МИНОБРНАУКИ РОССИИ
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«САРАТОВСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ ИМЕНИ Н.Г.ЧЕРНЫШЕВСКОГО»

Кафедра дискретной математики и информационных технологий

Информационные технологии выявления аномальных значений в бизнес-данных

АВТОРЕФЕРАТ БАКАЛАВРСКОЙ РАБОТЫ

студента 5 курса 521 группы
направления 09.03.01 «Информатика и вычислительная техника»
факультета компьютерных наук и информационных технологий
Лачинова Вадима Александровича

Научный руководитель
д.э.н, профессор _____ Л.В. Кальянов
подпись, дата

Зав. кафедрой
к. ф.-м.н., доцент _____ Л.Б. Тяпаев
подпись, дата

Саратов 2019

ВВЕДЕНИЕ

В настоящее время человек ежедневно сталкивается с большим объёмом информации. Когда стоит задача проанализировать эту информацию, высокую актуальность имеют технологии интеллектуального анализа данных (data mining). Интеллектуальный анализ данных состоит из методов классификации, моделирования и прогнозирования, совокупности статистических методов. Эти технологии открывают новые возможности исследователям данных, бизнес-аналитикам, инженерам, работающим с машинным обучением. Системы интеллектуального анализа данных успешно применяются в образовании, различных научных исследованиях, для анализа результатов производства и продаж, в оценке различных рисков. Разработка инструментов текстовой аналитики и методов выявления аномальных значений в данных являются актуальными задачами в области интеллектуального анализа. Актуальность работы заключается в том, что с помощью алгоритмов web-анализа и выявления аномальных значений в данных специалист получает требуемый ему результат анализа при минимальных затратах времени и ресурсов.

Целью бакалаврской работы является разработка информационной технологии выявления аномальных значений в бизнес-данных средствами платформы RapidMiner.

Поставленная цель определила следующие задачи:

1. Изучение технологий веб-краулинга и текстовой аналитики;
2. Обзор функциональных возможностей платформы RapidMiner, обзор её основных и дополнительных операторов;
3. Реализация алгоритма выявления аномальных значений в бизнес-данных средствами платформы RapidMiner;
4. Анализ аномалий на примере набора данных со статистикой аккаунта в социальной сети Twitter.

Выпускная квалификационная работа состоит из введения, трех разделов, заключения, списка используемых источников и трех приложений.

Общий объем работы – 49 страниц, из них 37 страниц – основное содержание, включая 31 рисунок, цифровой носитель в качестве приложения, список использованных источников информации – 20 наименований.

Методологические основы разработки и реализации были представлены в работе Маркуса Голдштейна «Anomaly Detection in Large Datasets», Маркуса Хоффманна и Ральфа Клинкенберга «RapidMiner Data Mining Use Cases and Business Analytics Applications».

Практическая значимость работы заключается в том, что разработанный алгоритм позволяет анализировать данные (набор данных) на наличие в них аномальных значений, что даёт аналитику возможность определить, какая часть значений является аномальной, с чем может быть связано появление аномалий, понять общую тенденцию.

КРАТКОЕ СОДЕРЖАНИЕ РАБОТЫ

Первый раздел «Методы интеллектуальной аналитики текста и виды аномалий». В разделе дается определение понятия «text mining» и сопутствующим понятиям. Рассмотрены средства интеллектуального анализа текста, виды аномалий, техники выявления аномальных значений.

Во втором разделе «Программные инструменты Text Mining и выявления аномалий» рассмотрены платформы для работы с данными Knime и RapidMiner, проведён сравнительный анализ этих платформ, по итогам которого для работы была выбрана платформа RapidMiner. Подробно рассмотрены и описаны все необходимые для работы функциональные операторы платформы, выполнен поиск необходимых дополнений в RapidMiner Marketplace, необходимых для реализации алгоритма.

В третьем разделе «Разработка информационной технологии веб-анализа и алгоритма выявления аномальных значений средствами RapidMiner» описываются созданные схемы, содержащие алгоритмы веб-анализа и выявления аномальных значений. Всего было разработано 4 алгоритма:

1. Алгоритм поиска веб-страниц и сохранения содержимого;
2. Алгоритм текстового анализа полученных данных;
3. Алгоритм поиска выбросов в наборе данных;
4. Алгоритм выявления аномальных значений в данных.

Для каждого алгоритма показана структура, принцип работы и область применения. Выведены результаты работы разработанных алгоритмов.

ЗАКЛЮЧЕНИЕ

В современном мире информация имеет большое значение для человека. Возможность анализа информации позволяет сократить время на поиск нужной информации, повысить качество работы в различных сферах деятельности человека. Инструменты интеллектуального анализа данных дают возможность специалистам, аналитикам и инженерам проводить моделирование и прогнозирование.

В ходе выполнения работы были решены следующие задачи:

- рассмотрение текстовой аналитики и технологии веб-краулинга как наиболее результативных методов поиска информации;
- изучение платформы RapidMiner, обзор стандартных и дополнительных функциональных операторов;
- реализация алгоритма выявления аномалий в наборе данных средствами платформы RapidMiner;
- анализ аномалий на примере набора данных со статистикой аккаунта в социальной сети Twitter.

Таким образом, стоит заключить, что платформа RapidMiner имеет широкий набор инструментов для подготовки данных, машинного обучения, глубокого обучения, анализа текста и прогнозной аналитики. Стандартный набор операторов RapidMiner может быть дополнен внешними расширениями из RapidMiner Marketplace.

Алгоритмы веб-анализа и веб-краулинга предоставляют информацию с интернет-страниц и различных ресурсов в удобном формате и большом количестве. Скорость работы алгоритма позволяет получать данные в короткий срок.

Стоит понимать, что алгоритм выявления аномальных значений в данных производит анализ документа, удаляет информацию, не представляющую интереса для аналитика, сортирует и упорядочивает значения в зависимости от заданных параметров. Выявление аномалий влияет на результаты научных исследований, позволяют повысить точность

экспериментов. На сегодняшний день проблема выявления аномальных значений является актуальной задачей для исследователей и аналитиков в разных сферах.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 Markus Goldstein – Anomaly Detection in Large Datasets. PhD-Thesis, Pages 248, Technische Universität Kaiserslautern, Dr. Hut Verlag München, 2014. ISBN: 978-3-8439-1572-4.

2 Markus Hofmann, Ralf Klinkenberg – RapidMiner Data Mining Use Cases and Business Analytics Applications. 2014. ISBN 978-1-4822-0550-3.

3 Soumen Chakrabati – Mining the Web: Discovering Knowledge from Hypertext Data. 2002. ISBN 1-55860-754-4.

4 Herbert Edelstein – Introduction to Data Mining and Knowledge Discovery. 1999. ISBN: 1-892095-02-5.

5 Andrew Chisholm – Exploring Data with RapidMiner. 2013. ISBN 978-1-78216-933-8.

6 UC Irvine Machine Learning Repository. [Электронный ресурс]: URL: https://archive.ics.uci.edu/ml/index.php (дата обращения 19.05.2019)

7 Хенрик Бринк, Джозеф Ричардс, Марка Феверолф – Машинное обучение. 2017. ISBN 978-5-496-02989-6.

8 Ronen Feldman, James Sanger – The text mining. 2007. ISBN:0521836573.

9 Michael W. Berry – Survey of Text Mining Clustering, Classification, and Retrieval Scanned by Velocity. University of Tennessee. 2004.

10 Roger Bilisoly – Practical Text Mining. Central Connecticut State University. 2008.