МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
**«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ
ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ
ИМЕНИ Н. Г. ЧЕРНЫШЕВСКОГО»**
Кафедра дискретной математики и информационных технологий

ОБНАРУЖЕНИЕ АНОМАЛИЙ ДАННЫХ В ПРОГНОЗНОМ ОБСЛУЖИВАНИИ

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студентки 4 курса 421 группы
направления 09.03.01 — Информатика и вычислительная техника
факультета КНиИТ
Сорокованової Светланы Алексеевны

Научный руководитель
профессор, к. э. н. _____ Л. В. Кальянов

Заведующий кафедрой
доцент, к. ф.-м. н. _____ Л. Б. Тяпаев

Саратов 2020

ВВЕДЕНИЕ

В современном мире мы всё чаще сталкиваемся с огромным количеством данных, проблемой их хранения и использования с разными целями. Программа Hadoop помогает справиться с этими проблемами. Hadoop – это проект с открытым исходным кодом, находящийся под управлением Apache Software Foundation. Используется для масштабируемых и распределённых вычислений, в исследовательских и производственных целях, а также применяется как хранилище файлов общего значения, способное вместить петабайты данных. Идея работы состоит в том, что увеличение информационных потоков ведет к увеличению объёмов не структурированной информации, что как правило приводит к увеличению «информационного шума», т.е. к уменьшению доли полезной информации в её общем объёме. Говоря более понятным языком, чем больше количества информации существует, а её с каждой минутой становится всё больше, тем больше становится излишней, бесполезной, ненужной, повторяющейся, одинаковой информации, которая мешает продуктивному поиску конкретной темы. Кроме того, такая информация не может храниться централизовано, то есть находится в рандомных частях хранилищ, и также препятствует скорости нахождения необходимой, полезной информации. Целью работы является изучение проблем доступа, методов обработки больших данных и применение технологий на практике для обнаружения аномальных данных.

В связи с этим возникают задачи:

1. организовать доступ и обработку больших объёмов не структурированной, децентрализованной информации
2. научиться анализировать аномальные данные

КРАТКОЕ СОДЕРЖАНИЕ

Первый раздел дипломной работы описывает основные понятия, связанные с технологиями Big Data. Данные раздел содержит два подраздела. В первом мы рассматриваем проблемы доступа и обработки больших данных. Второй подраздел содержит информацию о технологиях работы с Big Data, описывает программу Hadoop, её историю и использование, включает в себя анализ данных.

Второй раздел дипломной работы посвящён реализации процесса обнаружения аномальных данных в среде KNIME. Раздел содержит в себе два подраздела. В первом мы рассматриваем анализ аномалий данных, и этот подраздел содержит свои три подраздела. В первый из них содержит информацию о том, что такое идеальное предсказание и как его воспроизвести. Второй раскрывает проблему нахождения данных для обнаружения аномалий данных в прогнозном обслуживании. В третьем подразделе подробно рассматриваем методы предварительной обработки данных. Этот раздел разбит на три подраздела. В первом речь идёт о построении модели «чтение данных», во втором о частотном преобразовании данных, в третьем о согласовании времени наблюдений. Второй подраздел второго раздела дипломной работы даёт представление о визуализации аномальных данных. Данные раздел содержит в себе пять подразделов. В первом можно наглядно увидеть аномалию данных с помощью метода визуализации «графики временных рядов». Во втором также видим аномалию, но уже с помощью другого способа визуализации – матрицы рассеяния. В третьем данная аномалия показана со стороны метода построения корреляционных карт. В четвёртом со стороны автокорреляционных карт. В пятом описывается построение тепловой карты и то, как с помощью неё можно увидеть аномалию.

Данные методы рассмотрены с помощью программы KNIME Analytics, на рисунках показаны примеры работы программы.

ЗАКЛЮЧЕНИЕ

Заданной целью являлось изучение проблем доступа, методов обработки больших данных и применение технологий на практике для обнаружения аномалий данных – она была достигнута. Поставленные задачи, а именно, организация доступа, обработка больших объёмов не структурированной, децентрализованной информации, анализ аномальных данных, были также достигнуты. В этой работе мы ознакомились с технологиями Big Data, проблемами доступа и обработки больших данных, узнали о технологиях работы с Big Data, а также рассмотрели на практике и визуализировали данные временных рядов из сети датчиков, отслеживающих рабочий ротор, с помощью программы Knime Analytics. Благодаря реализованному процессу анализа аномальных данных мы смогли выявить элементы технического устройства, нуждающиеся в обслуживании, и предотвратить выход сложной технической системы из строя.

Поставленные перед нами задачи были решены, а именно была произведена работа в несколько этапов анализа данных: сначала мы разработали процесс для чтения данных и их предварительной обработки, затем преобразовали данные по частотам и разработали процесс для согласования времени, также разработали процесс построения графиков временных рядов и процесс построения матрицы рассеяния, далее создали корреляционные, автокорреляционные и тепловые карты. После этого мы визуально исследовали изменения временных рядов, используя пять различных методов визуализации: линейные графики, матрицу рассеяния, карты автокорреляции, карты корреляции и тепловые карты. На полученных изображениях чётко видно эпизод аномалии в некоторых полосах частот. Описанные процедуры могут быть легко использованы для аналогичных проблем в другой сфере, например, при отслеживании потребительского спроса на определённые виды товаров в конкретный период времени в магазине или отслеживании данных об авариях на участках дороги за последние годы для дальнейшего программирования навигатора с объездом

наиболее опасных зон.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Выявление аномалий [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/%D0%92%D1%8B%D1%8F%D0%B2%D0%BB%D0%B5%D0%BD%D0%B8%D0%B5_%D0%B0%D0%BD%D0%BE%D0%BC%D0%B0%D0%BB%D0%B8%D0%B9;

2 Поиск аномалий [Электронный ресурс]. URL: https://dyakonov.org/2017/04/19/%D0%BF%D0%BE%D0%B8%D1%81%D0%BA-%D0%B0%D0%BD%D0%BE%D0%BC%D0%B0%D0%BB%D0%B8%D0%B9-anomaly-detection/;

3 Методы обнаружения аномалий: обнаружение нормального [Электронный ресурс]. URL: https://www.knime.com/blog/anomaly-detection-techniques-defining-normal;

4 Обнаружение аномалий в прогнозном обслуживании с анализов временных рядов [Электронный ресурс]. URL: https://www.knime.com/blog/anomaly-detection-in-predictive-maintenance-with-time-series-analysis;

5 Корреляция Пирсона [Электронный ресурс]. URL: http://statsoft.ru/home/textbook/glossary/gloss_k.html;

6 Jack Grieve «Regional Variation in Written American English. Spatial analysis» / Cambridge university press // 2016. -335 c;

7 Тепловая карта [Электронный ресурс]. URL: https://ru.wikipedia.org/wiki/%D0%A2%D0%B5%D0%BF%D0%BB%D0%BE%D0%B2%D0%B0%D1%8F_%D0%BA%D0%B0%D1%80%D1%82%D0%B0;

8 Матрица рассеяния [Электронный ресурс]. URL: https://www.booksite.ru/fulltext/1/001/008/074/417.htm;

9 Матрица рассеяния или s – матрица [Электронный ресурс]. URL: https://ru.qwe.wiki/wiki/S-matrix;

10 Обнаружение аномалий и мониторинг состояния [Электронный ресурс]. URL: https://www.machinelearningmastery.ru/how-to-use-machine-learning-for-anomaly-detection-and-condition-monitoring-6742f82900d7/.

11 Hadoop [Электронный ресурс]. URL: http://www.taskdata.com/index.php?option=com_content&view=article&id=26&Itemid=5&lang=ru

12 История Hadoop [Электронный ресурс]. URL: https://www.bigdataschool.ru/wiki/hadoop