НАУЧНО-ТЕХНИЧЕСКИЙ ВЕСТНИК ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ
сентябрь–октябрь 2017 Том 17 № 5 ISSN 2226-1494 http://ntv.ifmo.ru/
SCIENTIFIC AND TECHNICAL JOURNAL OF INFORMATION TECHNOLOGIES, MECHANICS AND OPTICS
September–October 2017 Vol. 17 No 5 ISSN 2226-1494 http://ntv.ifmo.ru/en

УДК 658.512.011.56:004.42

МЕТОД ОБРАБОТКИ В РЕАЛЬНОМ ВРЕМЕНИ ОТКРЫТЫХ ДАННЫХ, СОДЕРЖАЩИХ ГЕОКОНТЕКСТНУЮ РАЗМЕТКУ

М.М. Заславский[a], Э.И. Блеес[b], С.И. Баландин[c]

[a] Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация
[b] Санкт-Петербургский государственный электротехнический университет «ЛЭТИ» им. В.И. Ульянова (Ленина), Санкт-Петербург, 197376, Российская Федерация
[c] Технологический университет Тампере, Тампере, FI-33101, Финляндия
Адрес для переписки: mark.zaslavskiy@gmail.com

Информация о статье
Поступила в редакцию 04.07.17, принята к печати 22.08.17
doi: 10.17586/2226-1494-2017-17-5-850-858
Язык статьи – русский

Ссылка для цитирования: Заславский М.М., Блеес Э.И., Баландин С.И. Метод обработки в реальном времени открытых данных, содержащих геоконтекстную разметку // Научно-технический вестник информационных технологий, механики и оптики. 2017. Т. 17. № 5. С. 850–858. doi: 10.17586/2226-1494-2017-17-5-850-858

Аннотация

Предложено решение проблемы интерфейса обработки и исследования открытых данных в реальном времени с точки зрения выделения информации о местоположении в платформе, использующей данные о местоположении (LBS-платформе), путем создания дополнения к платформе, реализующего импорт и сопоставление открытых данных из нескольких источников для дальнейшего выделения дополнительного признака в элементах данных и обработки с использованием статистических показателей и кластеризации. В качестве апробации был разработан плагин определения популярности открытых точек доступа Wi-Fi с помощью данных социальной сети ВКонтакте для LBS-платформ Geo2Tag, реализующий предложенный метод решения проблемы производительности обработки открытых данных в LBS-платформах. Данный плагин осуществляет импорт и сопоставление набора открытых данных Правительства Санкт-Петербурга и архива записей в определенных районах города с выделением дополнительного признака при помощи вычисления медианы, среднего арифметического или центров кластеров по методу кластеризации k-means для количества записей. Для определения скорости работы плагина была проведена серия экспериментов по измерению его производительности. Экспериментальное исследование показало, что общее время работы плагина в первую очередь определяется скоростью загрузки открытых данных из источника, поскольку время обработки на порядок меньше времени загрузки. Результаты показывают, что плагин может осуществлять анализ открытых данных из удаленного источника практически в реальном времени. Разработанный метод может быть применен не только в LBS-платформе Geo2Tag, но и для широкого класса подобных систем, так как его реализация полагается только на наличие подсистемы импорта открытых данных, которая, в свою очередь, может быть реализована в любой LBS-платформе. Метод также создает конкурентное преимущество для LBS-платформы, так как позволяет расширить качественный состав данных за счет результатов анализа импортированных открытых данных, причем способы и методики анализа, а также конечная форма представления результатов могут определяться не только администраторами LBS-платформы, но и ее пользователями-разработчиками, поскольку реализация метода опирается на подсистему пользовательских дополнений.

Ключевые слова

Location-based services, Geo2Tag, LBS-платформа, открытые данные

METHOD FOR REAL TIME PROCESSING OF OPEN DATA CONTAINING GEOCONTEXT MARKUP

M.M. Zaslavskiy[a] , E.I. Blees[b], S.I. Balandin[c]

[a] ITMO University, Saint Petersburg, 197101, Russian Federation
[b] Saint Petersburg State Electrotechnical University (“LETI”), Saint Petersburg, 197376, Russian Federation
[c] Tampere University of Technology, Tampere, FI-33101, Finland
Corresponding author: mark.zaslavskiy@gmail.com

Article info
Received 04.07.17, accepted 22.08.17
doi: 10.17586/2226-1494-2017-17-5-850-858
Article in Russian

For citation: Zaslavskiy M.M., Blees E.I., Balandin S.I. Method for real time processing of open data containing geocontext markup. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2017, vol. 17, no. 5, pp. 850–858 (in Russian). doi: 10.17586/2226-1494-2017-17-5-850-858

Abstract

The paper proposes solution for the problem of real time processing and analysis interface for Open Data containing geocontext markup inside location-based services (LBS) platform. Solution method is based on providing the ability to extend a platform by the addition that implements import and mapping of several Open Data sets. This addition also should perform calculation of additional data attributes and processing of set element by set statistical indicators and clustering methods. The plug-in for Geo2Tag LBS-platform was developed for the proposed method approval. The solution determines popularity of open WiFi hot spots with the use of VK social network data. The plug-in performs import and mapping of Saint Petersburg government Open Data set and VK posts archive related to particular city districts. The mapping is performed by calculation of additional data attribute related to posts distribution - median, arithmetical mean or cluster centers by k-means clustering method. For plug-in performance evaluation, a series of experiments was performed. Analysis of experiment results showed that plug-in work time firstly depends on Open Data downloading speed because the time of Open Data processing by plug-in is an order of magnitude less than the loading time. This fact demonstrates that plug-in can perform almost real-time analysis of remote Open Data source. Developed method can be applied not only for Geo2Tag LBS-platform but also for broad set of similar systems because the solution depends only on Open Data import subsystem implementation that can be implemented on any LBS-platform. The method also gives competitive advantage for LBS-platform because it gives the possibility to extend qualitative composition of platform data by imported Open Data analysis results wherein analysis methods can be defined not only by LBS-platform administrators but also by platform users who are also the developers due to the fact that the method is based on user's plug-in subsystems.

Keywords

Location-Based Services, Geo2Tag, LBS-platform, Open Data

Введение

Проблема интерфейса обработки и исследования открытых данных с точки зрения выделения информации о местоположении актуальна в связи с быстрым ростом объема открытых данных [1] и активным внедрением технологий контекстных вычислений [2]. Согласно прогнозу на 2019 год, этот фактор приведет к серьезному росту рынка Location Based Services (LBS) за счет интеграции технологии контекстной разметки с существующими технологиями LBS-платформ [3]. Существующие на текущий момент инструменты исследования не могут решить данную проблему, поскольку они предлагают либо достаточно низкий уровень взаимодействия с данными (базы геоданных [4]), либо в общем случае не учитывают специфики геоданных и процедуры геоконтекстной разметки (BigData-решения [5]). Однако для BS-платформ в силу их предметной области нет подобных ограничений, благодаря чему в предыдущей работе [6] был продемонстрирован принципиальный метод решения данной проблемы путем организации подсистемы импорта открытых данных с пользовательскими расширениями для конкретных источников. В работе будет показано применение данного подхода с точки зрения обработки открытых данных LBS-платформой Geo2Tag в реальном времени. Под термином «в реальном времени» подразумевается такая скорость обработки, при которой время обработки сопоставимо или меньше, чем время передачи данных по сети между источником и LBS-платформой.

Обзор подходов к решению задачи

Для построения решения по обработке открытых данных в реальном времени с учетом геоконтекстной разметки необходимо изучить существующие инструменты. Тремя наиболее типичными способами решения поставленной задачи являются инструменты обработки Big Data, геоинформационные системы (ГИС) и LBS-платформы.

На сегодняшний день наиболее популярной методологией обработки больших объемов данных является Big Data, поэтому целесообразно начать обзор принципиальных инструментов геоконтекстной разметки именно с нее. Big Data рассматривает наборы данных, которые не могут быть обработаны традиционными средствами работы с данными ввиду сложности самих данных (слабая структурированность данных [7]), высокой скорости генерации и объема, многократно превышающего объем памяти обрабатывающего устройства[1]. Big Data решениями называют программные инструменты обработки данных, позволяющие вести их высокопроизводительную обработку и хранение. Примерами решений, позволяющими работать с Big Data, являются такие программы, как Hadoop[2], Spark[3], Cassandra[4], Oracle Big Data [8].

Говоря об обработке открытых данных, содержащих геоданные, необходимо также рассмотреть ГИС. ГИС – информационная система, обеспечивающая сбор, хранение, обработку, доступ, отображение

[1] The world's technological capacity to store, communicate, and compute information // Martin Hilbert. URL: http://www.martinhilbert.net/WorldInfoCapacity.html/ (дата обращения 01.07.2017).
[2] Apache Hadoop. URL: http://hadoop.apache.org/ (дата обращения 01.07.2017).
[3] Apache Spark. URL: https://spark.apache.org/ (дата обращения 01.07.2017).
[4] Apache Cassandra. URL: http://cassandra.apache.org/ (дата обращения 01.07.2017).

и распространение пространственно-координированных данных (пространственных данных)[1]. Понятие геоинформационной системы также используется в более узком смысле – как программного продукта, позволяющего пользователям искать, анализировать и редактировать как цифровую карту местности, так и дополнительную информацию об объектах [9]. Примерами таких систем являются различные геобазы данных, такие как ArcGIS[2]. К ним же можно отнести инструменты GIS as a service[3] – Google Maps[4], OpenStreetMap[5], CartoDB[6], MapBox[7].

Другим механизмом обработки больших объемов данных являются LBS-платформы, представляющие собой информационные системы (ИС), решающие такие же задачи, как и ГИС, но при этом основными пользователями являются разработчики прикладных программ и пользователи LBS [10]. Благодаря этому LBS-платформы позволяют проводить достаточно сложную обработку данных, используя интерфейсы достаточно высокого уровня. Наиболее популярными LBS-платформами, согласно данным поисковой системы Google[8], являются Geo2Tag[9], Anagog[10], Azoft LBS-platform[11], TomTom[12].

Сравнивая основные подходы к обработке больших объемов геоконтекстных данных, можно сделать вывод о том, что предпочтительной технологией являются LBS-платформы, так как они сочетают в себе возможность сложного анализа данных и предоставление интерфейсов высокого уровня. Инструменты для работы с BigData имеют общеее назначение и не учитывают специфику геоконтекстных данных [11]. Недостатком ГИС является то, что предоставляемые ими интерфейсы в основном предназначены для программистов, а интерфейсы более высокого уровня, как правило, предоставляют функции визуализации данных[5].

Обзор существующих решений задачи импорта и обработки открытых данных в LBS-платформах

Для определения принципиальной архитектуры предлагаемого интерфейса обработки и исследования открытых данных необходимо исследовать существующие интерфейсы работы с открытыми данными в LBS-платформах. Ограничим сравнение наиболее популярными LBS-платформами, перечисленными в предыдущем разделе:

- LBS-платформа Geo2Tag является наиболее популярной LBS-платформой с открытым исходным кодом[13]. Данное решение представляет собой набор базовых интерфейсов для построения приложений, использующих данные о местоположении, включая хранение, обработку и визуализацию данных. Более подробная характеристика LBS-платформы Geo2Tag дается в предыдущей работе [6].
- Сервис Anagog представляет собой высокоуровневый интерфейс к данным сенсоров, массово собираемым с реальных мобильных устройств. Пользователи данной LBS-платформы взаимодействуют с ней с помощью специального Software Development Kit (SDK), который осуществляет сбор данных и их отправку в облачные ресурсы Anangog, где происходит обработка и геоконтекстная разметка.
- Azoft LBS-platform – закрытая платформа, собирающая данные с мобильных устройств и предоставляющая готовые решения для клиентских сценариев использования.
- LBS-платформа TomTom представляет собой облачную программную платформу, позволяющую разработчикам использовать готовые решения в области навигации, построения маршрутов, отображения карт и получения актуальной информации о пробках. Для разработчиков предоставляется SDK.

Сравнение LBS-платформ будет проводиться с помощью критериев, определяющих удобство и эффективность обработки открытых данных в рамках LBS-платформы. Для настоящей работы были выбраны следующие критерии.

1. Возможность сохранения данных для анализа. Этот критерий подразумевает наличие интерфейсов для плагинов, позволяющих не только получать данные, но и сохранять их в LBS-платформу. Необходимо также отметить, что в более широком понимании практически каждая LBS-платформа имеет такие интерфейсы для своих пользователей, поэтому в данном критерии рассматриваются именно пользовательские дополнения. Соблюдение этого критерия обеспечивает дополнительное удобство решения задач обработки открытых данных.

[1] ГИС // ГИС ассоциация. URL: http://www.gisa.ru/13058.html (дата обращения 01.07.2017).
[2] ArcGIS. URL: https://www.arcgis.com/features/index.html (дата обращения 01.07.2017).
[3] Gis as a service // Thundermaps. URL: https://learn.thundermaps.com/blog-posts/what-is-gis-as-a-service-gaas/ (дата обращения 01.07.2017).
[4] Google Maps. URL: https://www.google.ru/maps (дата обращения 01.07.2017).
[5] OpenStreetMap. URL: https://www.openstreetmap.org/ (дата обращения 01.07.2017).
[6] CartoDB. URL: https://carto.com/ (дата обращения 01.07.2017).
[7] MapBox. URL: https://www.mapbox.com/ (дата обращения 01.07.2017).
[8] LBS Platform query // Google. URL: https://www.google.ru/?q=lbs+platform (дата обращения 01.07.2017).
[9] Geo2Tag. URL: http://www.geo2tag.com/ (дата обращения 01.07.2017).
[10] Anagog LBS. URL: http://anagog.com (дата обращения 01.07.2017).
[11] Azoft LBS. URL: http://www.azoft.com/solutions/lbs/ (дата обращения 01.07.2017).
[12] TomTom LBS. URL: http://www.tomtom.com/lib/doc/licensing/I.LBS.EN.pdf (дата обращения 01.07.2017).
[13] LBS Platform query // Google. URL: https://www.google.ru/?q=lbs+platform (дата обращения 01.07.2017).

2. Возможность создавать пользовательские дополнения (плагины) к платформе. Системы, реализующие данное требование, позволяют решать поставленную задачу наиболее простым и безопасным способом – с помощью программ, создаваемых пользователями и работающих в изолированной среде.
3. Открытость платформы, под которой подразумевается доступ к исходному коду. С точки зрения задачи импорта открытых данных в LBS-платформе это требование является одним из важнейших, так как при отсутствии механизмов подключения плагинов единственным способом расширить функциональность LBS-платформы остается модификация ее исходного кода. Кроме того, даже при наличии подсистемы плагинов в ряде случаев доступ к исходному коду может ускорить процедуру импорта открытых данных за счет использования недокументированных возможностей системы и оптимизации ее поведения для конкретного сценария.
4. Возможность импорта и обработки данных в реальном времени, что подразумевает под собой любые интерфейсы прикладного программирования для решения указанных задач, причем эти интерфейсы могут быть и не связаны между собой. Выполнение этого требования является одним из базовых критериев решения задачи обработки открытых данных в LBS-платформе, так как при его невыполнении реализация механизмов импорта и анализа ложится на сторонних разработчиков, которые, в свою очередь, могут ее решить неоптимальным образом.

Для сравнения существующих решений построим таблицу оценок по критериям.

	Возможность сохранения данных для анализа	Возможность создавать плагины к платформе	Открытость платформы	Возможность импорта и обработки данных в реальном времени
Anagog	Нет	Нет	Нет	Есть
TomTom	Нет	Нет	Нет	Есть
AzoftLBS	Нет	Нет	Нет	Нет
Geo2Tag	Есть	Есть	Есть	Есть

Таблица. Сравнение LBS-платформ

Как видно из сравнения, основной задачей LBS-платформы является выдача и обработка данных в реальном времени, что связывает их с Big Data решениями. Совокупности критериев наиболее полно удовлетворяет LBS-платформа Geo2Tag, так как она обладает интерфейсами создания плагинов, а также предоставляет возможность сохранения данных для последующей обработки.

Постановка задачи

Проведенный в предыдущем разделе обзор показал, что существующие методы работы с открытыми данными не позволяют вести их обработку эффективно в смысле использования специфичных для геоданных методов. Наиболее близкой к обозначенной проблеме технологией являются LBS-платформы, однако ее применение в решении требует расширения существующих платформ. По этой причине целью данной работы поставлено создание инструмента анализа открытых данных в виде расширения LBS-платформы Geo2Tag. Расширение должно обладать следующими свойствами:

– осуществлять импорт открытых данных из нескольких источников;
– реализовывать геоконтекстную разметку импортированных данных;
– предоставлять интерфейсы для обработки импортированных данных с помощью наиболее востребованных среди исследователей методов обработки;
– работа решения должна происходить в реальном времени, т.е. скорость обработки должна быть сопоставимой по скорости со временем передачи открытых данных по сети.

Предлагаемый способ решения

Решением поставленной задачи является разработка плагина импорта открытых данных, реализующего, помимо самого импорта, функции обработки самих данных. Данная форма решения наиболее подходит для задачи, так как плагины импорта открытых данных являются единственным способом расширения указанной LBS-платформы, что было показано в предыдущей работе [6]. Кроме того, плагин удовлетворяет всем требованиям к решению:

– по результатам работы [6] в LBS-платформу Geo2Tag добавлены программные интерфейсы, позволяющие вести импорт и обработку данных из произвольных источников;
– подсистема плагинов позволяет проводить в фоновом режиме операции любой сложности над данными платформы[1].

Недостатком выбранной формы решения являются риски с точки зрения информационной безопасности, так как на текущий момент LBS-платформа Geo2Tag не предусматривает методов ограничения потенциально опасных действий плагинов.

[1] Geo2Tag plugins // Geo2Tag. URL: http://geo2tag.org/?tag=plugins&lang=ru_RU (дата обращения 01.07.2017).

Выбор метода решения

Для решения задачи обработки открытых данных в реальном времени необходимо определить типичные задачи, решаемые потребителями открытых данных вручную. Поскольку наиболее структурированное описание методик работы с данными содержится в научных текстах, был проведен обзор актуальных статей, посвященных исследованию открытых данных с точки зрения содержащегося в них геоконтекста. Для изучения отбирались научные статьи за 2016 год, чья тематика сильнее всего пересекается с поставленной проблемой. В результате были отобраны две статьи – [12] и [13].

В статье [12] было проведено исследование, основанное на обработке нескольких наборов данных, полученных из социальной сети Instagram. Предметом исследования статьи являлся анализ регулярных перемещений жителей Амстердама и Копенгагена с точки зрения выявления закономерностей. В исследовании использовались различные методы анализа данных – анализ связности, анализ сетей, классификация, кластеризация. При этом данные анализируются как с помощью отдельных их атрибутов, так и с помощью дополнительно вычисляемых признаков.

В статье [13] дается подробное описание технологий анализа больших геопространственных наборов данных в контексте исследования концепции Smart City. В качестве материала для исследования используются наборы открытых данных, содержащие информацию о перемещениях людей в городах. Методом обработки выступает вычисление статистических квантилей и моментов – медиан и математических ожиданий для выборок.

На основании анализа статей можно сделать выводы о том, что общими характеристиками обоих исследований являются:

1. сопоставление нескольких наборов данных;
2. вычисление дополнительных признаков для элементов наборов открытых данных, в том числе и для отдельных подмножеств выборки.

Эти характеристики определяют функциональное наполнение разрабатываемого решения. Для определенности будем считать, что в качестве дополнительных признаков плагин будет вычислять числовые статистические характеристики выборки, а также производить кластеризацию элементов. Реализация методов теории графов (анализ связности и анализ сетей) выходит за рамки задач, решаемых LBS-платформой Geo2Tag, и в силу отсутствия необходимых интерфейсов в СУБД MongoDb[1] может привести к сильному снижению производительности, что, в свою очередь, может привести к невыполнению требования обработки в реальном времени. Процедуры классификации объектов в неявном виде присутствуют в процессе импорта открытых данных, поэтому их реализация не требуется [6].

Поскольку существующие интерфейсы для разработки плагинов импорта и обработки открытых данных в LBS-платформе подразумевают асинхронную работу, это накладывает дополнительные ограничения на принципиальную архитектуру плагина, помимо требования обработки в реальном времени. В этих условиях он должен снизить время задержки между запуском и сохранением в платформу данных, пригодных для пользователей. Для достижения этого эффекта предлагается разделение процедуры обработки на две асинхронные задачи:

1. импорт данных из всех необходимых наборов и приведение их к формату LBS-платформы;
2. сопоставление наборов и вычисление числовых статистических характеристик выборки, кластеризация.

Такое разделение обеспечивает достижение необходимого уровня производительности, так как при выполнении указанных задач последовательно пользователи уже в процессе выполнения задачи импорта имеют доступ к данным.

Разработанное решение

В качестве решения поставленной задачи в рамках LBS-платформы Geo2Tag был реализован плагин обработки и импорта открытых данных в базу данных (БД) платформы, сопоставляющий два набора открытых данных и осуществляющий вычисление дополнительных признаков для элементов данных. Архитектура плагина изображена на рис. 1.

В качестве модельной задачи, решаемой плагином, было выбрано определение популярности открытых точек доступа Wi-Fi в г. Санкт-Петербурге с помощью статистики записей в социальной сети ВКонтакте. Источником данных об открытых точках доступа Wi-Fi был выбран соответствующий набор с портала открытых данных Санкт-Петербурга[2]. Указанная задача является актуальной, так как ограниченный ресурс сетей расходуется различным образом из-за разного количества пользователей. Кроме того, задача определения популярности открытых Wi-Fi подходит для демонстрации решения поставленной ранее проблемы обработки открытых данных LBS-платформами в реальном времени, так как она подра-

[1] MongoDB. URL: https://www.mongodb.com/ (дата обращения 01.07.2017).
[2] Free SPb Wi-Fi // Spb Open data. URL: http://data.gov.spb.ru/opendata/7825457753-free_wi-fi/ (дата обращения 01.07.2017).

зумевает сопоставление данных двух различных наборов, и при этом используется достаточно большой объем данных.

Данные, получаемые из источников, имеют формат JSON. Работа с этим форматом в Python ведется с помощью инструментов модуля JSON. После обработки данных и получения дополнительных сведений формируется JSON, соответствующий схеме Geo2Tag, также с помощью инструментов модуля JSON.

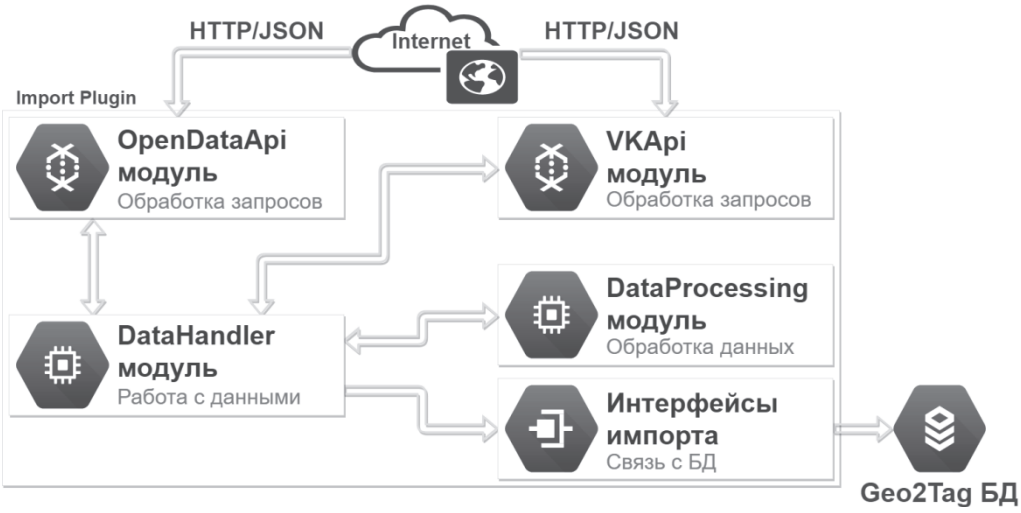

Рис. 1. Архитектура плагина

OpenDataApi module выполняет загрузку данных о бесплатных точках Wi-Fi по Санкт-Петербургу, которые размещены в открытом доступе на сайте открытых данных Санкт-Петербурга[1]. VKApi module выполняет загрузку данных о записях в социальной сети ВКонтакте за определенный промежуток времени в указанной области, используя методы VK API. DataHandler module получает данные из двух других модулей – OpenDataApi module и VKApi module и обрабатывает их. Интерфейсы импорта – реализованные интерфейсы импорта данных в платформу Geo2Tag. Исходный код плагина опубликован в открытом репозитории [14].

Первым этапом анализа импортированных объектов является использование координат точек Wi-Fi из первого источника для получения данных о записях в социальной сети ВКонтакте за определенное время. Вычисляется среднее по городу значение записей в области, и в сравнении с этим значением генерируется новый отличительный признак точки Wi-Fi – ее популярность. Формируется JSON в формате, хранимом в Geo2Tag. Данные загружаются в базу.

Дополнительный признак будет считаться следующим образом: получив данные об открытых городских точках доступа Wi-Fi по Санкт-Петербургу, система также получает данные из социальной сети ВКонтакте о количестве записей за определенное время в радиусе действия этих точек Wi-Fi. Затем, в зависимости от выбранного метода обработки данных, могут вычисляться:

1. среднее арифметическое количества записей на точку. В зависимости от полученного значения оценивается популярность точки Wi-Fi;
2. медиана количества записей на точку. В зависимости от полученного значения оценивается популярность точки Wi-Fi;
3. кластеризация точек по количеству записей. В зависимости от выделенных кластеров оценивается популярность точки Wi-Fi.

Исследование свойств решения

Для исследования скорости работы предлагаемого решения были проведены автоматизированные тестовые запуски плагина с разными входными данными. Измерялось два типа зависимостей: зависимость времени работы от количества точек Wi-Fi и от размера временного промежутка, в рамках которого запрашивались данные из социальной сети ВКонтакте. Тесты проводились в следующих условиях: ноутбук с характеристиками, соответствующими средним в доступном ценовом сегменте (процессор Intel® Core™ i5-3210M с частотой 2,5 Ггц, 4 ГБ оперативной памяти; операционная система Ubuntu 16.04.2 LTS (Xenial Xerus); заявленная скорость интернет-соединения 100 Мбит/сек. Размер одного JSON-ответа на запрос к VK API составляет от 0 до 40 КБ. Следует добавить, что использовалась реализация алгоритма кластеризации из библиотеки Python-cluster[2], количество выделяемых кластеров – 2. В качестве функции

[1] Spb open data. URL: http://data.gov.spb.ru/ (дата обращения 01.07.2017).
[2] Python-cluster. URL: https://github.com/exhuma/python-cluster (дата обращения 01.08.2017).

расстояния между элементами использовалась евклидова метрика, условием остановки итерационного процесса служило стандартное ограничение на изменение положений главных точек кластеров[1].

По результатам экспериментов было проведено исследование зависимостей времени работы плагина от количества точек Wi-Fi для всех методов обработки (рис. 2 и рис. 3). Перед сравнением времени работы всех методов необходимо исследовать на примере одного из них (базового) вид данной зависимости. В качестве базового метода было выбрано «среднее арифметическое» как самое простое для вычислений. Исследуемая зависимость изображена на рис. 2, для нее также были построены линии тренда с помощью метода наименьших квадратов (МНК). Необходимо отметить, что данные содержат шум, который обусловлен в первую очередь нестабильностью сетевого соединения, что следует из идентичных форм кривых на рис. 2.

Рис. 2. График зависимости времени работы плагина от количества анализируемых точек Wi-Fi, метод обработки «среднее арифметическое»: 1 — работа плагина, 2 — работа плагина (МНК), 3 — загрузка из ВКонтакте, 4 — загрузка из ВКонтакте (МНК))

Как видно из рис. 2, разница между полным временем обработки данных плагином по методу «среднее арифметическое» ($0{,}1036497003x + 0{,}5379161301$) и собственно временем загрузки данных по сети ($0{,}1032361531x + 0{,}0402164269$) на порядок меньше обеих величин ($0{,}0004135472x + 0{,}4976997032$), следовательно, большую часть времени работы плагина занимает именно загрузка данных из социальной сети ВКонтакте. При этом наблюдается линейный рост времени загрузки данных в зависимости от количества точек.

На рис. 3 изображен график зависимости времени работы плагина от количества анализируемых точек Wi-Fi при каждом методе обработки данных.

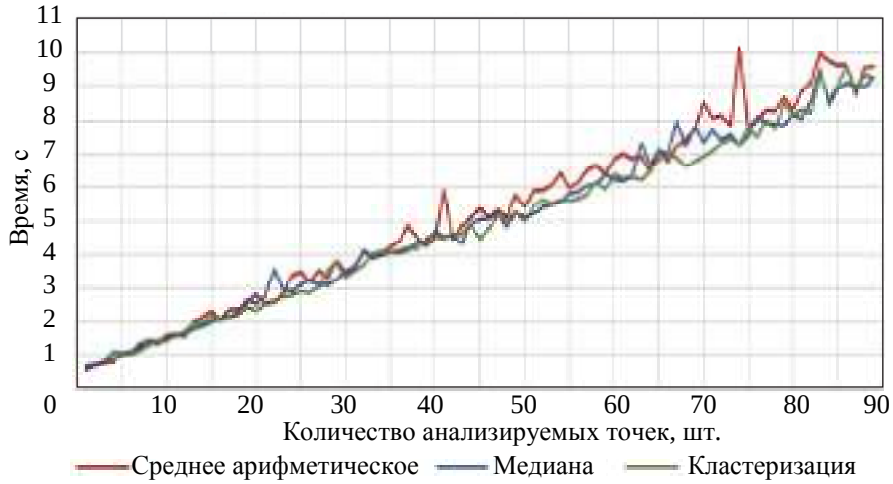

Рис. 3. График зависимости времени работы плагина от количества анализируемых точек Wi-Fi при каждом методе обработки данных

[1] A Tutorial on Clustering Algorithms. K-Means Clustering. URL: https://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html (дата обращения 01.08.2017)

Как видно на рис. 3, время работы плагина при разных методах обработки данных практически идентично. Это объясняется тем, что, согласно показанным выше измерениям, большую часть времени работы плагина занимает загрузка данных из социальной сети ВКонтакте. В связи с этим стоит сравнить время обработки данных разными методами – график изображен на рис. 4.

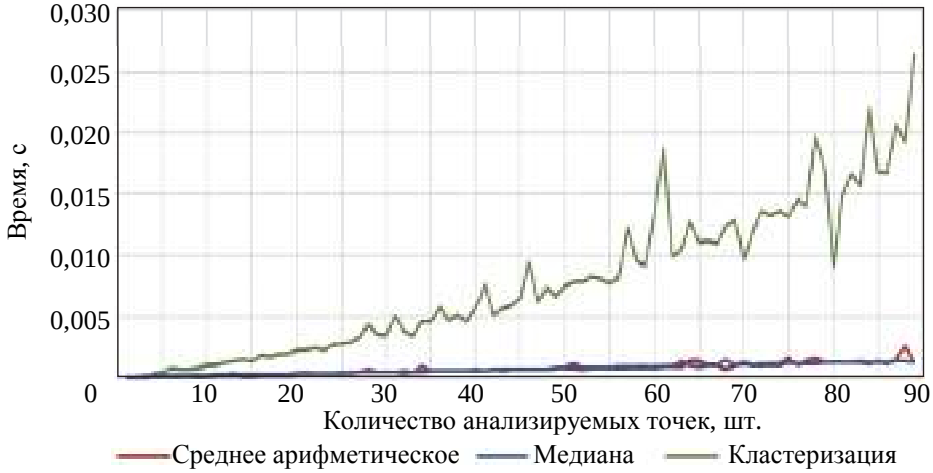

Рис. 4. График зависимости времени обработки данных различными методами от количества анализируемых точек Wi-Fi

Как видно из рис. 4, время обработки данных методом кластеризации больше, чем время обработки данных с выделением среднего арифметического или медианы набора. При наибольшем количестве анализируемых точек, использованных в эксперименте, разница во времени обработки данных методом кластеризации и методами «среднее арифметическое» и «медиана» достигает 15–20 раз. Это связано с тем, что зависимость времени выделения дополнительного признака при использовании методов среднего арифметического и медианы от количества анализируемых точек Wi-Fi – линейная, а зависимость времени выделения дополнительного признака при использовании метода кластеризации – степенная. Отметим, что выбранный алгоритм кластеризации имеет тот же самый характер зависимости скорости работы от объема обрабатываемых данных [15].

Заключение

В ходе работы было показано, что решение по импорту и обработке открытых данных в реальном времени с активной обработкой геоинформации возможно только в рамках реализации дополнения к LBS-платформе.

Для апробации выводов был создан плагин обработки и импорта открытых данных в LBS платформу Geo2Tag. Плагин загружает и сопоставляет данные о местоположении открытых Wi-Fi сетей Санкт-Петербурга с соответствующими по географическому местоположению данными о записях в социальной сети ВКонтакте для определения популярности отдельных сетей. Обработка данных после сопоставления производится с помощь вычисления численных статистических характеристик (медианы и среднего арифметического) и кластеризации по методу k-means.

Скорость обработки данных плагином была изучена экспериментально. По результатам исследования можно сделать вывод, что время осуществления процедур анализа данных самим плагином на порядок меньше, чем время загрузки данных из источника. Следовательно, предложенный подход позволяет организовать обработку открытых данных для LBS-платформ в реальном времени.

Разработанный метод был интегрирован в LBS-платформу Geo2Tag, и его внедрение показало, что метод позволяет расширить качественный состав данных, предоставляемых LBS-платформой, результатами анализа открытых данных. Кроме того, специфика метода позволяет реализовывать методы анализа и форматы представления его результатов силами пользователей-разработчиков LBS-платформы, что создает ей дополнительное конкурентное преимущество.

В дальнейших работах планируется расширить функции обработки данных за счет более сложных алгоритмов исследования данных, подключить новые форматы источников и более детально исследовать расход ресурсов LBS-платформы в процессе работы плагина.

Литература

1. Marr B. Big Data: 20 mind-boggling facts everyone must read // Forbes Magazine. 2015.

References

1. Marr B. Big Data: 20 mind-boggling facts everyone must read. *Forbes Magazine*, 2015.

2. Dey A.K. Understanding and using context // Personal and Ubiquitous Computing. 2001. V. 5. N 1. P. 4–7. doi: 10.1007/s007790170019
3. Basiri A. et al. Challenges of location-based services market analysis: current market description // Lecture Notes in Geoinformation and Cartography. 2014. P. 273–282. doi: 10.1007/978-3-319-11879-6_19
4. Peterson M.P. (ed.) Online Maps with APIs and WebServices. Springer, 2012. 318 p. doi: 10.1007/978-3-642-27485-5
5. Dittrich J., Quiane-Ruiz J.A. Efficient big data processing in Hadoop MapReduce // Proc. VLDB Endowment. 2012. V. 5. N 12. P. 2014–2015.
6. Заславский М.М., Баландин С.И. Метод импорта и обработки открытых данных в LBS-платформе // Научно-технический вестник информационных технологий, механики и оптики. 2016. Т. 16. № 5. С. 816–822. doi: 10.17586/2226-1494-2016-16-5-816-822
7. Abiteboul S. et al. The Lorel query language for semistructured data // International Journal on Digital Libraries. 1997. V. 1. N 1. P. 68–88.
8. Dijcks J.P. Oracle: Big data for the enterprise // Oracle White Paper. 2012.
9. Foote K.E., Lynch M. Geographic Information Systems as an Integrating Technology: Context, Concepts, and Definitions. 1996.
10. Weinreich A.P., Levy J.E., Barron G. Location-Based Services Platform. Patent US8447332. 2013.
11. Nandimath J. et al. Big data analysis using Apache Hadoop // Proc. IEEE 14th Int. Conf. on Information Reuse and Integration (IRI). San Francisco, USA, 2013. P. 700–703. doi: 10.1109/iri.2013.6642536
12. Boy J.D., Uitermark J. How to study the city on Instagram // PloS One. 2016. V. 11. N 6. P. e0158161. doi: 10.1371/journal.pone.0158161
13. Lohan E.S., Kauppinen T., Chandra Debnath S.B. A survey of people movement analytics studies in the context of smart cities // Proc. 19th FRUCT Conference. Jyväskylä, Finland, 2016. doi: 10.23919/fruct.2016.7892195
14. Import Plugin Repository. URL: https://bitbucket.org/EduardBlees/importplugin (дата обращения: 17.04.2017).
15. Inaba M., Katoh N., Imai H. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering // Proc. 10th ACM Symposium on Computational Geometry. NY, 1994. P. 332–339. doi:10.1145/177424.178042

2. Dey A.K. Understanding and using context. *Personal and Ubiquitous Computing*, 2001, vol. 5, no. 1, pp. 4–7. doi: 10.1007/s007790170019
3. Basiri A. et al. Challenges of location-based services market analysis: current market description. *Lecture Notes in Geoinformation and Cartography*, 2014, pp. 273–282. doi: 10.1007/978-3-319-11879-6_19
4. Peterson M.P. (ed.) *Online Maps with APIs and WebServices*. Springer, 2012, 318 p. doi: 10.1007/978-3-642-27485-5
5. Dittrich J., Quiane-Ruiz J.A. Efficient big data processing in Hadoop MapReduce. *Proc. VLDB Endowment*, 2012, vol. 5, no. 12, pp. 2014–2015.
6. Zaslavskiy M.M., Balandin S.I. Method of open data import and processing in LBS-platform. *Scientific and Technical Journal of Information Technologies, Mechanics and Optics*, 2016, vol. 16, no. 5, pp. 816–822. (In Russian) doi: 10.17586/2226-1494-2016-16-5-816-822
7. Abiteboul S. et al. The Lorel query language for semistructured data. *International Journal on Digital Libraries*, 1997, vol. 1, no. 1, pp. 68–88.
8. Dijcks J.P. Oracle: Big data for the enterprise. *Oracle White Paper*, 2012.
9. Foote K.E., Lynch M. *Geographic Information Systems as an Integrating Technology: Context, Concepts, and Definitions*. 1996.
10. Weinreich A.P., Levy J.E., Barron G. Location-Based Services Platform. *Patent US8447332*, 2013.
11. Nandimath J. et al. Big data analysis using Apache Hadoop. *Proc. IEEE 14th Int. Conf. on Information Reuse and Integration (IRI)*. San Francisco, USA, 2013, pp. 700–703. doi: 10.1109/iri.2013.6642536
12. Boy J.D., Uitermark J. How to study the city on Instagram. *PloS One*, 2016, vol. 11, no. 6, p. e0158161. doi: 10.1371/journal.pone.0158161
13. Lohan E.S., Kauppinen T., Chandra Debnath S.B. A survey of people movement analytics studies in the context of smart cities. *Proc. 19th FRUCT Conference*. Jyväskylä, Finland, 2016. doi: 10.23919/fruct.2016.7892195
14. *Import Plugin Repository*. URL: https://bitbucket.org/EduardBlees/importplugin (accessed: 17.04.2017).
15. Inaba M., Katoh N., Imai H. Applications of weighted Voronoi diagrams and randomization to variance-based k-clustering. *Proc. 10th ACM Symposium on Computational Geometry*. NY, 1994, pp. 332–339. doi:10.1145/177424.178042

Авторы

Заславский Марк Маркович – аспирант, Университет ИТМО, Санкт-Петербург, 197101, Российская Федерация, mark.zaslavskiy@gmail.com

Блеес Эдуард Игоревич – студент, Санкт-Петербургский государственный электротехнический университет (ЛЭТИ) им. В.И. Ульянова (Ленина), Санкт-Петербург, 197376, Российская Федерация, edw252@gmail.com

Баландин Сергей Игоревич – кандидат технических наук, адъюнкт-профессор, адъюнкт-профессор, Технологический университет Тампере, Тампере, FI-33101, Финляндия, Sergey.Balandin@fruct.org

Authors

Mark M. Zaslavskiy – postgraduate, ITMO University, Saint Petersburg, 197101, Russian Federation, mark.zaslavskiy@gmail.com

Eduard I. Blees – student, Saint Petersburg State Electrotechnical University (“LETI”), Saint Petersburg, 197376, Russian Federation, edw252@gmail.com

Sergey I. Balandin – PhD, Adjunct Professor, Adjunct Professor, Tampere University of Technology, Tampere, FI-33101, Finland, Sergey.Balandin@fruct.org