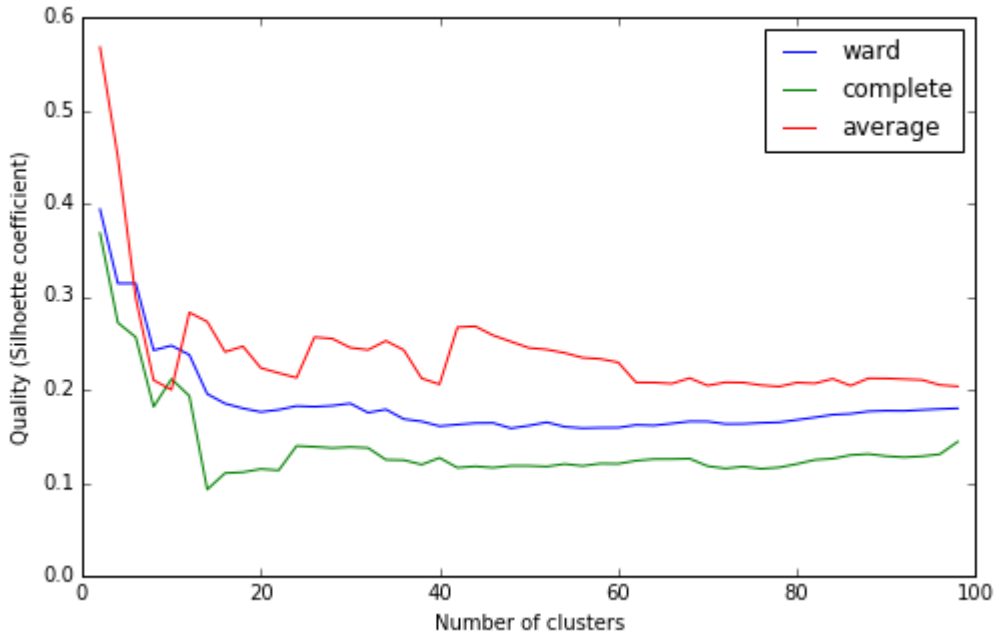
Рис. 2: Примеры кластеров в алгоритме K-Means

2.3.2. Алгоритм агломеративной кластеризации

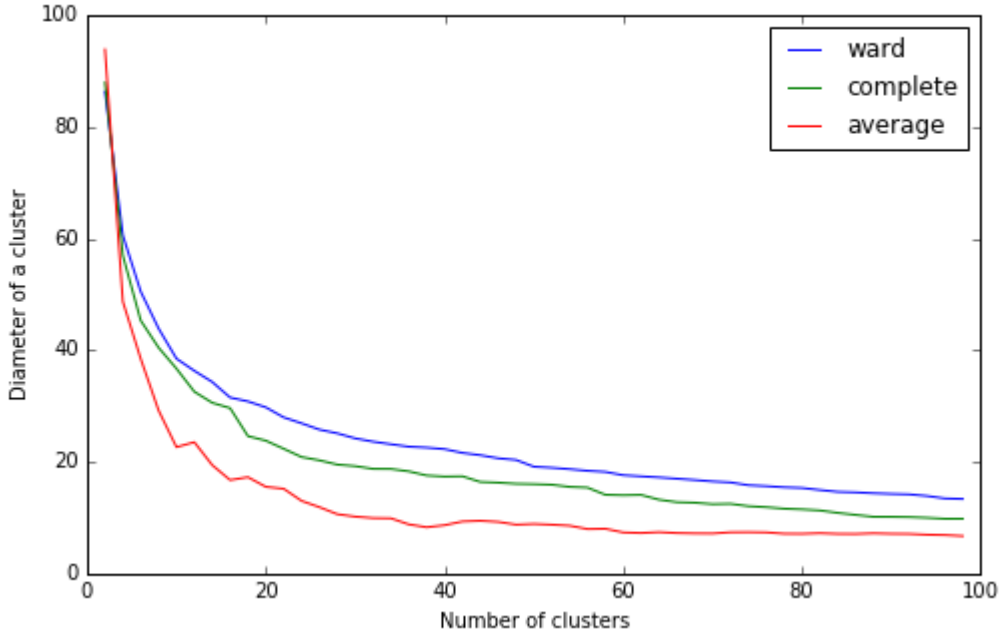
Был применен и принципиально отличающийся алгоритм кластеризации – агломеративная кластеризация. В этом методе изначально каждому объекту присваивается свой отдельный кластер, а затем на

каждом шаге самые схожие кластеры объединяются, пока не будет достигнуто требуемое количество кластеров. Была использована реализация алгоритма, предложенная в библиотеке scikit-learn [6]. Алгоритм имеет настраиваемые параметры: $n\_clusters$ – количество кластеров, $linkage$ – способ определения кластеров для слияния.

Был произведен перебор параметров, построены графики оценки качества кластеризации 3.



(a) Зависимость коэффициента силуэта от количества кластеров



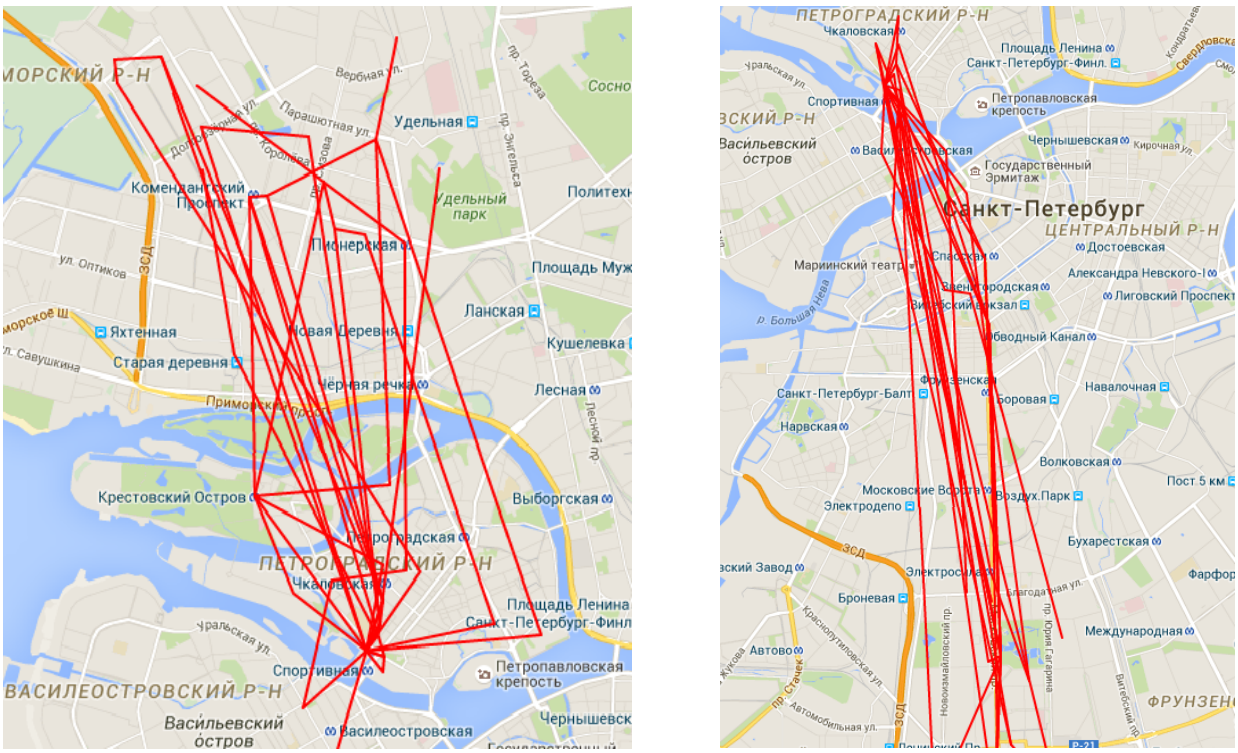
(b) Зависимость диаметров кластеров от количества кластеров

Рис. 3: Перебор количества кластеров в алгоритме агломеративной кластеризации и сравнение способов выбора кластеров для слияния

Стоит отметить, что данный алгоритм в среднем распределял пути по кластерам более равномерно. Для кластеризации с числом кластеров

от 20 до 50, в среднем 9% кластеров состояли из малого количества путей (менее 0.25% от объема выборки). Такие кластеры были исключены из дальнейшего анализа.

Была проведена визуализация кластеров 4.



(a) (b)

Рис. 4: Примеры кластеров в алгоритме агломеративной кластеризации

2.4. Определение самых популярных путей

Для определения самых популярных путей решено было уточнить маршруты пользователей, так как исходные данные предоставляли недостаточно полную информацию: для большого количества абонентов число активностей было небольшим, а, соответственно, расстояния между местами пользования мобильной связью – значительными. Для этого было использовано средство прокладки маршрутов Google Maps Directions API [1], между каждой парой действий пользователя был проложен маршрут.

Полученные данные были визуализированы с учетом разбиения абонентов по кластерам, были выделены наиболее популярные маршруты

в каждом кластере и в выборке в целом.

Рис. 5: Пример визуализации популярных путей в кластере

2.5. Оценка качества путей для эффективного размещения рекламы

Из результатов выделения наиболее популярных путей были определены наиболее популярные отрезки маршрутов, находящиеся в непосредственной близости от стадиона. Для расчета пригодности данных отрезков для размещения рекламы была выбрана оценка Daily Effective Circulation (D.E.C.) [4], определяющая среднее число людей, которые увидят новую рекламу компании. Для вычисления данной оценки был исследован другой массив данных об пользователях, включающий данные о перемещениях абонентов за 20 октября 2015 года. Данные были подготовлены аналогичным образом. Результаты вычисления D.E.C. были нормированы.

Количество отрезков	1	2	3	4
Нормированный D.E.C.	24.2%	44.5%	59.7%	69.6%

Где самым популярным маршрутом в направлении от стадиона являлся пр. Добролюбова и Биржевой мост, вторым по популярности – Тучков мост, третьим – Малый проспект Петроградской стороны, четвертым – Большой проспект Петроградской стороны.

Рис. 6: Самые популярные отрезки путей около стадиона

Заключение

В данной работе были разработаны и применены различные способы сравнения путей абонентов. Также были изучены и применены различные алгоритмы кластеризации, сделан перебор параметров алгоритмов для достижения наилучших оценок качества. Для визуализации результатов кластеризации были разработаны средства наглядного отображения путей на карте.

Пути абонентов кластеризованы, выделены различные группы абонентов, найдены наиболее популярные пути. Проведена оценка популярных путей с точки зрения пригодности их для размещения рекламы компании.

Список литературы

[1] Google. Google Maps Directions API. — URL: https://developers.google.com/maps/documentation/directions/ (online; accessed: 2016-05-23).

[2] Laasonen Kari. Clustering and Prediction of Mobile User Routes from Cellular Data // Knowledge Discovery in Databases: PKDD 2005. — 2005. — P. 569–576.

[3] NUMFocus. pandas. — URL: http://pandas.pydata.org/ (online; accessed: 2016-05-23).

[4] Outdoor Advertising Association of America Inc. Advertising glossary of terms. — URL: https://www.oaaa.org/OutofHomeAdvertising/OOHGlossaryofTerms.aspx (online; accessed: 2016-05-23).

[5] Saravanan M Pravinth Samuel V Pavan Holla. Route Detection and Mobility Based Clustering // Internet Multimedia Systems Architecture and Application (IMSAA), 2011 IEEE 5th International Conference on. — 2011.

[6] scikit-learn. — URL: http://scikit-learn.org/stable/ (online; accessed: 2016-05-23).