МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение
высшего образования

**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»**

Кафедра Математического и компьютерного моделирования

Применение технологий BigData при выборе бизнес-проектов

АВТОРЕФЕРАТ МАГИСТЕРСКОЙ РАБОТЫ

студента 2 курса 247 группы

направление 09.04.03 — Прикладная информатика

механико-математического факультета

Левого Ивана Валерьевича

Научный руководитель
проф., д.э.н Л.В. Кальянов

Зав. кафедрой
зав. каф., д.ф.-м.н., доцент Ю.А. Блинков

Саратов 2020

Введение. Настоящая работа посвящена применению технологии Big Data с целью выбора бизнес-проектов. Актуальность этой темы обусловлена огромными объемами информации, которые накапливаются человечеством, что несёт как позитивные, так и негативные последствия. С одной стороны, большой объем данных о различных сферах жизнедеятельности человека открывает перспективы извлечения из них полезной и ценной информации, которая явно может даже не содержаться в данных. С другой - информационная «зашумленность» создает препятствия для улавливания ключевой и ценной информации, которая может оставаться незамеченной.

В условиях современного мира возникла потребность в методе управления - управления слабыми сигналами, который предполагает, что руководство уже при первых слабых сигналах об изменениях на рынке не ждет, а начинает действовать. При этом конкретные шаги со стороны управления, которые направлены на использование пока еще нечетко проявившихся рыночных возможностей или угроз, становятся все более определенными, увеличивают своё влияние по мере поступления более точной и объёмной информации об изменении ситуации во внешней среде. К тому моменту, когда слабые сигналы набирают свою силу до очевидных для всех, компания уже успевает занять лидирующую позицию, что предоставляет ей преимущество в удержании позиции на рынке.

В условиях уже упомянутой информационной перегруженности о применении подхода управления слабыми сигналами возможно говорить только с практическим применением Big data в бизнесе.

Среди применений анализа информационного потока на сегодняшний день одними из самых известных примеров является мониторинг упоминаний каких-либо объектов и анализ отношения к ним в СМИ и социальных сетях. Очевидно, что Интернет ежедневно генерирует огромное количество текстовой информации — новости, пресс-релиза, блоги, социальные сети. Поэтому даже сама задача сбора и хранения такой информации, не говоря уже о переработке является отдельным бизнесом. Отдельную ценность представляют именно методы обработки текстов, в особенности мониторинг упоминаний в текстах отношения к конкретному объекту или компании.

Целью работы является изучение инструментов BigData с целью выбора бизнес-проектов. Задачей является выявление класса новостей, несущих значимую информацию, но теряющихся в информационном шуме.

Основная часть. Основная часть работы состоит из 4 разделов:

1. Постановка проблемы управления слабыми сигналами
2. Описание технологий BigData
3. Анализ и выбор программного обеспечения для решения поставленной в рамках работы задачи
4. Разработка модели выбора бизнес-проектов с помощью технологий BigData

Управление слабыми сигналами подразумевает определение постепенных шагов при планировании и в последующих действиях, которые могут быть реализованы при различных вариантах развития событий, вместо того, чтобы ожидать поступления конкретной и точной информации.

На начальном этапе проявления потенциальной возможности, когда информация не определена в достаточной мере, ответные меры (действия) могут иметь общий характер. По мере того, как поступает более точная и конкретная информация, ответные меры так же должны конкретизироваться. Итогом этих действий должно стать использование возможности, либо устранение опасности. Такое последовательное постепенное усиление ответных мер на трудно предсказуемые события можно определить как управление слабыми сигналами.

Сильные и слабые сигналы есть две крайние точки осведомленности, которая отражает полноту и достоверность доступной информации об определенной ситуации.

Выделяют пять уровней осведомленности по мере возрастания силы их сигнала:

- Первый уровень осведомленности соответствует наименьшему объему значимой и полезной информации. Известно лишь то, что не исключено возникновение какой-либо опасности (возможности), природа и источник которой на данный момент неизвестны.

- Второй уровень осведомленности — ситуация, когда источник возможных новых явлений известен, а сами явления — нет. Например, научные исследования в области информационных технологий.
- При осведомленности третьего уровня известен источник и сами явления, но есть неясности в области их применения и, соответственно, в необходимых мерах реагирования со стороны компании. На этом уровне для надежной оценки воздействия и эффективности ответных мер информации недостаточно. Например, после того как стали известны свойства новой информационной технологии, область применения понятна не до конца.
- Четвертый уровень осведомленности: информации достаточно для разработки и принятия конкретных мер, но из-за отсутствия опыта возможные последствия определить сложно. Например, компании, которые первыми использовавшие оптоволоконную технологию в телекоммуникационной сфере, инвестировали крупные средства в эту технологию с надеждой, что с допустимым в таких случаях риском их вложения окупятся.
- Осведомленность пятого уровня заключается в том, что получены данные о результативности принятых мер. Например, фирмы-первопроходцы получили достаточные данные, чтобы определить прибыльность новой технологии. Те, кто не попал в их число, должны были нести большие расходы для того, чтобы внедриться в эту область производства.

По мере того, как сигналы из внешней среды набирают силу, открывается больше возможностей для принятия активных действий. Принимаемые меры по интенсивности должны быть адекватны текущим уровням осведомленности.

Резюмируя, можно сформулировать преимущества фирмы при управлении по слабым сигналам, которые позволяют:

- заблаговременно узнавать о резких изменениях во внешней среде
- своевременно реагировать на трудно предсказуемые события
- на ранней стадии появления потенциальных опасности и возможности принять конкретные ответные меры, конечной целью которых станет

либо устранение опасности, либо использование так называемых стратегических окон - создавшихся возможностей

Большие данные (от английского big data) — обозначение структурированных и неструктурированных данных огромных объёмов и значительного многообразия, эффективно обрабатываемых горизонтально масштабируемыми программными инструментами, появившимися в конце 2000-х годов и альтернативных традиционным системам управления базами данных и решениям класса Business Intelligence.

Совокупность подходов и инструментов для обработки больших данных, с точки зрения информационных технологий, включала средства параллельной обработки неопределённо структурированных данных: системы управления базами данных категории NoSQL, алгоритмы MapReduce и реализующими их проектами - Hadoop. С течением времени к серии технологий Big Data стали относить разнообразные ИТ-решения, обеспечивающие сходные по характеристикам возможности по обработке сверхбольших массивов данных.

Преимущества, которые предоставляет Big Data:

- Сбор данных из разных источников
- Улучшение бизнес-процессов через аналитику в реальном времени
- Хранение огромного объема данных
- Big Data более проницательна к скрытой информации при помощи структурированных и полуструктурированных данных
- Большие данные помогают уменьшать риск и принимать умные решения благодаря подходящей риск-аналитике

Формы больших данных:

- Структурированная. Структурированными данными называются данные, которые могут храниться, быть в любой момент доступными и могут обрабатываться в форме с фиксированным форматом.
- Неструктурированная - данные неизвестной структуры. Ярким примером неструктурированных данных является неоднородный источник, содержащий комбинацию документов, изображений, аудио и видео.
- Полуструктурированная. Эта категория содержит комбинацию описанных выше форм. Примером этой категории могут служить персональные данные, представленные в XML файле.

Большие данные различаются по объему, скорости генерации, разнообразию и изменчивости.

В настоящее время существует большое разнообразие методик анализа массивов данных, лежащий в основе инструментарий которых заимствован из статистики и информатики.

Отдельное внимание уделено технологии Text Mining, которая является одной из разновидностей методов Data Mining, суть процессов которой заключается в извлечении из текстовых массивов знаний и значимой информации. Извлечение знаний становится возможным благодаря выявлению тенденций и шаблонов с помощью средств статистического изучения паттернов.

Text Mining – выявление и обнаружение в необработанных данных с помощью определенных алгоритмов ранее неизвестных взаимосвязей, до момента анализа неизвестных практически полезных знаний, с целью их интерпретации и использования для принятия решений в различных сферах человеческой деятельности. Эти результаты используются для многих целей, например: математического прогнозирования, маркетингового анализа, изучения рынков и так далее.

Фактически, анализ текстов и Text Mining – это набор лингвистических, статистических техник, а также техник машинного самообучения, которые способны моделировать и структурировать информационный контент и текстовые источники в целях бизнес-аналитики, анализа данных, исследований. Отдельно или совместно с другими средствами, технологии анализа текстов и Text Mining широко используются в повседневной практике с целью управления знаниями для решения разного рода бизнес-задач. По статистике, порядка 80% важной и ключевой для бизнеса информации существует в неструктурированной текстовой форме. Описанные выше технологии позволяют извлекать из данных ценные знания, которые невозможно получить какими-либо иными автоматизированными средствами: факты, бизнес-правила, взаимосвязи и так далее.

Неизмеримое множество знаний и информации содержится в Интернете. Подобное обилие часто создает сложности при поиске необходимой информации. Подобного рода проблемы могут иметь различный характер, например:

1. Для пользователя не всегда представляется возможным быстро найти необходимые ему источники информации, так как не все ссылки ведут туда, куда они указывают. Более того, информацию, которая не проиндексирована поисковыми системами таким способом найти и вовсе не представляется возможным.
2. Имея большое количество информации, пользователь часто испытывает сложности с тем, чтобы извлечь из нее полезные знания и интерпретировать их.
3. В случае с изучением информации о потребителях, возникает необходимость предоставлять интересные конкретно им сведения. Например, предоставлять пользователю рекомендации по выбору нужного товара.

Исходя из вышесказанного, появилась необходимость разработки специальных технологий с целью извлечения полезных знаний из сети Интернет. Именно этим обусловлено появление технологии Web Mining.

Выделяют 3 разновидности технологии Web Mining в зависимости от типа решаемых задач: анализ использования веб-ресурсов, извлечение веб-структур, извлечение веб-контента.

Под анализом использования веб-ресурсов подразумевается извлечение данных из логов веб-серверов тех или иных ресурсов сети Интернет с целью лучшего понимания предпочтений пользователей.

Извлечение веб структур – анализирует взаимосвязи между веб-страницами, с помощью анализа связи между ними.

Извлечение веб-контента – анализ содержание электронных документов, путем нахождения схожих по смыслу слов и их количеств, с целью проведения классификации, либо кластеризации и сгруппировать документы по смысловой близости.

Для анализа информационного потока полезным будет понимать его эмоциональную окраску - эмоциональное отношение автора высказывания, выраженное в тексте, к некоторому объекту: объекту реального мира, событию, процессу, их свойствам, атрибутам.

Анализ эмоциональной окраски (Sentiment analysis) — класс методов контент-анализа в компьютерной лингвистике, предназначенный для автоматизированного выявления в текстах эмоционально окрашенной лексики и

эмоциональной оценки авторов (их мнений) по отношению к объектам, речь о которых идёт в тексте.

Цель анализа тональности - вычленение мнений автора в тексте и определение их эмоциональных свойств.

Стоит отметить, что эмоциональная окраска является лишь одной из характеристик мнения. Однако именно задача классификации тональности наиболее распространена на практике в настоящее время.

Говоря о технологиях BigData, нельзя не упомянуть технологии распределенных вычислений.

Hadoop — это революционная база данных для больших данных, которая способна сохранять любые формы данных и обрабатывать их в кластере узлов, благодаря чему существенно снижается стоимость обслуживания данных.

К основным особенностям Hadoop относятся:

- Масштабируемость. При помощи Hadoop становится возможным организовать надежное хранение и обработку больших объемов данных, которые могут до петабайт
- Экономичность. Благодаря тому, что информация и вычисления распределяются по кластерам, которые реализованы на рядовом оборудовании. В кластер могут входить тысячи узлов
- Эффективность. Это свойство достигается благодаря распределению данных и их параллельную обработку на множестве компьютеров, что кратно ускоряет процесс обработки информации
- Надежность. Подход гарантирует сохранность информации в случае сбоев системы благодаря хранению нескольких копий на разных узлах
- Кроссплатформенность. Благодаря тому, что основным языком программирования в системе является Java, имеется возможность реализовать решение на базе любой операционной системе с поддержкой Java Virtual Machine

Hadoop имеет распределенное хранилище, а также распределенную систему процессов, таких как MapReduce.

Сохранение данных в Hadoop возможно с помощью двух различных средств — HDFS и HBase.

Hadoop-кластер состоит из нод трех типов: NameNode, Secondary NameNode, Datanode.

Отличительные характеристики распределенной файловой системы Hadoop:

- Больший размер блока по сравнению с другими файловыми системами (более 64MB)
- ориентация на недорогие и, поэтому не самые надежные сервера
- зеркалирование и репликация осуществляются на уровне кластера, а не на уровне узлов данных
- репликация происходит в асинхронном режиме – информация распределяется по нескольким серверам прямо во время загрузки
- HDFS оптимизирована для потоковых считываний файлов
- клиенты могут считывать и писать файлы HDFS напрямую через программный интерфейс Java
- файлы пишутся однократно, что исключает внесение в них любых произвольных изменений
- принцип Write-once and read-many: один раз записать – много раз прочитать
- HDFS оптимизирована под потоковую передачу данных
- сжатие данных и рациональное использование дискового пространства позволило снизить нагрузку на каналы передачи данных
- самодиагностика — каждый узел данных через определенные интервалы времени отправляет диагностические сообщения узлу имен, который записывает логи операций над файлами в специальный журнал
- все метаданные сервера имен хранятся в оперативной памяти

Были проанализированы основные программные комлпексы, подходящие под цели работы:

1. RapidMiner — это бесплатная open-source среда для прогнозной аналитики. Функционал RapidMiner может быть расширен при помощи дополнений, часть из которых также доступна бесплатно. RapidMiner поддерживает результирующую визуализацию, проверку и оптимизацию.

2. Платформа IBM SPPS Modeler — коммерческий конкурент RapidMiner, который характеризуется низким порогом входа для начинающих. Понятность и доступность для новичков обеспечивается режимами «автопилота». Авто-модели (Auto Numeric, Auto Classifier) перебирают несколько возможных моделей с разными параметрами, определяя среди них те, которые дают лучший результат. Благодаря этому начинающий и неопытный аналитик может построить на таком решении достаточно адекватную модель.
3. KNIME — свободно распространяемая система для решений в области интеллектуального анализа данных, обладающая достаточным функционалом уже в младшей версии. Аналогично с RapidMiner, KNIME предлагает интуитивно понятную рабочую среду без необходимости использовать программный код. Здесь также есть ряд операторов, аналогичных RapidMiner.
4. Компания Qlik - лидер на рынке визуальной аналитики, чья платформа для анализа данных позволяет осуществлять разработку как пользовательских приложений, так и приложений под заказчика. Qlik Analytics Platform обладает всеми необходимыми инструментами разработки для управления данными.
5. STATISTICA Data Miner является платформой российской разработки. В части функционала программное решение предоставляет наиболее полный набор инструментов и методов для проведения Data Mining анализа. В частности, в STATISTICA Data Miner имеются инструменты для предварительной обработки, фильтрации и чистки данных.
6. Informatica Intelligent Data Platform предоставляет интеллектуальные и управляющие сервисы, способные работать с большим количеством популярных форматами данных, среди которых: web, social network, машинные журналы и другие.
7. World Programming System на рынке занимает позицию основного конкурента программных продуктов компании SAS. Более того, World Programming System поддерживает написанные на языке SAS решения. Поддерживаемый синтаксис актуальной версии WPS охватывает ядро,

статистические и графические возможности приложений, созданных с помощью языка SAS.

8. Deductor — аналитическая платформа, разработанная компанией BaseGroup Labs, обладает самыми необходимыми алгоритмами анализа: деревья решений, нейронные сети, самоорганизующиеся карты, а также многочисленными способами визуализации. Помимо прочего реализована интеграция со многими источниками данных и приемников.
9. SAS Enterprise Miner — это программный продукт, который на основании больших объемов информации способен создавать точные предсказательные и описательные модели. SAS Enterprise Miner в большей степени ориентирован на бизнес-аналитиков. Среди возможных вариантов использования программного продукта наиболее распространены: оценка и минимизация рисков, выявление мошенничества, снижение оттока клиентов.

Исходя из проведенного анализа и имеющихся ограничений на использование программного обеспечения для разработки модели в рамках настоящей работы был выбран программный комплекс Rapidminer.

С целью реализации прикладной модели анализа информации из сети Интернет была построена модель анализа эмоциональной окраски текста сообщений пользовательской сети Twitter. Для поиска «твитов» использовался сайт https://www.trendsmap.com.

Модель находит «твиты» пользователей с нужным хэштегом, который определяется в параметрах. Затем данные с помощью оператора Write Excel записываются в таблицу.

Полученные данные с помощью оператора Analyze Sentiment проходят процедуру оценки эмоциональной окраски текста и его объективности. На выходе получается excel-файл. Затем полученные результаты кластеризуются с помощью метода k-средних.

Результаты кластеризации продемонстрировали, что наибольшее число «твитов» попало в нулевой кластер - сообщения с позитивной эмоциональной окраской, а в остальные 4 кластера попали по одному «твиту» - сообщения с негативной окраской. С полученными данными желаемого результата - информацию, со значительными последствиями, неожиданную для эксперта,

имеющую рационалистическое объяснение в ретроспективе - обнаружить не удалось.

Сообщения пользователей Twitter имеют ряд преимуществ, например, быстрый отклик на события. Но вместе с тем, как оказалось, не всегда удаётся получить качественные и значимые данные. С целью развития модели в качестве данных были проанализированы RSS каналы крупных зарубежных новостных агентств.

В качестве анализируемого текста использовался RSS канал новостного агентства Russia Today, с целью выявить значимые новости среди общего шума.

Краткое резюме результатов кластеризации:

1. Cluster 0 - содержит новости UFC (Ultimate Fighting Championship)
2. Cluster 1 - спортивные новости: футбол, спортивная стрельба
3. Cluster 2 - политические новости
4. Cluster 3 - новости про Covid19
5. Cluster 4 - новости ко Дню Победы в России

По результатам проведенного анализа видно, что модель достаточно точно кластеризует новостной поток, но для качественного анализа RSS-канал даёт слишком мало наблюдений - 100 новостных статей.

В качестве набора данных для дальнейшего анализа был выбран архив новостей с февраля по август 2017 года, данные взяты с сайта kaggle.com.

Проанализировав содержание каждого кластера становится понятно, что для целей работы наиболее интересен самый многочисленный кластер - Cluster 11, содержащий 1708 наблюдений (информационных статей). Это класс, касающийся экономических новостей. Остальные кластеры в основном содержат информацию про спорт, политику и происшествия.

Дальше кажется целесообразным декомпозировать Cluster 11 на дополнительный уровень с целью наиболее детального анализа содержащейся в нем информации.

Родительский кластер - массив всех экономических новостей, полученный на предыдущем этапе кластеризации всего архива новостных статей.

Полученные дочерние кластеры - экономические новости, но уже разбитые по тематикам, благодаря чему можно вычленить значимую и интересу-

ющую нас информацию и попытаться уловить слабые сигналы, которые не заметны во всём массиве.

Коротко резюмируя смысловое содержание значимых кластеров и отсеяв информационный шум, можно выделить основные темы:

- Сотрудничество РЖД и Министерства железных дорог Индии по программе «Скоростные железные дороги». В рамках программы планируется построить и электрифицировать 400 км железных дорог.
- Реформа правительством налога на добавленную стоимость
- Переговоры о создании зоны свободной торговли между ЕАЭС и Индией. Из 138 мер, действующих в мире в отношении товаров из Евразийского экономического союза, 13 приходится на Индию.

Таким образом, исходя из проведенного анализа можно сделать вывод, что в связи со строительством железных дорог могут возникнуть возможности организации бизнеса в транспортной логистике, ЖД перевозках, аренде и обслуживании железнодорожных вагонов. Это и есть слабый сигнал, который неуловим в информационном потоке.

Заключение. В магистерской работе был поставлен вопрос необходимости управления слабыми сигналами в современном мире, изучены основы BigData, в частности: методики анализа больших данных BigData, Text Mining, анализ эмоциональной окраски текста - Sentiment Analysis, а так же основы распределенных вычислений Hadoop. На практике применены инструменты Text Mining, проанализировано существующее ПО для анализа текстовой информации. В практической части построены модели анализа информации из сети Интернет с целью принятия бизнес-решений:

1. Модель анализа сообщений пользователей Twitter
2. Модель анализа новостей из RSS каналов
3. Модель анализа архива новостей с многоуровневой декомпозицией

Проведенный анализ и построенная модель позволили на реальных данных проверить наличие скрытых бизнес-возможностей среди «информационного шума».

Проделанная работа позволила применить и закрепить полученные теоретические знания на практике.