УДК 004.622

А. А. Пономарев

ООО Скартел
ул. Савушкина, 126Б, Санкт-Петербург, 197374, Россия

artem.ponomarev@gmail.com

ИСПОЛЬЗОВАНИЕ БОЛЬШИХ ДАННЫХ СОТОВЫМИ ОПЕРАТОРАМИ НА ПРИМЕРЕ ПОСТРОЕНИЯ МАРШРУТОВ АБОНЕНТОВ

Рассматриваются общие проблемы и вопросы использования больших данных в коммерческих организациях на примере телекоммуникационного бизнеса. Обсуждается вопрос функционала bigdata, подход к ним в зарубежной и российской практике. Автор акцентирует внимание на решении конкретной задачи в работе сотового оператора, которая может быть решена с использованием имеющихся данных в информационных системах компании. Проанализированы и построены наиболее популярные клиентские маршруты в определенные дни, которые позволяют оптимизировать рекламное размещение и повысить эффект от проводимых рекламных кампаний. Выявлены кластеры маршрутов, для которых получены наиболее успешные результаты анализа.

Ключевые слова: большие данные, bigdata, телекоммуникации, мобильная связь, маршруты, графы, оптимизация, кластеризация, рекламный охват.

Введение

Термин «большие данные» продолжает свое триумфальное шествие по научным конференциям в области технологий и инноваций, появляется все больше презентаций на тему развития bigdata, а у российских компаний, пытающихся внедрить инструментарий по обработке своих массивов данных, появляется все больший выбор поставщиков таких решений. Учитывая активное распространение мобильной связи и гаджетов (в большей степени смартфонов) среди населения страны, можно предполагать, что дальнейшая интернет-социализация населения будет только продолжаться и усиливаться. Следовательно, и количество данных о пользователях этих аппаратов и об их действиях в сети будет только накапливаться. Это позволит компаниям, которые анализируют данные о своих существующих и потенциальных клиентах или планируют этим заниматься, связывать имеющуюся у них информацию с данными социальных сетей, получая большой пласт дополнительной информации о своей клиентской базе. Как распорядиться полученным срезом и правильно его анализировать и, тем более, как правильно результаты анализа интерпретировать и принять на основе этого анализа коммерческие решения – это вторая сторона вопроса. Лежит ли она в сфере функционала больших данных – вопрос спорный. На мой взгляд, – да, поскольку я считаю, что область действия термина «большие данные» – это не только сам объем какой-либо информации и ее хранение, но и обработка этого объема, его оценка, интерпретация и соответствующие выводы из анализа и интерпретации, на основе которых должно приниматься бизнес-решение.

Различие во взглядах по отношению к функционалу больших данных хорошо видно в сравнении российского и западного подходов. Если в наших публикациях этот вопрос в принципе обсуждается, т. е. существуют определенные сомнения и расходящиеся мнения, то цель

Пономарев А. А. Использование больших данных сотовыми операторами на примере построения маршрутов абонентов // Вестн. НГУ. Серия: Информационные технологии. 2017. Т. 15, № 1. С. 70–78.

ISSN 1818-7900. Вестник НГУ. Серия: Информационные технологии. 2017. Том 15, № 1

больших данных в иностранных публикациях сводится практически везде к более широкому пониманию – анализ больших данных, в том числе методами машинного обучения [1], с целью нахождения скрытых паттернов поведения, зависимостей и инсайтов потребителей [1]. А понять инсайт клиента – это значит дать ему то, что он хочет. Это значит привлечь клиента, удержать его при меньших усилиях и затратах на дополнительные маркетинговые исследования, используя только ту информацию, которую сам клиент тебе осознанно или неосознанно уже предоставил.

Постановка задачи

Если вернуться к теме телекоммуникационного бизнеса в России и опыта применения bigdata в этой сфере, то можно упомянуть о спектре задач, которые рассматривались в первой публикации нашей рабочей группы [2]. В ней мы говорили о задачах сотового оператора, связанных с удержанием оттока существующей клиентской базы. Анализ данных строился на основе имеющейся информации о потреблении голосового трафика, трафика SMS и передачи данных обезличенных клиентов в течение нескольких месяцев. Решение данной задачи является достаточно востребованным для многих западных компаний, связанных со сферой обслуживания и работой с клиентами, поэтому неудивительно, что подобный вопрос встал и на российском сервисном рынке. Однако есть и ряд других, менее популярных задач, имеющих важное прикладное значение для бизнеса мобильных операторов. В данной статье мы попробуем проанализировать маршруты абонентов с целью выяснения и нахождения наиболее популярных из них. Нахождение подобных маршрутов может помочь спланировать размещение рекламных носителей, с тем чтобы их увидело как можно большее количество существующих и потенциальных клиентов. Некоторое подобие данной задачи было рассмотрено в работе [3], в которой были исследованы закономерности временных рядов для предсказания движения клиентов.

Сформулируем задачу следующим образом: найти наиболее популярные маршруты абонентов на пути их следования на спортивные мероприятия на стадионе «Петровский» (Санкт-Петербург) с целью оптимального размещения рекламных носителей крупных форматов для повышения индексов просмотра рекламной информации.

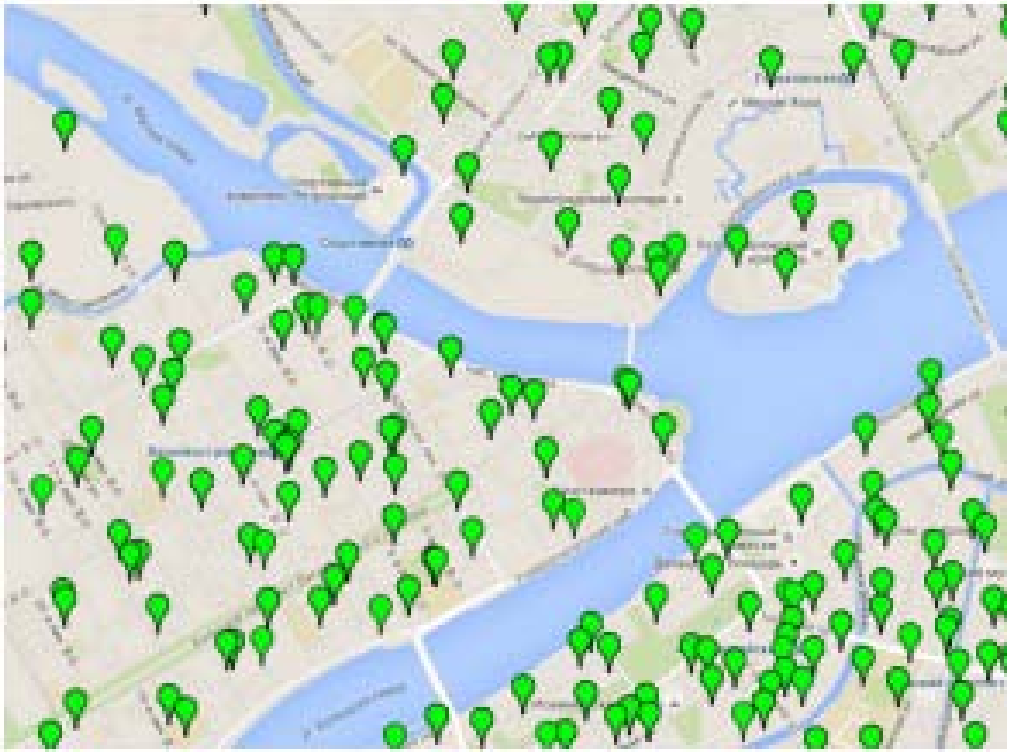
Актуальность данной задачи обусловливается следующим: стоимость аренды рекламных мест и носителей в центре городов по естественным причинам значимо превышает стоимость на его окраинах. В условиях ограничения рекламных и маркетинговых бюджетов требуется наиболее оптимальное использование подобных носителей в центре города с целью максимального охвата аудитории. Таким образом, решается задача экономии коммерческих расходов без сокращения эффективности рекламного охвата. Выбор именно спортивных мероприятий и именно данной точки города связан с тем, что на этом стадионе выступает футбольный клуб «Зенит», с которым у исследуемого оператора связи периодически проводятся совместные рекламные акции и кампании, на основе которых можно судить об их постоянных партнерских отношениях. Основываясь на этом выводе, можно говорить о том, что данный оператор видит свою целевую аудиторию, в том числе, и среди болельщиков этого клуба.

Описание решения

Для анализа и решения задачи мы использовали обезличенные данные по голосовому трафику, трафику SMS или передачи данных абонентов и информацию о местоположении

[1] См.: SAS – Analytics Software & Solutions. URL: http://www.sas.com/en_us/insights/analytics/big-data-analytics.html; а также: Simple; Learn, How Applications of Big Data Drive Industries. URL: https://www.simplilearn.com/big-data-applications-in-industries-article (дата обращения 25.01.2017).

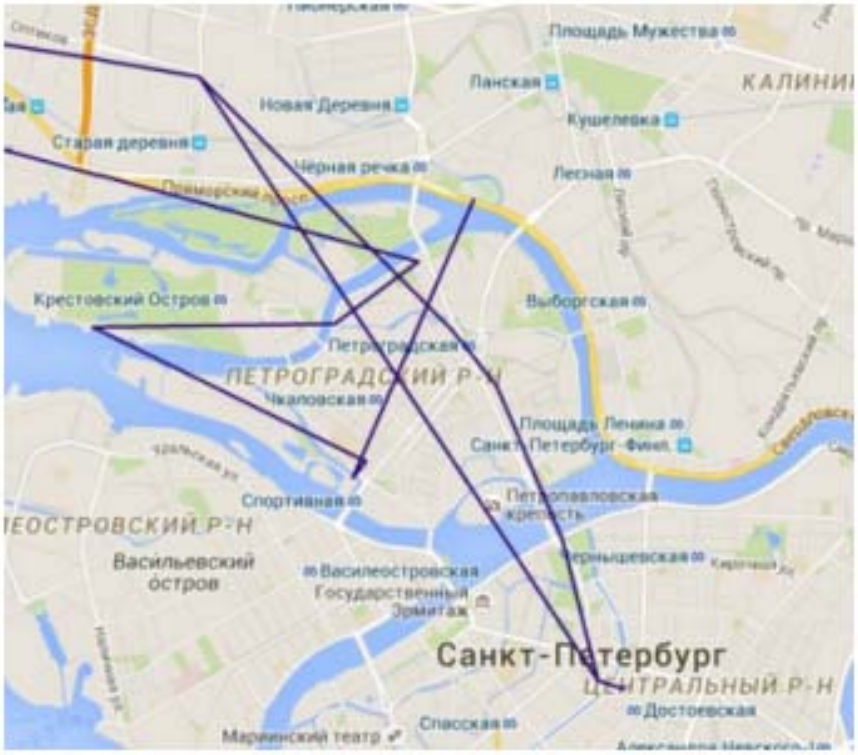
базовых станций, на которых происходила регистрация этого трафика. Данные по трафику брались за временной период, когда на стадионе проводилось спортивное мероприятие с участием футбольного клуба. Мы брали день проведения матча и часовой промежуток: за три часа до старта матча и два часа после матча. Данные по базовым станциям были представлены в виде пары (LAC, CellID), где LAC – это уникальный код городской зоны, а CellID – уникальный номер базовой станции в этой зоне. Таким образом, мы смогли идентифицировать месторасположение базовых станций на карте города:



Клиентские данные были взяты в разрезе UserID со временем регистрации любого перечисленного выше трафика на базовой станции. Поскольку пара LAC, CellID для каждой базовой станции была определена, фактически по каждому клиенту мы получили таблицу следующего вида:

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

Иначе говоря, мы получили возможность изучать расположение клиента в различные моменты времени и его передвижение исходя из временной последовательности регистрации на базовых станциях. Условное передвижение в течение дня теперь можно было наблюдать в следующем виде:



Понятно, что данные маршруты достаточно условны, поскольку вряд ли клиент передвигался исключительно по прямой. Если была поездка на метро – это один маршрут, если на автомобиле – другой. Но, тем не менее, общие очертания движения и, самое главное, вершины и ребра графа передвижения абонента были определены, направления движения от одной вершины до другой тоже. Учитывая узкий временной промежуток времени (три часа до матча и два часа после матча), допущение, что человек выходил за пределы текущего графа, при этом не отправляя сигналы на другие базовые станции, маловероятно.

Далее, учитывая имеющиеся перечисленные данные в разрезе идентификаторов каждого клиента, мы попробовали выполнить кластеризацию вершин и ребер различными методами: кластеризацию EM-алгоритмом, агломеративную кластеризацию с неевклидовыми метриками, GRGPF, алгоритмами кластеризации K-Means и Mini-BatchK-Means, а также алгоритмом кластеризации графов MCL [4]. Задачей каждого алгоритма кластеризации являлось нахождение так называемых важных ребер, т. е. участков передвижения абонентов, которые потенциально представляют наибольший интерес с коммерческой точки зрения – размещения рекламных носителей. Проще говоря, надо было определить ребра и вершины, через которые прошло (зарегистрировалось с трафиком) наибольшее количество абонентов. Поскольку для каждого пользователя (UserID) у нас было только соответствие базовой станции в момент трафиковой транзакции, нам пришлось сделать несколько допущений.

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

2. Знание координат абонента не требуется непрерывно, т. е. перемещение дискретное, и нам достаточно знать, где пользователь находится через определенный фиксированный интервал времени.

3. Базовые станции за пределами центральных районов не рассматривались, поскольку трафика в эти часы у рассматриваемых абонентов практически не было.

4. Данные с пустыми позициями (там, где не определились позиции базовой станции, где не связался UserID с трафиковой транзакцией) были приняты за погрешность выгрузки и также не учтены в анализе.

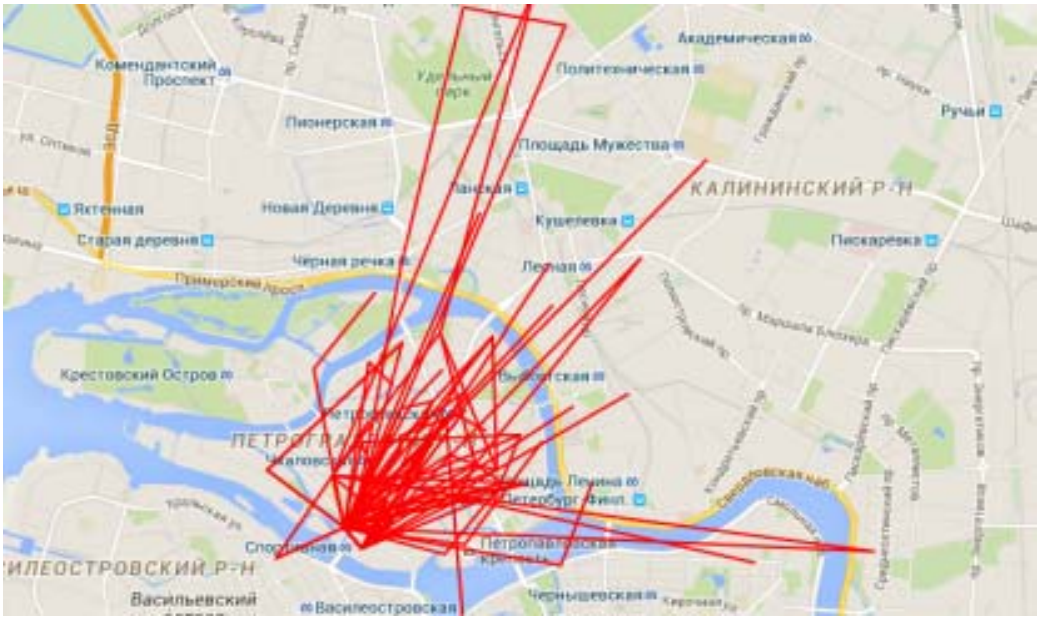
В итоге, если обратиться к результатам кластеризации графов MCL, нам удалось выделить набор важных ребер и итоги кластеризации можно представить следующим образом:

День	1	5	10	20
03.10.2015	26	13	6	2
20.10.2015	30	15	7	2
24.10.2015	27	14	6	2
31.10.2015	24	11	5	2
21.11.2015	23	10	5	2
24.11.2015	30	16	7	2

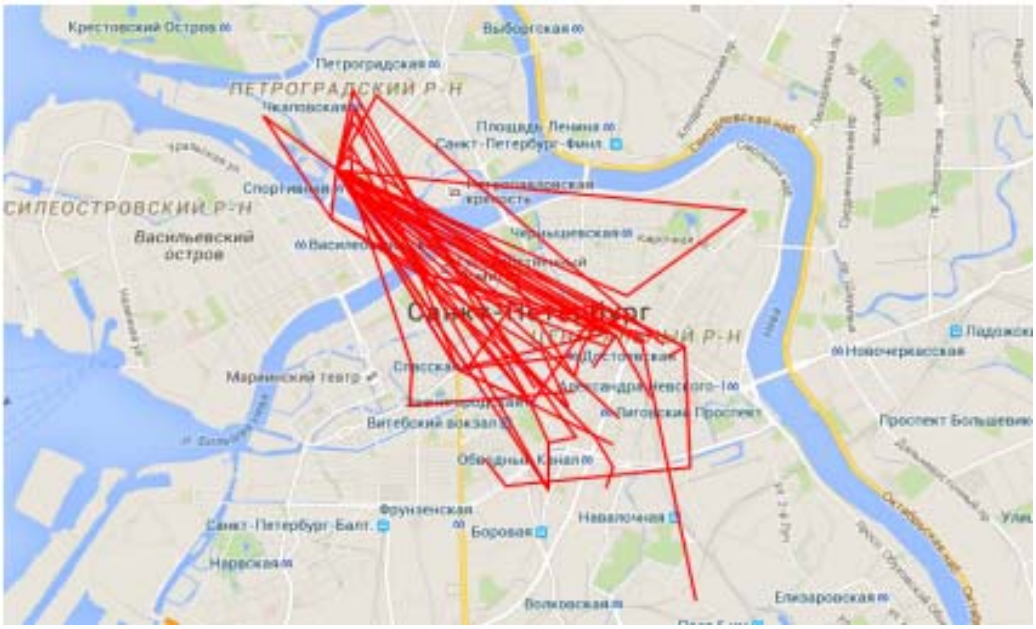
Значения 1, 5, 10 и 20 – это количество предложенных важных маршрутов, а значения в ячейках – это процент абонентов, которые прошли хотя бы по 1, 5, 10 и 20 маршрутам в указанный день. Иначе говоря, интерпретация показателя 6 % в первой строке в третьем столбце следующая: по первому маршруту шло всего 26 % абонентов, далее на каком-то перекрестке абоненты разделились и пошли часть налево своим вторым маршрутом, часть прямо своим вторым маршрутом и так далее до 10-го маршрута. Таким образом, все вместе по одному набору 10 маршрутов прошло 6 % абонентов. Сами маршруты представлены так:

Дополнительно стал интересен разброс в популярности маршрутов по дням. Изначально разброс мы связали с погодой, однако погода во все рассмотренные дни была примерно одинаковой – типично осенней и петербуржской. Затем мы связали это с днями недели, полагая, что в выходные дни большее количество болельщиков предпочитает пройти большие расстояния. Однако по факту оказалось, что 20.10 и 24.11 – дни, когда маршруты были загружены в большей степени, являлись будними днями. В эти дни проходили игры Лиги Чемпионов, соответственно, приезжали европейские клубы, что вызвало больший интерес у болельщиков и, соответственно, большее перемещение у стадиона.

Если обратиться к результатам кластеризации методами K-Means и агломеративной кластеризации, результирующие кластеры выглядели следующим образом соответственно:

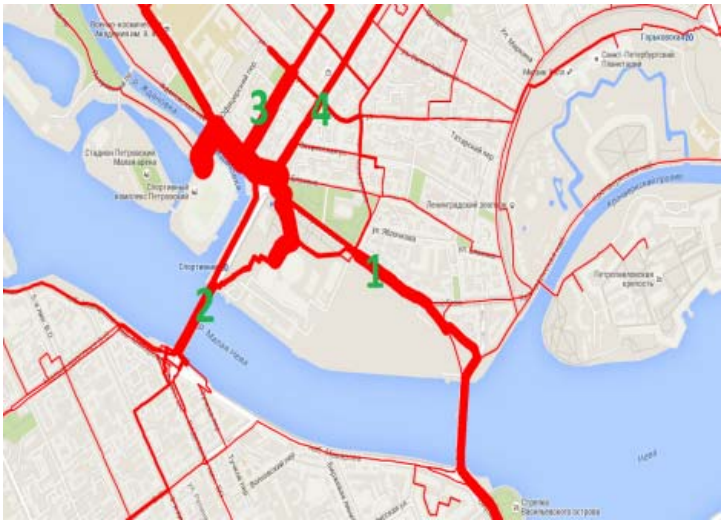


K-Means



Агломеративная кластеризация

В попытке выделить популярные маршруты этими методами мы столкнулись с проблемой недостаточно точной информации. Для большого количества абонентов число активностей было сравнительно малым, а расстояния между этими активностями (потреблением любого вида трафика) – большим. Поэтому для уточнения самых популярных путей в данном случае мы решили прибегнуть к средству прокладывания маршрутов из GoogleMapsDirectionsAPI. Мы для каждой пары действий абонента проложили маршрут и полученные данные для каждого ребра и пары вершин визуализировали с учетом разбиения абонентов по кластерам. Таким образом, нам удалось выделить 4 наиболее популярных маршрута:



1 – пр. Добролюбова и Биржевой мост
2 – Тучков мост
3 – Малый проспект Петроградской стороны
4 – Большой проспект Петроградской стороны

Схожие результаты показала визуализация маршрутов на основе метода кластеризации GRGPF:

Качество кластеризации было оценено с помощью SSQ – квадраты расстояний между объектами каждого кластера, а схожие карты популярных маршрутов, построенных двумя различными методами кластеризации, позволяют говорить о правильной группировке полученных клиентских данных.

Заключение

В заключение хотелось бы уточнить, что рассмотренные примеры представляют не только теоретический, но и практический интерес, о котором говорилось в начале статьи. В услови-

ях необходимости сокращения коммерческих затрат, повышения показателей эффективности бизнеса, относительной стагнации рынка мобильной связи в России компании вынуждены искать дополнительные меры по оптимизации своих затрат. Данный анализ может позволить не только сэкономить на маркетинговых исследованиях, которые, скорее всего, должны были показать аналогичные результаты и выдать схожие рекомендации по размещению рекламных носителей, но и использовать эти наработки в решении других задач: размещение торговых точек, предоставляющих продукты и сервис оператора, планирование дополнительных размещений базовых станций в часы пиковых нагрузок и т. д.

Итоги полученной работы в целом логичны. Все подходы к стадиону показывают достаточно высокую проходимость, однако результаты и ожидания приоритетов внутри популярных маршрутов разошлись. Если ожидания от маршрута по Тучкову мосту совпали с результатами, и он действительно оказался одним из двух самых популярных маршрутов, то ожидаемо вторые по популярности проспекты Петроградской стороны сильно уступили Биржевому мосту. На момент анализа никаких ремонтных работ на этих участках не велось, что говорит о том, что на результаты не влияли внешние городские факторы. Биржевой мост является хорошей рекламной площадкой, и на подходах к нему и на самом мосту сконцентрировано много рекламных носителей Оператора. Проведенный анализ говорит о том, что концентрацию на этих маршрутах стоит усилить, снизив ее соответственно на Петроградской стороне, по крайней мере в дни проведения рекламных акций, связанных с кобрендингом «Зенита» и Оператора. Низкая концентрация потоков на дальних подходах к стадиону также объяснима: достаточно высокий процент зрителей добирается до стадиона на автотранспорте, но в связи с проблемами с парковочными местами оставляет автомобили далеко от стадиона. За рулем трафиковая активность клиентов, разумеется, ниже, плюс перемещение между базовыми станциями происходит достаточно быстро, что не позволяет регистрировать клиента на каждом отрезке его пути. Этот вывод также может служить рекомендацией Оператору, направленной на концентрацию рекламных носителей на пути к стадиону. Просматриваемость носителей за рулем, если клиент не стоит в пробке, на порядок ниже, поэтому и их эффективность ниже, чем если бы они стояли на пути его пешеходного маршрута. Поэтому размещение на подобных носителях в дни проведения акций и рекламных кампаний стоит производить на маршрутах, по которым клиент идет пешком, уже оставив автомобиль.

Несомненно, построение маршрутов и кластеризации путей абонентов – задача достаточно объемная и может представлять интерес не только для оператора, но и для государственных органов, для других коммерческих организаций. Результаты кластеризации маршрутов можно связать с задачами роста проникновения продуктов на основе технологий RFID и iBeacon, задачами разделения транспортных и пешеходных потоков и др. В своих дальнейших исследованиях мы обязательно вернемся к задачам построения маршрутов, поскольку в их решении есть несомненная практическая значимость.

Список литературы

1. *Невоструев К. Н.* Обзор литературы по методам машинного обучения (Machine Learning) // Компьютерные инструменты в образовании. 2015. № 4. С. 19–26.

2. *Пономарев А. А.* Варианты использования больших данных в телекоммуникационном бизнесе // Компьютерные инструменты в образовании. 2015. № 4. URL: http://ipo.spb.ru/ journal/index.php?article/1781/ (дата обращения 25.01.2017).

3. *Laasonen Kari.* Clustering and Prediction of Mobile User Routes from Cellular Data // Knowledge Discovery in Databases: PKDD, 2005. P. 569–576.

4. *Leskovec J. A., Rajaraman J. D.* Ullman. Mining of Massive Datasets. Cambridge University Press, 2011.

Материал поступил в редколлегию 02.02.2017

A. A. Ponomarev

Scartel Ltd.
126B Savushkina Str., St. Peterburg, 197374, Russian Federation

artem.ponomarev@gmail.com

BIG DATA USAGE IN SUBSCRIBERS ROUTES CONSTRUCTING BY MOBILE OPERATORS

The article observes general questions and variants of big data usage in commercial enterprises, particularly in telecommunications. The problem of big data functional is discussed, and the approach of this functional as well. The author makes an accent on the certain problem of the cell operator, that can be solved with the help of data, the company has at its disposal. He analyzes and constructs most popular clients' routes during certain days to optimize advertizing placement and increase the effect of advertisement campaigns. The author identifies those route clusters, that show the most successful results

Keywords: big data, telecommunications, cell operator, mobile, routes, graphs, optimization, advertisement placement, clusters, advertisement campaigns.

References

1. Nevostruev K. N. Literature Review of Machine Learning Methods (Machine Learning). *Computer Instruments in Education*, 2015, no. 4, p. 19–26. (in Russ.)

2. Ponomarev A. Big data usage in telecommunications. *Computer Instruments in Education*, 2015, no. 4. URL: http://ipo.spb.ru/journal/index.php?article/1781/ (Application Date 25.01.2017). (in Russ.)

3. Laasonen Kari. Clustering and Prediction of Mobile User Routes from Cellular Data. *Knowledge Discovery in Databases*: PKDD, 2005, p. 569–576.

4. Leskovec J. A., Rajaraman J.D. Ullman. Mining of Massive Datasets. Cambridge University Press, 2011.

For citation:

Ponomarev A. A. Big Data Usage in Subscribers Routes Constructing by Mobile Operators. *Vestnik NSU*. *Series*: *Information Technologies*, 2017, vol. 15, no. 1, p. 70–78. (in Russ.)