Министерство образования Республики Беларусь
Учреждение образования
Белорусский государственный университет
информатики и радиоэлектроники

УДК_____

Якубович
Федор Владимирович

ИСПОЛЬЗОВАНИЕ BIG DATA ДЛЯ СБОРА И АНАЛИЗА ИНФОРМАЦИИ ИЗ СОЦИАЛЬНЫХ СЕТЕЙ

АВТОРЕФЕРАТ
на соискание академической степени магистра информатики и вычислительной техники

по специальности 1-40 81 04 – Обработка больших объемов информации

_____

Научный руководитель
Волорова Н. А.
к. т. н., доцент

_____

Минск 2017

КРАТКОЕ ВВЕДЕНИЕ

Увеличение пользования социальными сетями, такими как Facebook, Twitter, Instagram производило и производит огромный объем данных. Коммерческие фирмы и другие организации заинтересованы в открытии новых бизнес кругов и в повышении эффективности бизнеса.

С помощью использования углубленной аналитики, предприятия могут анализировать большие объемы данных, чтобы узнать о связях лежащих в основе социальных сетей, которое характеризует социальное поведение отдельных лиц и групп. Используя полученные данные, описывающие отношения, можно определять общественных лидеров, которые влияют на поведение других людей в сети, и, с другой стороны, для определения людей, которые наиболее подвержены воздействию других участников сети.

В широком смысле, понятие большие объемы данных представляет собой совокупность методов, инструментов и подходов обработки структурированных и неструктурированных данных больших объемов [1]. Результатом обработки больших объемов данных являются данные, которые можно обработать стандартными средствами. С точки зрения информационных технологий к подходам и инструментам относятся средства массово-параллельной обработки неопределенно структурированных данных, прежде всего, системами управления базами данных категории NoSQL, алгоритмами MapReduce и реализующими их программными каркасами и библиотеками проекта Hadoop.

Сам термин «большие объемы данных» в компьютерных технологиях появился 3 сентября 2008 года, когда Клиффорд Линч подготовил статью, в которой были собраны материалы о феномене взрывного роста объемов и многообразия обрабатываемых данных и технологических перспективах в парадигме вероятного скачка «от количества к качеству» и проблемах, которые возникнут при обработки такого объема данных [2]. Уже в 2010 годах стали появляться первые продукты и решения, относящиеся непосредственно и исключительно к проблеме обработки больших данных. В 2011 году большие данные были отмечены как тренд номер два, после виртуализации.

В данный момент технологии больших данных наибольшее влияние оказали на информационные технологии в торговле, образовании, производстве, здравоохранении и государственном управлении. Также большие данные, как академический предмет изучаются в программах высших учебных заведений, как одно из наиболее перспективных направлений. Основными источниками больших объемов информации являются социальные медиа и интернет вещей. В качестве примеров источников возникновения больших данных приводятся непрерывно поступающие данные с измерительных устройств, события от радиочастотных идентификаторов, потоки сообщений из социальных сетей.

ОБЩАЯ ХАРАКТЕРИСТИКА РАБОТЫ

Актуальность темы магистерской диссертации

В настоящее время социальные сети играют большую роль в жизни каждого человека. Ежедневно люди получают, а также делятся большим количеством информации из социальных сетей, через свои социальные профили. Данная информация, а именно реакция людей на тот или иной продукт, представляет собой неограниченный источник знаний для маркетинга и коммерческого применения.

Задача процесса автоматизации анализа больших объемов информации из социальных сетей, требует автоматизации процесса сбора информации из различных источников. При этом информации может быть представлена в неформализованном, неструктурированном виде. Обработка такой информации является весьма сложной задачей, требующей использования алгоритмов, математических методов и современных компьютерных технологий.

Цель и задачи исследования

Целью диссертационной работы является автоматизация процесса анализа полученной информации из социальных сетей на предмет настроения.

Для достижения поставленной цели необходимо решить следующие задачи:

1. Проанализировать существующие решения, позволяющие анализировать социальные сети.
2. Определить их основные характеристики и ограничения.
3. Проанализировать средства, позволяющие собирать информацию из социальных сетей.
4. Провести анализ методов и алгоритмов, позволяющих анализировать текст на предмет настроения.

Задачами исследования являются изучение предметной области, ознакомление с подходами и технологиями обработки больших объемов неструктурированной информации, построение алгоритмов подбора и анализа, проектирование приложения, исследования алгоритмов анализа текста на предмет настроения.

Объектом исследования являются социальные сети, методы, технологии и алгоритмы обработки больших объемов информации, алгоритмы анализа настроений текста.

Предметом исследования являются алгоритмы анализа больших объемов данных.

Связь работы с приоритетными направлениями научных исследований и запросами реального сектора экономики

В связи с быстрым ростом популярности социальных сетей, количество информации в них стремительно увеличивается. Социальные сети служат для объединения людей, обеспечивая им возможность обмениваться информацией, делиться мнением, оставлять отзывы о тех или иных продуктах, выкладывать информацию о себе. Ежедневно в социальных сетях выкладывают информацию о тех или иных событиях, новинках, которые они приобрели или просто делятся своими впечатлениями от увиденного. Такая информация представляет собой важность, с той точки зрения, что ее можно анализировать на предмет настроения, с целью получения информации о реакции человека, его количественной оценки удовлетворенности о событии или продукте. Данная информация может быть применена в коммерческих целях, например для маркетинга, рекламы. Также можно повысить эффективность бизнеса, путем анализа мнений пользователей, которые высказались о продукте неудовлетворительно.

С учетом всего выше сказанного, в реальном секторе экономики существует реальная необходимость разработки средства, позволяющего анализировать социальные сети на предмет настроения.

Личный вклад соискателя

Результаты, приведенные в диссертации, получены соискателем лично. Вклад научного руководителя Н. А. Волоровой, заключается в формулировке целей и задач исследования.

Публикация результатов диссертации

По теме диссертации опубликовано 2 печатные работы, из них 2 работы в международном журнале «Наука, образование и культура»

Структура и объем диссертации

Диссертация состоит из общей характеристики работы, введения, трех глав, заключения, списка использованных источников. В первой главе представлен анализ предметной области, выявлены основные проблемы в рамках тематики исследования, показаны направления их решения. Также в данной

главе приведен анализ существующих инструментов, позволяющих выполнять анализ социальных сетей, а также дана их сравнительная характеристика. Вторая глава посвящена анализу существующих методов и алгоритмов, анализирующих настроение текста и emoji.

В третьей главе происходит описание реализации предложенных ранее методов и алгоритмов, приводятся результаты проведения испытаний реализованных методов. В данной главе также рассматривается реализация системы сбора и анализа при помощи различных инструментов и методов по обработке больших объемов информации.

Общий объем работы составляет 64 страницы, из которых основного текста – 45 страниц, 27 рисунков на 26 страницах, 1 таблицы на 1 страницах и список использованных источников из 30 наименований на 2 страницах.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

В первой главе диссертационной работы подробно рассмотрена предметная область. В данной главе было дано определение понятию Big Data, были рассмотрены основные характеристики Big Data, а также были рассмотрены основные принципы работы с большими объемами информации. Также в данной главе проводился анализ социальных сетей, которые использовались в работе, а именно Twitter и Instagram. Был проведен анализ аудитории данных социальных сетей и была дана сравнительная характеристика двух данных социальных сетей. Далее проводился анализ средств, похожих по своей направленности на данный проект. Были проанализированы следующие средства Hootsuite Insights, Twitter Advanced Search, Brandwatch, Semantria, Rapidminer. Из обзора можно сделать вывод, что большинство данных инструментов являются платными и не позволяют использовать все возможности в бесплатном режиме, а также данные сервисы являются сторонними и не предоставляют своего API, что делает их непригодными для интеграции в другие сервисы. В дополнении данные инструменты не анализируют настроения emoji. Также в данной главе было рассмотрено понятие социальных сетей, их суть, их процесс анализа и многое другое.

Во второй главе были рассмотрены алгоритмы анализа настроений текста и emoji. Для анализа настроений текста был выбран VADER, который используется для анализа чувств, чувствительности к полярности и интенсивности эмоций. Алгоритмы оценки emoji и текста практически одинаковы. Их основной задачей является составление словаря путем эмпирической проверки с использованием человека-оценщика. Для каждого алгоритма было выбрано определенное количество людей, являющихся носителями языка, которые проводили оценку настроений каждого слова или emoji в отдельности. Далее были составлены свои словари, в которые были записаны значения для каждого слова или emoji.

В третьей главе были рассмотрены инструменты, которые применялись при реализации данного проекта. Была приведена архитектура приложения, на которой были показаны все основные структурные блоки приложения и дано объяснение каждому из этих блоков. Было показано создания приложения в Twitter Streaming API, которое позволяет получать данные от Twitter в режиме реального времени. Был приведен код программы позволяющий оценивать настроение emoji. Также в заключении данной главы, было продемонстрирована и объяснена, работа с большими объемами информации. Для этого были выбраны брокер задач RabbitMQ и асинхронная очередь заданий, основанная на распределенной передачи сообщений Celery. Также была продемонстрирована упрощенная архитектура приложения.

ЗАКЛЮЧЕНИЕ

В современном обществе социальные сети представляют собой большие объемы структурированной и неструктурированной информации, которую можно анализировать и использовать. Информацию можно задействовать в различных целях, например для предугадывания поведения, настройки рекламных компаний, совершенствования диагностики, создания автопилотов.

С помощью использования углубленного анализа больших объемов информации, предприятия могут использовать социальные сети, чтобы узнать о связях лежащих в их основе, которые характеризуют социальное поведение отдельных лиц и групп. Благодаря анализу социальных сетей, предприятия могут извлекать необходимую информацию с целью бизнес-аналитики, увеличения эффективности маркетинга, отслеживания распространения информации, увеличения прибыли.

В процессе разработки данного проекта были использованы современные средства разработки программного обеспечения. Были использованы инструменты для быстрого и легкого развертывания приложения. Приложение построено по принципу «микросервис». Это сделано для легкой интеграции данного приложения с другими сервисами или системами. При реализации данного проекта были задействованы современные методы обработки больших объемов информации, а также способы обработки неструктурированных данных. Также были исследованы и использованы алгоритмы оценки настроений текста и emoji. Были использованы способы увеличения скорости обработки больших объемов информации, для повышения производительности приложения.

В ходе разработки данного проекта были проанализированы существующие средства и системы сбора и анализа информации из социальных сетей. Были выявлены их сильные и слабые стороны.

Результаты, полученные в ходе реализации данного проекта, могут быть использованы для интеграции в системы, предназначенные для анализа поведения рынка, с целью увеличения эффективности. Также результаты могут быть положены в основу более глубокого и детального анализа социальных сетей, с целью извлечения информации для бизнес-аналитики и использованы для отслеживания распространения информации в социальных сетях в режиме реального времени.

Продолжением разработки данного проекта может служить добавление нескольких новых возможностей: увеличение источников получаемой информации, создание пользовательского интерфейса, улучшение отчетности, планирование времени сбора информации, создание личного кабинета пользователей.

СПИСОК ОПУБЛИКОВАННЫХ РАБОТ

1. Обнаружение передачи несанкционированного траффика посредством туннелирования DNS. Калабухов Е.В., Недведский А.Ю., Масензов В.В., Якубович Ф.В. Номер 05(20) 2017 года журнала «Наука, образование и культура».

2. Технология HealthCloud на базе одноименной CRM-системы Salesforce. Калабухов Е.В., Масензов В.В.,Недведский А.Ю., Якубович Ф.В. Номер 05(20) 2017 года журнала «Наука, образование и культура».