БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Факультет прикладной математики и информатики

Кафедра информационных систем управления

ЗИНЬКОВИЧ ИЛЬЯ АНДРЕЕВИЧ

ПЛАНИРОВАНИЕ ПАССАЖИРСКИХ ПЕРЕВОЗОК НА ОСНОВЕ ПРЕДИКТИВНОГО АНАЛИЗА ДАННЫХ СОЦИАЛЬНЫХ СЕТЕЙ

Отчет по преддипломной практике

студента 5 курса 2 группы

|  |  |
| --- | --- |
|  | **Руководитель**  Совпель Игорь Васильевич,  профессор, доктор технических  наук |

Минск 2017

**Содержание**

|  |  |  |
| --- | --- | --- |
|  | **ВВЕДЕНИЕ** | 3 |
|  | **ГЛАВА 1. ПОСТАНОВКА ЗАДАЧИ** | 4 |
|  | **ГЛАВА 2. ОБЗОР** | 5 |
|  | **ГЛАВА 3. РЕЗУЛЬТАТЫ** | 11 |
|  | **ГЛАВА 4. РЕКОМЕНДАЦИИ** | 16 |
|  | **БИБЛИОГРАФИЧЕСКИЙ СПИСОК** | 17 |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

**ВВЕДЕНИЕ**

Рынок пассажирских перевозок является одним из перспективных и быстрорастущих секторов высокоразвитых экономик. Развитие науки и техники позволило увеличить прибыльность и сократить расходы при использовании пассажирского транспорта. Большая конкуренция обеспечила улучшение сервиса и условий обслуживания клиентов. Появление информационных ресурсов, обеспечивающих высокий уровень интеграции с сопутствующими сервисами, такими как например гостиничный и ресторанный, поспособствовали увеличению пассажирооборота и развитие экономических связей не только внутри но и вне индустрии. Сильнейшие игроки из целого ряда областей заинтересованы в развитии рынка пассажирских перевозок. Выигрывает тот, кто лучше знает свой рынок, умеет распознать тенденции и предлагает то, в чем клиенты более всего заинтересованы.

На сегодняшний день едва ли кто-то станет отрицать повсеместное проникновение информационных устройств в жизнь современных людей.

То, что сейчас происходит можно оценивать положительно или отрицательно, но нельзя отрицать то, что всеобщая «интернетизация» создает как новые угрозы, так и новые перспективы в различных областях жизни человека.

Конкретно «цифровой след», который оставляют каждый, кто пользуется электронными средствами связи, открывает новые перспективы в развитии наук о человеке: обществе, демографии и социально-экономической географии.

Очень большим потенциалом обладают данные, которые доступны операторам сетей сотовой связи, которая стала обыденным явлением по всему миру, а количество действующих контрактов на услуги сотовой связи давно превысило население. Данные, имеющиеся в распоряжении оператора сотовой сети, такие как наличие подключения телефонного аппарата к базовой станции позволяют довольно точно определить местоположение абонента сети в каждый момент времени.

Эти факторы вместе создают предпосылки для беспрепятственного отслеживания перемещений всех абонентов сети, что значит большей части населения Земли. На данный момент каинформация о перемещениях каждого отдельно взятого человека защищена законом, но уже известны случаи использования агрегированных данных от операторов сотовой связи для анализа перемещений жителей внутри отдельно взятых городов[6].

Детальному анализу данных о перемещениях пользователей сотовых сетей препятствуют не только законы, но и трудоёмкость вычислений. Такие объемы информации — яркий пример «больших данных» (англ. «big data»), методы и алгоритмы обработки которых стали развиваться ускоренными темпами и привлекают внимание ученых и исследователей в области информационных технологий в последние годы.

В то же время кроме очень детальных и объемных данных, которые могут быть получены из сотовых сетей, имеются и другие источники «электронного следа» человека, не такие детальные и полные (но при этом менее трудоемкие в обработке), и размещенные самим человеком в открытом доступе. Речь идет о профилях пользователей социальных сетей. Возникновение термина «социальная сеть» в социологии произошло еще до появления современных вычислительных машин и обозначал модель взаимодействия акторов в обществе [9].

Социальные сети, появившиеся в начале 2000-х годов в сети Интернет, можно так или иначе считать «проекцией» социальных сетей в том смысле, как их понимают в социологии. Социальные сети, существующие на данный момент в Интернете, довольно разнообразны по своим целям, локации и преобладающему возрасту пользователей. Вместе с тем, сама природа социальной сети диктует наличие элементов, которые присутствуют в каждой социальной сети: профилей пользователей, отношений между пользователями и другими субъектами сети.

Данные в пользовательском профиле заполняются самим пользователем добровольно и размещаются в открытом доступе. Поэтому доступ к этим данным с точки зрения закона ничем не отличается от доступа к любым другим данным из сети Интернет. Извлечение и обработка этих данных в больших количествах требует серьезных технических средств, но это гораздо доступнее, чем данные из сетей мобильной связи.

Анализ социальных данных стремительно набирает популярность во всём мире благодаря появлению в 1990-х годах онлайновых сервисов социальных сетей (Facebook, Twitter, YouTube и другие). С этим связан феномен социализации персональных данных: стали публично доступными факты биографии, переписка, дневники, фото-, видео-, аудиоматериалы, заметки о путешествиях и т.д. Таким образом, социальные сети являются уникальным источником данных о личной жизни и интересах реальных людей. Это открывает беспрецедентные возможности для решения исследовательских и бизнес-задач (многие из которых до этого невозможно было решать эффективно из-за недостатка данных), а также создания вспомогательных сервисов и приложений для пользователей социальных сетей. Кроме того, этим обуславливается повышенный интерес к сбору и анализу социальных данных со стороны компаний и исследовательских центров.

С одной стороны, в данных из социальных сетей нет ничего необычного по сравнению с данными из других источников. Достоверность таких данных сравнима с достоверностью данных социальных опросов, которые до сих пор играли огромную роль в изучении и анализе различных социальных процессов, а также в анализе рынков, превнося некоторую уверенность в решения компаний, которые основывались на результатах такого анализа. Таким образом для анализа данных социальных сетей можно использовать давно устоявшиеся и зарекомендовавшие себя методы анализа социальных опросов, только масштабы уже совсем другие и на входе зачастую имеется информация, которую невозможно было достать привычными методами в таких количествах.

С другой стороны, данные из социальных сетей порой выглядят отлично от формата, привычного исследователям социальных опросов. Данные могут содержать в себе различные связи, о существовании которых ранее можно было только предполагать. Это меняет наше представление о данных вцелом, порождая новые способы их анализа, приводящие впоследствии к неожиданным результатам. Среди классических способов представления данных выделяется табличное представление, хорошо знакомое всем, кто имел дело со статистическими данными, где в результатах сбора данных становится известно о некотором множестве качественных и количественных характеристик объектов исследования. Способы анализа таких структур данных включают в себя нахождение корреляции между определенным подмножеством исходных характеристик объектов исследования. В решении таких задач сильно преуспели как классические статистические методы, так и методы машинного обучения, работающие на больших корпусах исходных данных. В то же время данные из социальных сетей зачастую имеют совершенно иную структуру, которая может быть представлена, например, в виде графа и отображающая отношения между объектами исследования. Причем отношения могут быть как простыми, означающими наличие или отсутствие связи между участниками, так и более сложными, имеющими определенные качественные и количественные характеристики такие как степень, с которой схожие участники формируют связи между собой в сравнении с несхожими (гомогенность), степень, с которой схожие участники формируют связи между собой в сравнении с несхожими (множественность), степень, с которой двое участников отвечают друг другу взаимностью в сфере дружеских или других взаимодействий (обоюдность), закрытость сети, степень, с которой двое участников отвечают друг другу взаимностью в сфере дружеских или других взаимодействий.

В частном секторе фирмы используют анализ социальных сетей для поддержки такой деятельности, как взаимодействие и анализ клиентов, маркетинг и бизнес-аналитика. Использование анализа социальных сетей государственным сектором включает в себя развитие стратегий участия руководства, анализ индивидуального и группового участия, использование средств массовой информации и основанное на сообществах решение проблем, а также их мониторинг. Мониторинг включает получение и структурирование первичных данных. Собираются тексты сообщений, связи между пользователями, ссылки на внешние ресурсы. Возможности таких систем во многом определяются богатством используемых данных и режимом их обработки. Системы, поддерживающие мониторинг в режиме реального времени, сложнее в реализации, чем использующие ретроспективный сбор данных.

Глобализация привела к увеличению популярности международных пассажирских перевозок. Границы стираются, люди проявляют больший интерес и имеют все больше возможности посетить другие страны, познакомиться с другими культурами. Всё это отражается в информации, появляющейся в социальных сетях. Информация о перемещении пользователей социальных сетей, а также оценка их собственных путешествий, общедоступна и может быть использована в качестве источника данных для исследования рынка пассажирских перевозок. Авиакомпании, операторы автобусного и железнодорожного транспорта и, как упоминалось ранее, гостиничный бизнес заинтересованы в получении информации о тенденциях рынка пассажирских перевозок. Современные средства и методы анализа данных могут помочь решить поставленную задачу.

**Глава 1. ПОСТАНОВКА ЗАДАЧИ**

Проблема существующих методов исследования рынка пассажирских перевозок, особенно международных, состоит в том, что игнорируется наличие общедоступной информации о перемещении пользователей социальных сетей.

Цель работы - разработать программное обеспечение для сбора информации о пользовательских перемещениях из социальных сетей и на основании собранной информации разработать программную модель для предсказания наиболее перспективных маршрутов пассажирских перевозок.

Объектом исследования являются общедоступные данные социальных сетей, содержащих информацию о перемещении пользователей, такие как место постоянного пребывания и данные о посещении других стран и городов, с учетом возможности подключения проприетарных источников подобной информации, таких как базы данных авиаперелетов, туристических туров, пассажирских перевозок.

Методология проведения работы: data mining, интеллектуальный анализ данных, анализ временных рядов, машинное обучение, big data.

Результаты работы: модели, алгоритмы, программный инструментарий и методика его применения.

Область применения результатов: логистика, бизнес-аналитика, маркетинг.

**Глава 2. ОБЗОР**

* 1. **Классический анализ пассажирских перевозок.**

Пассажирские перевозки - представляют собой перемещения людей с помощью различных видов транспорта [1]. Среди пассажирских перевозок можно выделить внутригородские, пригородные, междугородные и международные пассажирские перевозки. Постепенное развитие и улучшение качества пассажирских перевозок во всех сообщениях и направлениях, в том числе международных, играет большую роль как в экономическом плане, в виде налоговых отчислений в бюджеты разных уровней, так и в качестве развития отношений между различными регионами.

Рынок пассажирских перевозок анализируется с целью определения структуры и динамики пассажирских перевозок различными видами транспорта на конкретной территории за определенный промежуток времени (год, месяц, неделю). Анализ может проводиться как сотрудниками транспортных компаний так и независимыми исследователями с привлечением пользователей услуг пассажирских перевозок посредством опросов, а также при помощи анализа данных из других многочисленных авторитетных источников.

Основные этапы проведения анализа рынка пассажирских перевозок:

1. определение цели исследования;
2. выбор методов проведения исследования;
3. сбор и анализ вторичной информации;
4. проведение первичных исследований;
5. анализ данных, полученных в результате проведения первичных исследований;
6. разработка выводов, стратегий развития, прогнозов рынка пассажирских перевозок.

Методы проведения анализа рынка пассажирских перевозок подразделяются на кабинетные и полевые исследования.

Кабинетные исследования:

* анализ данных официальной статистики;
* анализ данных компаний-перевозчиков;
* анализ отраслевых СМИ;
* анализ результатов исследования рынка пассажирских перевозок, проведенных ранее.

Полевые исследования:

* проведение опросов потребителей в местах формирования пассажирского потока;
* проведение опросов потребителей в транспорте, при совершении ими поездок;
* наблюдение, организованное в транспорте;
* экспертное интервью с участниками рынка пассажирских перевозок, организация и проведение фокус-групп.

В ходе исследования рынка пассажирских перевозок могут быть решены следующие задачи:

* определение величины пассажиропотока на рассматриваемой территории за определенный промежуток времени (год, месяц, неделя, сутки);
* определение величины входного и выходного пассажиропотоков на рассматриваемой территории (год, месяц, неделя, сутки);
* определение структуры пассажиропотока в разрезе отдельных видов транспорта: воздушный, железнодорожный, автомобильный (автобусный), внутренний водный;
* определение наиболее загруженных видов транспорта и участков на транспорте.

Решение этих задач будет способствовать разработке и принятию мер по повышению эффективности деятельности различных видов транспорта на региональном, федеральном и международном уровнях.

В целях обеспечения единого понятийного аппарата введем определение понятия «пассажиропоток». Пассажиропоток представляет движение пассажиров в одном направлении маршрута. Таким образом, пассажиропоток может быть в прямом направлении (входной) и в обратном (выходной). Основные характеристики пассажиропотока — мощность (напряженность) и объем перевозок пассажиров. Мощность пассажиропотока, напряженность, определяется количеством пассажиров, которое проезжает в определенное время (год, месяц, неделя) на рассматриваемой территории в одном направлении (любым видом транспорта).

Объем перевозок пассажиров определяется количеством пассажиров, перевозимых рассматриваемым видом транспорта за определенный промежуток времени (год, месяц, неделя, сутки, час).

Кроме того, перевозки пассажиров различными видами транспорта по дальности осуществления перевозки можно классифицировать на четыре группы:

* Внутригородские перевозки пассажиров представляют перемещение пассажиров по территории города различными видами транспорта.
* Пригородные перевозки — это пассажирские перевозки между городами и прочими населенными пунктами в рамках одной области.
* Междугородние пассажирские перевозки предполагают осуществление пассажирских перевозок между населенными пунктами в рамках одной страны.
* Международные пассажирские перевозки предполагают пассажирские перевозки между населенными пунктами различных стран, позволяют осуществлять взаимодействие людей, проживающих на территориях различных стран.

Анализ пассажиропотока международных пассажирских перевозок является наиболее трудоемким, так классические методы их анализа усложняются за счет увеличения сложности получения данных для исследования.

* 1. **Анализ данных из социальных сетей**

Аналитическое агентство Gartner в 2016 году опубликовало отчёт под названием "Цикл ажиотажа для развивающихся технологий". Согласно отчёту, технологии "Социальная аналитика" и "Большие данные" в настоящее время находятся на т.н. "пике завышенных ожиданий". Компании-владельцы сервисов онлайновых социальных сетей (Facebook, Twitter) активно инвестируют в разработку усовершенствованных инфраструктурных и алгоритмических (новые алгоритмы поиска и рекомендации пользователей, товаров и услуг) решений для обработки больших массивов пользовательских данных. Возникают и успешно развиваются коммерческие компании, предоставляющие услуги по доступу к хранилищам социальных данных, сбору социальных данных по заданным сценариям, социальной аналитике, а также расширению существующих платформ с помощью социальных данных. Таким образом, специалисты из исследовательских центров и компаний по всему миру используют данные социальных сетей для моделирования социальных, экономических, политических и других процессов от персонального до государственного уровня с целью разработки механизмов воздействия на эти процессы, а также создания инновационных аналитических и бизнес-приложений и сервисов. Вместе с тем, при работе с социальными данными нужно принимать во внимание такие факторы, как нестабильность качества пользовательского контента (спам и ложные аккаунты), проблемы с обеспечением приватности личных данных пользователей при хранении и обработке, а также частые обновления пользовательской модели и функционала. Всё это требует постоянного совершенствования алгоритмов решения различных аналитических и бизнес-задач. Обработка социальных данных требует также разработки соответствующих алгоритмических и инфраструктурных решений, позволяющих учитывать их размерность. К примеру, база данных социальной сети Facebook на сегодняшний день содержит более 1 миллиарда пользовательских аккаунтов и более 100 миллиардов связей между ними. Каждый день пользователи добавляют более 200 миллионов фотографий и оставляют более 2 миллиардов комментариев к различным объектам сети. На сегодняшний день большинство существующих алгоритмов, позволяющих эффективно решать актуальные задачи, не способны обрабатывать данные подобной размерности за приемлемое время. В связи с этим, возникает потребность в новых решениях, позволяющих осуществлять распределённую обработку и хранение данных без существенной потери качества результатов.

Развитие информационных технологий, связанное с возможностью поддержки удалённого взаимодействия между пользователями сети интернет, благотворно влияет на стремительное развитие социальных сетей. Каждый год популярность и важность социальных сетей повышается. Этот процесс непосредственно имеет связь с использованием виртуальных сетей, которые позволяют общаться между собой людям, удалённым на огромных расстояниях. Использование социальных сетей стало важной частью жизни каждого из нас. Впоследствии онлайн сети оказывают непременное воздействие на жизнь человека и его взаимоотношения с окружающими его людьми. Сейчас наблюдается рост количества исследований, связанных с созданием алгоритмов, методов, а также приложений для изучения взаимоотношений между пользователями социальных сетей.

Компьютерные социальные сети становятся все более популярными. Одним из результатов взаимодействия людей посредством таких сетей является получение огромного количества информации различных форматов: тексты, картинки, аудио, видео. Для того чтобы делать определенные выводы о процессах, протекающих в обществе, прогнозировать поведение его участников, моделировать социальное взаимодействие, необходимо осуществлять анализ информации, существующей в социальных сетях. Поэтому создание новых и интеграция уже созданных методов и моделей анализа компьютерных социальных сетей представляет интерес для исследования.

Анализ социальных сетей – это активно развивающаяся область социологии. Исследования, ставящие перед собой задачу изучения социальных сетей, используют различные подходы для оценки зрелости онлайн сообществ их пользователей. В то же время изучается вопрос автоматизации жизни современного человека при помощи социальных сетей. Рассматриваются следующие направления автоматизации: информатизация бизнеса и деятельности государства и увеличение количества общения с друзьями, родственниками и коллегами. Автоматизация позволяет более полно удовлетворять потребности в самовыражении и в информации, а также значительно увеличивает количество инструментов самообслуживания.

Интерес разработчиков программного обеспечения, ученых и исследователей к данному направлению связан с тем, что оно предоставляет новый набор аналитических инструментальных средств, а также объяснительных моделей, результаты использования которых невозможно достичь использованием одних лишь количественных методов. Также в данной области сформирован отличный математический аппарат, включающий в себя различные расширения классических статистических методов и основ, позволяющий строить довольно сложные и в то же время исчерпывающие модели социальных взаимодействий, описывающие практически любые социальные системы. Социальные сети вероятно могут стать инструментом для модернизации общества, распространения различных форм знаний, формирования общественных движений и коллективов. На данный момент они используются обычно для общения, исследований рынка и продвижения продукции и услуг различных компаний. Социальные сети в сети Интернет в то же время представляют собой очень большой источник данных о развивающихся связях в обществе. Экономическая теория создания сети формирует гипотезы о том, какими ключевыми правилами и законами связей и экономических стимулов может быть сформирована архитектура сети, и помогает определить их эффективность. Эти гипотезы можно опытным путем протестировать используя данные из социальных сетей.

На данный момент активно изучается влияние таких социальных сетей как Facebook и Twitter на формирование социальных групп у студентов и молодежи. В то же время рассматриваются связи между социальной сетью и психологическим состоянием её участника, осуществляющего в ней взаимодействие с другими пользователями.

Twitter – социальная сеть для обмена краткими сообщениями – уже сейчас является обширной платформой для исследований, предоставляющей различные механизмы доступа к пользовательским данным, а также профессиональную поддержку со стороны разработчиков этой социальной сети.

С технической точки зрения массовое извлечение данных из социальных

сетей осуществимо достаточно просто, но все-таки находится за пределами возможностей рядового пользователя персонального компьютера. В задаче извлечения данных можно выделить три этапа:

1. Формирование списка необходимых профилей.

2. Выгрузка профилей из социальной сети.

3. Обработка профилей для извлечения данных.

Для проведения широкомасштабных исследований может иметь смысл создание «зеркала» социальной сети, т. е. сначала выгрузить все доступные профили, а затем уже выбирать среди них нужные, тем самым выполнив сначала пункты 2 и 3, а затем уже многократно выполнять пункт 1. Если же исследуемая генеральная совокупность мала по сравнению с общим массивом профилей (в настоящее время крупнейшие социальные сети насчитывают десятки миллионов пользователей), то разумнее выгружать и обрабатывать только требуемые в исследовании профили.

Кроме социальных сетей на данный момент изучаются другие инструменты, с помощью которых можно обрабатывать содержимое сети и собирать информацию о её участниках. Прикладные программные интерфейсы и методы также активно используются в современных исследованиях. Используя программные интерфейсы программисты могут создавать приложения, которые могут взаимодействовать с социальными сетями.

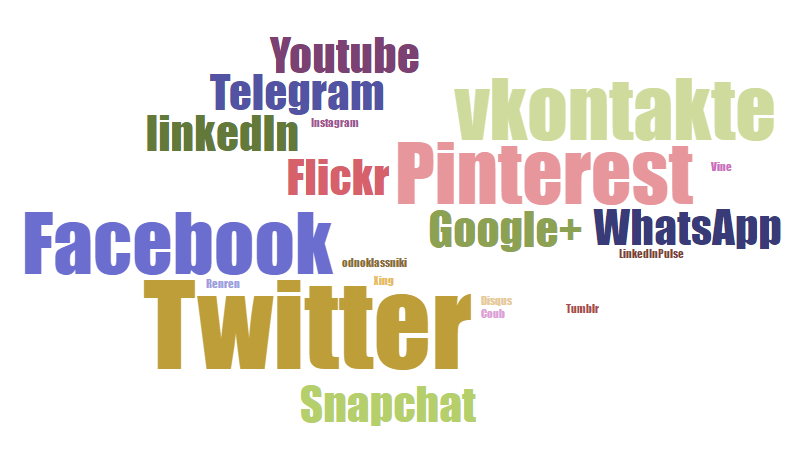
За последние годы популярность социальных сетей в мире неуклонно растет, а также увеличивается и их количество. Наиболее крупные из них, такие как Twitter, Facebook, MySpace, ВКонтакте, Одноклассники, Flickr (облако тегов изображено на рисунке 1), насчитывают миллионы зарегистрированных пользователей и гостей (незарегистрированные пользователи), которые каждый день генерируют огромные потоки информации в виде публикации различных заметок, аудио-, фото- и видеофайлов, ведения обсуждений в различных сообществах и форумах и т.д. Уже сейчас информации в социальных сетях столько, что пользователи не в силах разобраться в ней и быстро находить то, что им нужно. Для решения этой проблемы требуется разработка механизмов "умного" поиска и персонифицированного представления информационных ресурсов сети.

Рисунок 1 – Исследованные социальные сети.

Данные, доступные в социальных сетях, можно использовать как инструмент для проведения социальных опросов. Не смотря на то, что социологи с некоторой долей осторожности относятся к подобной практике, потому как обеспечить и доказать репрезентативность выборки в этом случае гораздо сложнее, чем при проведении классических видов опросов, в то же время данная практика все чаще используется в социологических исследованиях. В таких исследованиях обычно объектом является сама социальная сеть. Результаты этих исследований имеют значение для владельцев социальных сетей, так как они способствуют повышению качества «таргетирования» рекламной информации в соответствии с предпочтениями пользователей. С этой точки зрения исследования направлены на определения причастности пользователей к различным социальным группам, а также влияние их пола, возраста и других отличительных качеств на принятия решений.

* 1. **Сбор данных из социальных сетей**

Существующие в настоящее время в Интернете социальные сети весьма разнообразны по своим целям, месту проживания и возрастному составу пользователей. Вместе с тем, сама природа социальной сети диктует наличие элементов, присутствующих во всех сетях: а именно — страниц (профилей) пользователей и связей (как минимум, между пользователями, а зачастую еще и между другими объектами, выступающими в качестве узлов сети).

Данные в профиле заполняются пользователем самостоятельно, добровольно и размещаются в открытом доступе с его ведома. Таким образом, доступ к этим данным с правовой точки зрения ничем не отличается от доступа к любым другим данным из сети Интернет. Массовое извлечение и обработка этих данных требует определенных технических средств, однако это несоизмеримо доступнее, чем данные из сетей мобильной связи.

Доступные данные о пользователе, варьируются в зависимости от социальной сети. Если причислять к социальным сетям популярные сервисы блогов, микроблогов, то, из всей доступной широким массам профильной информации о пользователе гарантированно можно рассчитывать только на название пользовательского аккаунта. В других социальных сетях, где пользователи в первую очередь находят друзей и вступают в различные сообщества пользовательские профили становятся более развернутыми - пользователи указывают пол, имя, дату рождения, место проживания, учебы и работы, а также еще более специфичискую информацию в виде семейного положения, политических взглядов, отношения к курению и алкоголю, любимые книги, музыка, фильмы. Заполнение всех этих полей, конечно, дело добровольное, но, учитывая, что основная цель социальной сети поиск новых контактов и общение, то указание довольно подробных сведений о себе становится разумным с точки зрения большинства пользователей, которые хотят, чтобы их было легче найти их одноклассникам, сокурсникам, коллегам по работе.

Кроме того, пользовательский профиль содержит информацию о «друзьях» - связях с другими пользователями социальной сети, и тематических сообществах в которых состоит пользователь.

Кроме данных профиля, страницы пользователей зачастую содержат иные пользовательские данные: фотографии, сообщения, видео, музыку и другие документы.

Веб-интерфейсы социальных сетей являются источниками данных реального времени и предназначены для просмотра и взаимодействия со страницами социальной сети в веб-браузере либо для использования данных пользователей специализированными приложениями. Поскольку сценарии использования интерфейсов социальных сетей не предполагают автоматического сбора данных множества пользователей с целью построения социального графа, то возникает ряд проблем [2]:

1. приватность данных - зачастую доступ к данным пользователей разрешён только для зарегистрированных и авторизованных участников сети, что требует поддержки эмуляции пользовательской сессии с помощью специальных учётных записей (аккаунтов);

2. слабая структурированность данных - во многих случаях программные интерфейсы (API) социальных сетей имеют ограниченный функционал, что требует поддержки получения с помощью пользовательского веб-интерфейса статических копий HTML-страниц, корректной обработки их динамической части (включая исполнение асинхронных запросов к серверу социальной сети), извлечения нужных данных с помощью алгоритма и/или шаблона и построения их структурированного представления, удобного для дальнейшей автоматической обработки;

3. ограничения доступа и блокировки - с целью предотвращения несанкционированного автоматического сбора данных и ограничения нагрузки на инфраструктуру сервиса социальной сети владельцы сервисов зачастую вводят явные или скрытые ограничения на допустимое количество запросов от одного пользовательского аккаунта и/или IP-адреса в единицу времени, что требует учёта количества посылаемых запросов, а также поддержки динамической ротации используемых для сбора данных пользовательских аккаунтов и IP-адресов;

4. размерность данных обуславливает необходимость в параллельном методе сбора данных, а также в методах получения репрезентативной выборки пользователей социальной сети (сэмплирование).

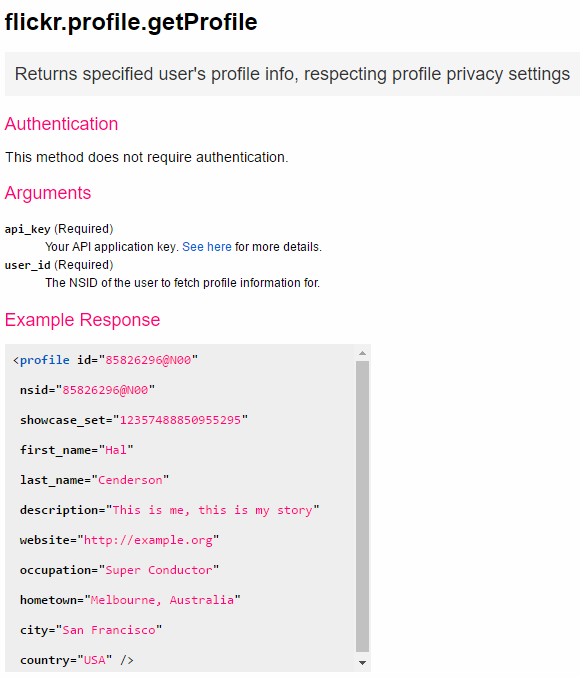
В связи с постоянной необходимостью получения больших наборов данных из социальных сетей, требуется разработка решения для сбора данных из различных интернет-сервисов.

Описываемые ниже технологии работы с социальными сетями основанына опыте по извлечению данных из сетей ВКонтакте и Facebook, однако, по-видимому, они могут быть, с соответствующими изменениями применены и для других социальных сетей.

Средством формирования списка необходимых профилей выступает обычный механизм поиска пользователей, имеющийся в социальных сетях. Обычно такие механизмы позволяют устанавливать фильтры (ограничивая отбираемые профили по полу, возрасту, месту проживания, местам учебы), после чего выдается список профилей пользователей, удовлетворяющих критериям запроса. В сети ВКонтакте фильтры представляют собой набор заполняемых полей, а в сети Facebook изначально пользователю предлагается ввести поисковый запрос на естественном языке (например, «люди, которые учились в МГУ»), который затем может быть дополнен фильтрами, заданным с помощью заполняемых полей. В обеих сетях список результатов имеет фиксированный размер (порядка тысячи профилей), и если число результатов превышает этот порог, какие-то профили не будут показаны. В связи с этим необходимо формировать поисковые запросы так, чтобы количество подпадающих под них профилей было заведомо небольшим.

При использовании данных ВКонтакте хорошей практикой оказалась выборка одновременно по возрасту и школе обучения - для каждого года рождения в конкретной школе учится лишь несколько десятков человек. При экспериментах с сетью Facebook, однако, использовать такие запросы оказалось гораздо менее удобно, потому что пользователи Facebook слишком часто не указывают свою дату рождения в профиле. Примечательно, что поисковая программа сети ВКонтакте имеет доступ в том числе и к данным, которые пользователь не делал публичными: в выборку двадцатилетних попадают и те профили, где дата рождения скрыта. На уровне браузера поисковые запросы транслируются в HTTP-запросы с достаточно простой структурой, что создает предпосылки для частичной автоматизации формирования таких запросов, и, как следствие, облегчает процесс формирования списка профилей.

После того, как получен список профилей пользователей, можно переходить к выгрузке профилейных данных. Это можно осуществить, как минимум, двумя способами. Во-первых, каждая профилейа — это просто веб-страница, которая может быть сохранена в файл. Во-вторых, социальные сети часто предоставляют программные интерфейсы для доступа создания приложений, работающих в рамках социальной сети (API - Application Programming Interface), которые также можно использовать для получения необходимых данных. Поскольку основное назначение этих интерфейсов иное, этот способ часто приходится комбинировать с предыдущим, однако, он может быть предпочтительнее, так как облегчает последующий разбор профилей.

Оба способа технически несложно реализуются и сводятся к отправке на сайт социальной сети последовательности HTTP-запросов. Единственный нюанс, который надо учитывать при автоматизации - это ограничение на число запросов в единицу времени. При слишком частых обращениях к сайту, действия могут быть восприняты как атака на сайт, что приведет к блокировке доступа. Данные, полученные через API социальной сети (пример API для получения профиля пользователя сети Flickr изображен на рисунке 1), заведомо структурированы, что существенно упрощает их обработку, хотя и в этом случае она может понадобиться. Так, например, населенные пункты из профили могут быть представлены числовыми идентификаторами, которые потребуют дополнительной «расшифровки». Для данных, полученных в виде HTML-страниц, неизбежно потребуется программная обработка, извлекающая из содержимого страницы значения интересующих полей профили (или другую имеющуюся там информацию). При работе с данными ВКонтакте для этого использовались стандартные средства POSIX-среды: программа текстового поиска и фильтрации grep, потоковый редактор sed и язык awk. Результаты извлечения данных из профилей собираются в таблицы, в которых каждая профилейа представлена строчкой, а поля профили — столбцами. К таким таблицам легко применять агрегирующие операторы и делать выборки (например, средствами SQL). При исследовании взаимосвязей пользователей социальной сети предпочтительной может оказаться графовая модель хранения, как более соответствующая внутренней структуре данных.

* 1. **Анализ временных рядов**

Временноой ряд – это собранный в разные моменты времени статистический материал о значении каких-либо параметров (в простейшем случае одного) исследуемого процесса. Каждая единица статистического материала называется измерением или отсчётом, также допустимо называть его уровнем на указанный с ним момент времени. Во временном ряде для каждого отсчёта должно быть указано время измерения или номер измерения по порядку. Временной ряд существенно отличается от простой выборки данных, так как при анализе учитывается взаимосвязь измерений со временем, а не только статистическое разнообразие и статистические характеристики выборки. В зависимости от наличия основной тенденции выделяют стационарные ряды, в которых среднее значение и дисперсия постоянны, и нестационарные, содержащие основную тенденцию развития[3].

В отличие от анализа случайных выборок, анализ временных рядов основывается на предположении, что последовательные значения в файле данных наблюдаются через равные промежутки времени (тогда как в других методах нам не важна и часто не интересна привязка наблюдений ко времени) [4].

Существуют две основные цели анализа временных рядов: определение природы ряда и прогнозирование (предсказание будущих значений временного ряда по настоящим и прошлым значениям). Обе эти цели требуют, чтобы модель ряда была идентифицирована и, более или менее, формально описана. Как только модель определена, с её помощью можно интерпретировать рассматриваемые данные (например, использовать для понимания сезонного изменения цен на товары). Не обращая внимания на глубину понимания и справедливость теории, можно экстраполировать затем ряд на основе найденной модели, т.е. предсказать его будущие значения.

4.1. Анализ тренда.

Не существует "автоматического" способа обнаружения тренда в временном ряде. Однако если тренд является монотонным (устойчиво возрастает или устойчиво убывает), то анализировать такой ряд обычно нетрудно. Если временные ряды содержат значительную ошибку, то первым шагом выделения тренда является сглаживание.

Сглаживание. Сглаживание всегда включает некоторый способ локального усреднения данных, при котором несистематические компоненты взаимно погашают друг друга. Самый общий метод сглаживания - скользящее среднее, в котором каждый член ряда заменяется простым или взвешенным средним n соседних членов, где n - ширина "окна". Вместо среднего можно использовать медиану значений, попавших в окно. Основное преимущество медианного сглаживания, в сравнении со сглаживанием скользящим средним, состоит в том, что результаты становятся более устойчивыми к выбросам (имеющимся внутри окна). Таким образом, если в данных имеются выбросы (связанные, например, с ошибками измерений), то сглаживание медианой обычно приводит к более гладким или, по крайней мере, более "надежным" кривым, по сравнению со скользящим средним с тем же самым окном. Основной недостаток медианного сглаживания в том, что при отсутствии явных выбросов, он приводит к более "зубчатым" кривым (чем сглаживание скользящим средним) и не позволяет использовать веса.

Относительно реже, когда ошибка измерения очень большая, используется метод сглаживания методом наименьших квадратов, взвешенных относительно расстояния или метод отрицательного экспоненциально взвешенного сглаживания. Все эти методы отфильтровывают шум и преобразуют данные в относительно гладкую кривую. Ряды с относительно небольшим количеством наблюдений и систематическим расположением точек могут быть сглажены с помощью бикубических сплайнов.

Подгонка функции. Многие монотонные временные ряды можно хорошо приблизить линейной функцией. Если же имеется явная монотонная нелинейная компонента, то данные вначале следует преобразовать, чтобы устранить нелинейность. Обычно для этого используют логарифмическое, экспоненциальное или (менее часто) полиномиальное преобразование данных.

4.2. Анализ сезонности

Периодическая и сезонная зависимость (сезонность) представляет собой другой общий тип компонент временного ряда. Это понятие было проиллюстрировано ранее на примере авиаперевозок пассажиров. Можно легко видеть, что каждое наблюдение очень похоже на соседнее; дополнительно, имеется повторяющаяся сезонная составляющая, это означает, что каждое наблюдение также похоже на наблюдение, имевшееся в том же самом месяце год назад. В общем, периодическая зависимость может быть формально определена как корреляционная зависимость порядка k между каждым -м элементом ряда и -м элементом. Ее можно измерить с помощью автокорреляции (т.е. корреляции между самими членами ряда); k обычно называют лагом (иногда используют эквивалентные термины: сдвиг, запаздывание). Если ошибка измерения не слишком большая, то сезонность можно определить визуально, рассматривая поведение членов ряда через каждые k временных единиц.

Автокорреляционная коррелограмма. Сезонные составляющие временного ряда могут быть найдены с помощью коррелограммы. Коррелограмма (автокоррелограмма) показывает численно и графически автокорреляционную функцию (AКФ), иными словами коэффициенты автокорреляции (и их стандартные ошибки) для последовательности лагов из определенного диапазона (например, от 1 до 30). На коррелограмме обычно отмечается диапазон в размере двух стандартных ошибок на каждом лаге, однако обычно величина автокорреляции более интересна, чем ее надежность, потому что интерес в основном представляют очень сильные (а, следовательно, высоко значимые) автокорреляции.

Исследование коррелограмм. При изучении коррелограмм следует помнить, что автокорреляции последовательных лагов формально зависимы между собой. Рассмотрим следующий пример. Если первый член ряда тесно связан со вторым, а второй с третьим, то первый элемент должен также каким-то образом зависеть от третьего и т.д. Это приводит к тому, что периодическая зависимость может существенно измениться после удаления автокорреляций первого порядка, т.е. после взятия разности с лагом 1.

Частные автокорреляции. Другой полезный метод исследования периодичности состоит в исследовании частной автокорреляционной функции (ЧАКФ), представляющей собой углубление понятия обычной автокорреляционной функции. В ЧАКФ устраняется зависимость между промежуточными наблюдениями (наблюдениями внутри лага). Другими словами, частная автокорреляция на данном лаге аналогична обычной автокорреляции, за исключением того, что при вычислении из нее удаляется влияние автокорреляций с меньшими лагами. На лаге 1 (когда нет промежуточных элементов внутри лага), частная автокорреляция равна, очевидно, обычной автокорреляции. На самом деле, частная автокорреляция дает более "чистую" картину периодических зависимостей.

Удаление периодической зависимости. Как отмечалось выше, периодическая составляющая для данного лага k может быть удалена взятием разности соответствующего порядка. Это означает, что из каждого -го элемента ряда вычитается -й элемент. Таким образом можно определить скрытые периодические составляющие ряда.

**Глава 3. РЕЗУЛЬТАТЫ**

**3.1 Сбор данных**

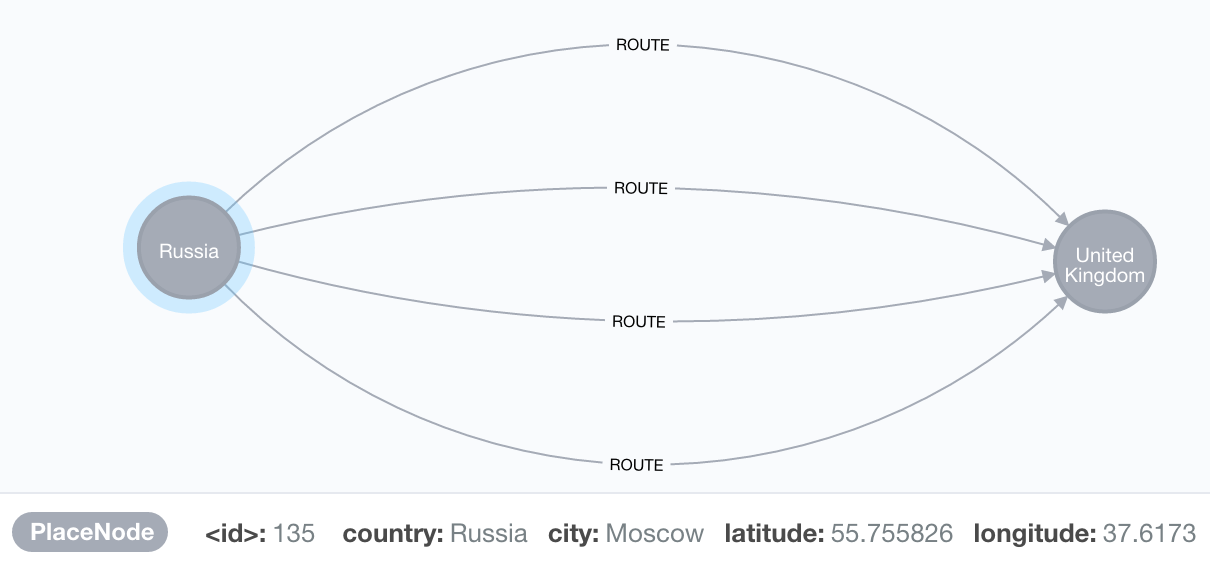
В качестве источника начальных данных для предиктивного анализа была выбрана социальная сеть Flickr, содержащая фотографии 122 миллионов пользователей более чем из 63 стран мира.

У Flickr есть общедоступный программный интерфейс, позволяющий искать фотографии по заданной геолокации, предоставляя информацию о времени и месте съемки, а также информацию о владельце фотографии, включая его привычное местоположение.

На основании этих данных можно судить о том, какие перемещения совершали пользователи сети и проводить их дальнейший анализ.

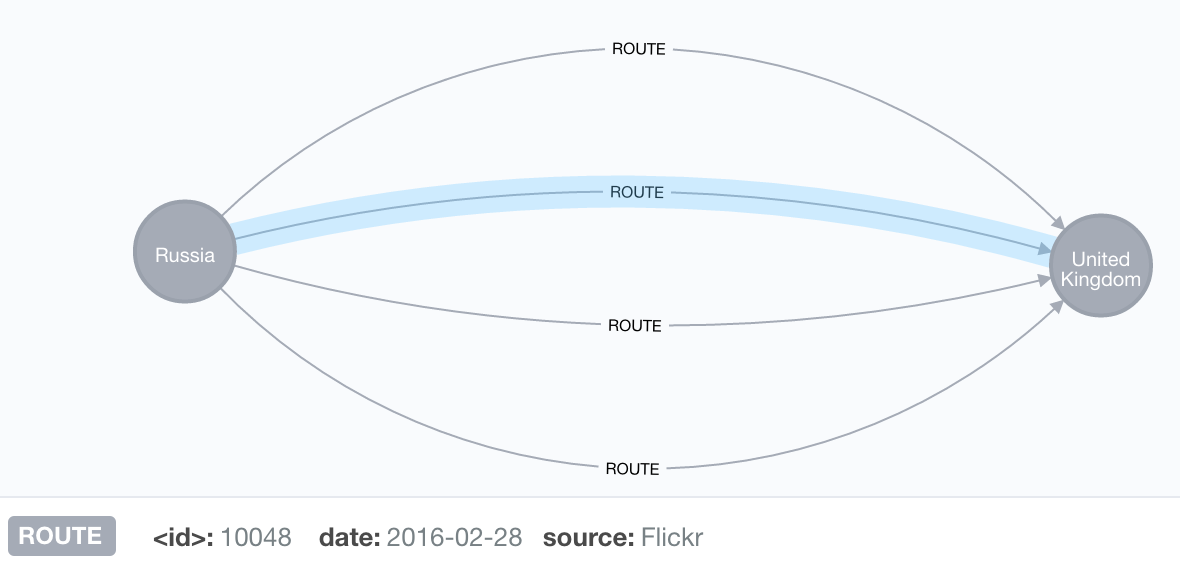
Для хранения маршрутов перемещения было выбрано их графовое представление, как наиболее естественное для такого типа данных.

В качестве базы данных было решено использовать NoSQL базу данных Neo4j, которая позволяет хранить данные в виде графа и выполнять над ним как простейшие операции, так и наиболее популярные алгоритмы поиска в графе.

Вершинами графа представлены города и соответствующие страны, дугами – маршруты с указанием времени. Пример визуализации графа для отладки изображен на рисунке 3.1 и 3.2.

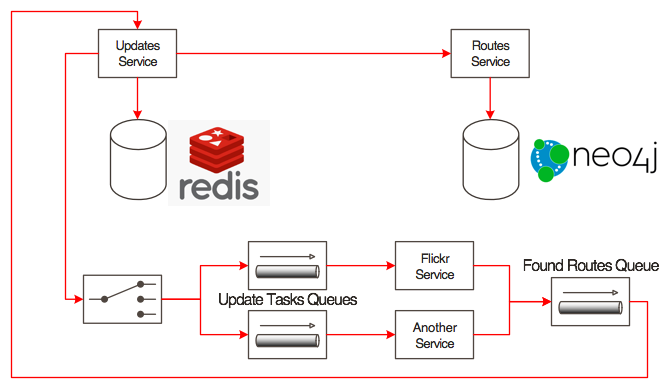
**Рисунок 3.1.**

**Рисунок 3.2.**

****

Для сбора данных было решено использовать микросервисную архитектуру, позволяющую масштабировать решение для использования других источников данных.

Архитектура решения для сбора данных изображена на рисунке 3.3.



**Рисунок 3.3.**

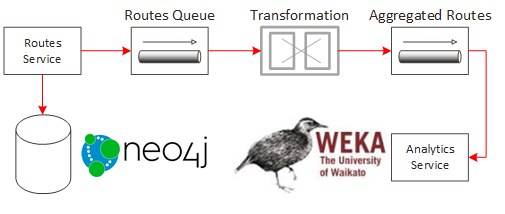
**3.2 Анализ данных**

Задача прогнозирования перспективных маршрутов пассажирских перевозок может быть эффективно решена с использованием методов теории анализа временных рядов.

В ходе исследования программных средств для анализа временных рядов были рассмотрены следующие программные пакеты: Weka, MacAnova, GMDH Shell, Zaitun, специализированные библиотеки языка программирования R.

Среди них был выбран Weka, так как в нем используется инновационный подход преобразования временных данных в форму, которая может быть подана на вход множеству хорошо изученных алгоритмов машинного обучения.

Данные о пользовательских перемещениях, хранящиеся в графовой базе данных, преобразуются в новый формат. Для каждой страны формируется свой файл, в который заносятся данные о количество посещений её жителями других стран в течение определенного периода времени с интервалом - день.

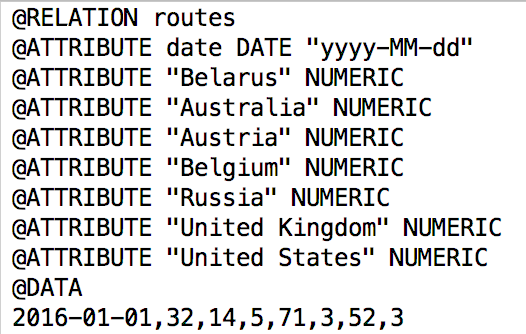
Архитектура модулей подготовки и анализа данных представлена на рисунке 3.4.

**Рисунок 3.4.**

Пример входных данных изображен на рисунке 3.5.

Далее было произведено построение модели временного ряда на основании данных о каждой стране за период 2016-2017 лет. На данном этапе были добавлены дополнительные временные признаки, такие как день недели, выходные, праздники, сезоны, и применены непосредственно методы машинного обучения.

Среди алгоритмов машинного обучения были использованы следующие:

* решающие деревья (Decision Tree);
* случайный лес решающих деревьев (Random Forest);
* SVM (Support Vector Regressor);
* гауссовские процессы (Gaussian Processes);

**Рисунок 3.5.**

Для оценки используемых алгоритмов данные были разделены на обучающую и тестовую выборку и был произведен анализ среднеквадратичного отклонения. Результаты изображены в таблице 1.

**Таблица 1.**

В итоге был выбран алгоритм, использующий метод опорных векторов как наиболее эффективный.

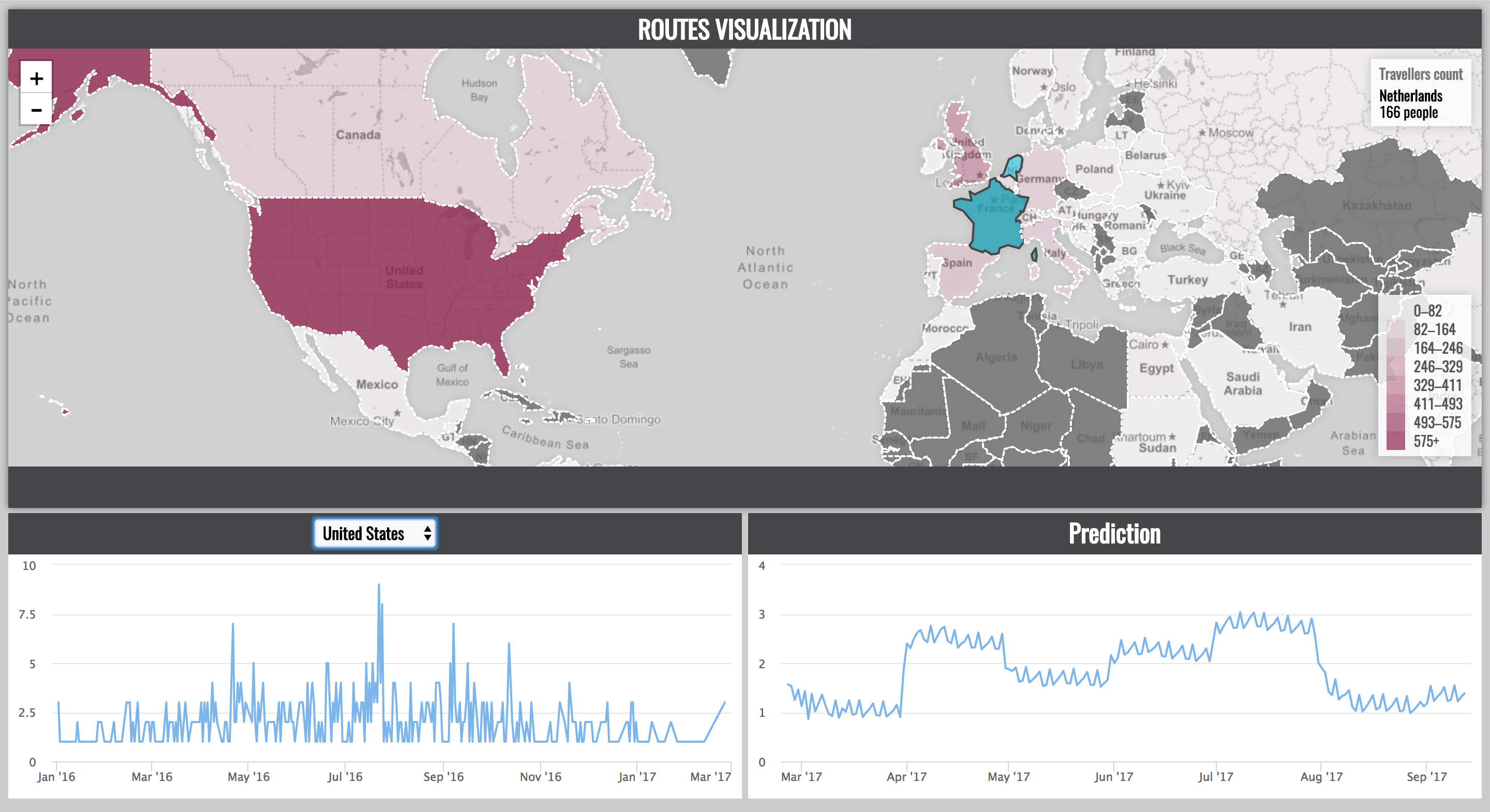
|  |  |
| --- | --- |
| Алгоритм | Среднеквадратичное отклонение |
| Decision Tree | 0.35 |
| Random Forest | 0.31 |
| SVM | 0.13 |
| Gaussian Processes | 0.27 |

**3.3 Визуализация результатов.**

Для визуализации результатов было решено использовать интерактивные карты для отображения количества собранных данных из различных источников, а также графики для изображения временных рядов соответствующих имеющимся данным и предсказанным результатам.

В качестве библиотеки для создания интерактивных карт был использован leaflet.js, имеющий интуитивный интерфейс, механизм расширений и поддержку стандарта хранения геоданных GeoJSON.

Для изображения графиков была использована библиотека highcharts.js предоставляющая различные форматы изображения временных рядов, а также возможность интерактивного масштабирования графика.

Результат визуализации изображен на рисунке 3.6.

**Рисунок 3.6.**

На карте различными цветами изображены проанализированные страны. Выбранная страна окрашивается в голубой цвет, страны, жители которых посетили выбранную страну - в оттенки бордового цвета, чем больше посетителей - тем темнее цвет. При наведении на любую страну, кроме выбранной, показывается, сколько её жителей посетили выбранную страну.

На графике в левом нижнем углу изображены собранные данные о перемещениях жителей каждой из проанализированных стран, в правом нижнем углу - предсказание.

**Глава 3. РЕКОМЕНДАЦИИ**

В результате был разработан программный продукт, представляющий собой аналитическую панель, которая может быть использована как инструмент маркетологами в сфере пассажирских перевозок и логистами.

В ходе исследований были выявлены следующие возможности улучшения полученных результатов:

1. Рассмотрены данные из одной социальной сети - Flickr, в которой содержится сравнительно мало информации о пользователях из стран СНГ. Необходимо использовать данные из других социальных сетей (Facebook, Instagram), в том числе популярных в Беларуси и России (Вконтакте, Одноклассники), и включить их в построенную систему.

2. В качестве нового источника данных исследовать туристические порталы, такие как TripAdvisor. Собрать большое количество данных и оценить на них построенную модель.

3. Включить в исследование методы opinion mining для поиска взаимосвязи между количествами посещений определенных стран и мнениями пользователей социальных сетей о них.

4. Для анализа были использованы методы машинного обучения. Необходимо рассмотреть также классические методы анализа временных рядов и сравнить их результаты с уже использованными методами.

**БИБЛИОГРАФИЧЕСКИЙ СПИСОК**

1. Юркевич М., Казунина А., Анализ рынка пассажирских перевозок в дальнем сообщении, 2013.
2. Gjoka M. et al. Practical recommendations on crawling online social networks //Selected Areas in Communications, IEEE Journal on. – 2011. – Т. 29. – №. 9. – С. 1872-1892.
3. Шмойлова Р. А. Общая теория статистики: Учебник. — М.: Финансы и статистика, 2002. — ISBN 5-279-01951-8.
4. Shumway Robert H., Time series analysis and its applications, 2000.
5. Bogorov V., Novikov A., Serova E. Self-exploration of the City <URL:http://issuu.com/mosurbanforum/docs/eng-uf_354-373_data__tm>
6. 1
7. 1
8. Granovetter M. S. The Strength of Weak Ties // The American Journal of Sociology. — 1973. — 78 (6). — Pp. 1360–1380.