DOI: http://dx.doi.org/10.21686/1818-4243-2019-2-42-49

К.В. Мулюкова, В.М. Курейчик

Инженерно-технологическая академия
Южного федерального университета, Ростов-на-Дону, Россия

Проблема анализа больших веб-данных и использование технологии Data Mining для обработки и поиска закономерностей в большом массиве веб-данных на практическом примере

Целью работы *является исследование современных проблем и перспектив решения обработки больших данных, получаемых или сохраняемых в сети Интернет (веб-данных), а так же возможность практической реализации технологии Data Mining для больших веб-данных на практическом примере.*

Материалы и методы. *Исследование включало в себя обзор библиографических источников по проблемам анализа больших данных.*

Была применена технология Data Mining для анализа больших веб-данных, а также компьютерное моделирование практической задачи с помощью языка программирования C# и создания структуры базы данных на языке описания данных DDL для накопления веб-данных.

Результаты. *В ходе работы описана специфика больших данных, были выделены основные характеристики больших данных, а также были проанализированные современные подходы к обработке больших данных. Дана краткая характеристика горизонтально-масштабируемой архитектуры и архитектуры BI-решения для обработки больших данных. Сформулированы проблемы обработки больших веб – данных: ограничение скорости доступа к данным, организация доступа по сетевым протоколам через сети общего назначения.*

Так же был реализован пример, показывающий подход к обработке больших веб-данных. На основе представления о больших данных, описанных сложностях обработки веб-данных и методах Data Mining, были предложены приёмы эффективного решения поставленной практической задачи обработки и поиска закономерностей в большом массиве данных.

Были разработаны следующие классы на языке программирования C#:

класс получения веб-данных через Интернет;
класс преобразования данных;
класс интеллектуальной обработки данных.

Создан DDL-скрипт, создающий структуру для накопления веб-данных.

Разработана единая UML-диаграмма классов.

Построенная система данных и классов позволяет решить основную часть проблем обработки больших веб-данных и выполнить интеллектуальную обработку по технологии Data Mining с целью решения поставленной задачи выявления определенных записей в большом массиве. Сочетание объектно-ориентированного подхода, нейронных сетей и BI-анализа для фильтрации данных позволит максимально ускорить процесс обработки данных и получения результата исследования

Заключение. *По результатам проведённого исследования, можно утверждать, что современное состояние технологии анализа больших веб-данных позволяет эффективно обрабатывать объекты данных, выявлять закономерности, получать скрытые данные и получать полноценные статистические данные.*

Полученные результаты могут использоваться как в целях первичного изучения технологий обработки больших данных, так и в качестве основы разработки уже реального приложения для анализа веб-данных. Использование нейронных сетей и созданных универсальных классов-обработчиков делает созданную архитектуру гибкой и самообучаемой, а декларации классов и DDL-структура базы существенно упростят разработку программного кода.

Ключевые слова: *большие данные, Data Mining, веб - данные, Business Intelligence (BI), DDL-структура. анализ данных, big date, интеллектуальная обработка данных*

Ksenia V. Mulyukova, Victor M. Kureichik

Engineering-Technological Academy of SFU, Rostov-on-Don, Russia

The problem of analysis of big web data and the use of data mining technology for processing and searching patterns in big web data on a practical example

The purpose of the work *is to study the current problems and prospects of the solution for processing big data received or stored in the Internet (web data), as well as the possibility of practical realization of Data Mining technology for big web data on practical example.*

Materials and methods. *The study included a review of bibliographic sources on big data analysis problems.*

Data Mining technology was used to analyze large web data, as well as computer modeling of a practical problem using the C # programming language and creating a DDL database structure for accumulating web data.

Results. *In the course of the work, the specifics of big data were described, the main characteristics of big data were highlighted, and modern approaches to processing big data were analyzed. A brief description of the horizontal-scalable architecture and the*

*Работа выполнена за счет частичного финансирования по гранту РФФИ ГР №18-07-00050

BI-solution architecture for big data processing is given. The problems of processing large web data are formulated: limiting the speed of access to data, providing access via network protocols through general-purpose networks.
An example showing the approach to processing large web data was also implemented. Based on the idea of big data, the described complexities of web data processing and the methods of Data Mining, techniques were proposed for effectively solving the practical problem of processing and searching patterns in a large data array.
The following classes have been developed in the C # programming language:
Class of receiving web data via the Internet;
Data conversion class;
Intelligent data processing class;
Created DDL script that creates a structure for the accumulation of web data.
A single UML class diagram has been developed.
The constructed system of data and classes allows to solve the main part of the problems of processing large web data and perform intelligent processing using Data Mining technology in order to solve the problem posed of identifying certain records in a large array. The combination of object-oriented approach, neural networks and BI-analysis to filter data will speed up the process of data processing and obtaining the result of the study
***Conclusion.** According to the results of the study, it can be argued that the current state of technology for analyzing large web data allows you to efficiently process data objects, identify patterns, get hidden data and get full-fledged statistical data.*
The obtained results can be used both for the purpose of the initial study of big data processing technologies, and as a basis for developing an already real application for analyzing web data. The use of neural networks and the created universal classes-handlers makes the created architecture flexible and self-learning, and the class declarations and the base DDL structure will greatly simplify the development of program code.

***Keywords:** big data, Data Mining, web data, Business Intelligence (BI), DDL structure. data analysis, big date*

Введение

За последнее десятилетие объем создаваемых данных стремительно растет. Каждую секунду формируется более 30 тысяч гигабайт данных, и скорость их создания только увеличивается. Нам постоянно приходится иметь дело с разнообразными данными. Пользователи создают контент вроде сообщений в блогах и социальных сетях, публикуя свои видео и фотографии [1]. Серверы постоянно регистрируют сообщения о выполняемых операциях и размещают контент пользователей. Интернет окончательно стал основным и необъятно большим хранилищем и источником данных.[2]

Обработка данных средствами вычислительной техники является одной из основных задач большинства информационных систем. Любая информация, структурированная определённым образом, может быть обработана как для получения непосредственных результатов вычислений, так и для подготовки к передаче по каналам связи или дальнейшей обработки. По мере развития средств хранения и коммуникаций, объёмы информации возрастают нелинейно. Количественное изменение массива обрабатываемой информации переходит в качественное новое состояние — большие данные.

До начала 2000-х годов, можно говорить об отсутствии больших данных в практической и теоретической областях знаний. Большая часть массивов данных на тот период была локальна, структурирована и сосредоточена, а не распределена между различными узлами. Качественный скачок в появлении больших данных связывают с двумя факторами:

- резкий рост с началом 2000-х годов объёма цифровой информации;
- массовое повышение скорости доступа в сеть Интернет, что сделало возможным не только передачу, но и хранение данных с постоянным доступом.

Активное формирование больших данных, как научного направления началось в 2008, когда Клиффорд Линч ввел термин «большие данные» в журнале «Nature».

В 2010 г. XXI в. в своих книгах Марц Натан и Уоррен Джеймс дают представление о теоретических основах больших данных, а также об их реализации на практике [3].

До перехода глобальной сети на новую методику проектирования, получившую название Web 2.0. Интернет предоставлял скорее услуги связи и передачи данных, чем глобальную систему хранения. Но с переходом на Web 2.0. Интернет становится глобальным хранилищем данных.

Авинаш Кошик в своей книге пишет, что «Интернет – совершенно уникальное явление, не похожее ни на что другое. И оно требует совершенно индивидуального подхода к проблеме обработки данных [4].»

К проблемам хранения и обработки больших данных, размещённых в сети Интернет, добавлены дополнительные особенности, обусловленные спецификой хранения веб-данных:

– веб-данные расположены в сети Интернет и к ним нет прямого доступа – вся обработка выполняется через межсетевое взаимодействие;

– скорость доступа к веб-данным существенно ниже скорости доступа к локальным данным;

– из-за нестабильных каналов связи, возможны ошибки передачи и последующей обработки этих данных, что потребует дополнительной интеллектуальной проверки результатов.

На данный момент не существует единых устоявшихся решений ни в области теоретических способов обработки, ни на рынке программных продуктов для подобных задач. Как правило, компании и разработчики, связанные с обработкой больших веб-данных, реализуют собственные решения или адаптируют существующие.

Актуальность выбранной темы обусловлена трендом по “цифровизации” всех аспектов жизни современного общества. Как показали исследования, каждые два года в течение последних трёх десятилетий количество информации увеличивается приблизительно в десять раз – темп, который оставляет далеко позади даже закон Мура об удвоении мощности процессоров. Соответственно, перед современным обществом с каждым годом всё острее и острее будет вставать проблема хранения и, главное – обработки больших объёмов информации. Данная проблема будет становиться всё острее, т.к. темп роста информации превышает темп роста вычислительных мощностей в мире, в связи с чем потребуется разработка соответствующих методик, которые были бы способны оперировать такими большими объёмами данных.

Целью работы является исследование современных проблем и перспектив решения обработки больших данных, получаемых или сохраняемых в сети Интернет (веб-данных), а так же возможность практической реализации технологии Data Mining для больших веб-данных на практическом примере.

Что такое большие данные?

Под большими данными обычно подразумеваются обобщенные наборы, включающие в себя структурированные и неструктурированные данные, существенные по объему и разнообразные по структуре [5].

Сформулируем характерные признаки больших данных, которые позволили бы не количественно, но качественно отличить их от обычных, традиционных массивов данных. Как и любое нечеткое понятие, определяемое чаще всего по факту, оно может быть описано несколькими признаками, дополняющими друг друга. Чем больше признаков соответствует исследуемому набору данных, тем более вероятно, что массив относится к большим данным. Выделим более подробно основные признаки:

Масштабность. Если обычные данные чаще всего локализованы как частные базы данных, хранящие узкие данные (бухгалтерия предприятия, списки сотрудников, отчеты о продажах) – то большие данные содержат масштабные записи, включающие в себя сведения о миллионах людей или об огромных территориях, например, общегосударственный реестр сделок с недвижимостью или база биллинга оператора сотовой сети.

Распределенность. Традиционно конкретные данные содержатся на одном-двух носителях, обрабатываются в едином адресном пространстве приложения и не превышают размеров файловой системы. Большие данные чаще всего распределены между многими носителями.

Многоструктурность. Классические структуры данных оформляются в виде кортежей «поля-записи» в случае реляционных таблиц, или в виде хранилища типовых документов в документо-ориентированных системах. Большие данные могут содержать сотни различных структур, часть из которых даже не имеет описания данных на уровне хранилища, что требует перестраивать обработку данных по мере выявления новых структур.

Протяженность по времени. Чаще всего, типовые массивы данных содержат информацию за небольшой период времени — текущий месяц, квартал, отчетный год. Этого достаточно для задач учета и первичного анализа. Большие данные могут хранить информацию за десятки лет, что требует введения особых методов вычислений, если расчеты касаются временной динамики.

Для более доступного восприятия данной информации была создана следующая таблица.

Таблица

Характерные признаки больших данных

Признак	Пример
Масштабность	Большие данные содержат огромные потоки информации
Распределенность	Большие данные распределены по многим носителям
Многоструктурность	Большие данные содержат множество различных структур
Протяженность по времени	Зачастую большие данные содержат информацию за большой временной срок

В данный момент на практике используются два основных подхода к обработке больших данных: горизонтально-масштабируемая архитектура обработки данных и архитектура BI-решения для обработки больших данных.

Горизонтально-масштабируемая архитектура обработки данных использует отдельные узлы для обработки отдельных частей данных, и применяет централизованный узел связи только для синхронизации своей работы [6]. С практической точки зрения это означает, что для увеличения мощности вычислительной системы в два раза, достаточно удвоить число вычислительных узлов без увеличения стоимости и сложности отдельного узла.

Альтернативным подходом к обработке больших данных является т.н. Business Intelligence (BI) [7], совокупность методов и технологий для интеллектуальной обработки в первую очередь деловых данных. Этот подход подразумевает, что данные переводятся в форму более удобную для дальнейшей обработки и затем анализируются традиционными инструментами [8].

Все вышеперечисленные подходы применимы к большим данным в обобщённой форме. При использовании веб-данных, перечис-

ленные технологии осложняются факторами специфики сетевых приложений. Наибольшей сложностью на практическом уровне использования является ограничение скорости доступа к данным, из-за чего даже относительно небольшие массивы могут обрабатываться недели и месяцы. Другую типичную проблему представляет организация доступа по сетевым протоколам через сети общего назначения, что с одной стороны понижает безопасность данных [9], а с другой — усложняет программную реализацию решения по обработке.

На примере постановки задачи для веб-ориентированного способа обработки больших данных, мы опишем решение подобной проблемы и разработаем архитектуру системы.

2. Использование технологии Data Mining для решения задач, которые содержат большие веб-данные

Обработка больших данных представляет собой комплексную задачу, не имеющую однозначного решения и осложненную рядом факторов. На эти факторы дополнительно накладываются проблемы канала доступа к информации и вопросы сетевых протоколов в процессе обработки веб-данных.

Сформулируем практическую задачу, подходящую по условиям к описанным ранее признакам для больших данных.

2.1. Пример практической реализации задачи поиска инвестиционных объектов в базе данных объявлений

Практическая задача обработки больших веб-данных—найти среди большого объема данных о продажах недвижимости (порядка 30-40 миллионов записей) такие объекты, которые покупались и продавались несколько раз в течение указанного периода времени. В основе такого анализа лежат следующие аппроксимации:

1) Покупка и продажа выполнялась не менее 2 раз за период.

2) Параметры объекта совпадают с точностью до адреса.

3) Каждый объект имеет уникальный код, который может быть найден в базе, но повторная покупка или продажа создает новый код.

4) Возможны различные вариации записи адреса.

Очевидно, что подобная задача, выполняемая традиционными способами через линейные запросы к базе данных, потребует времени, которое пропорционально как минимум второй степени количества записей для двух продаж или покупок. Применительно к веб-данным, задача уточняется следующими двумя положениями:

1) Различные записи находятся на различных источниках в сети Интернет, доступ к которым может быть ограничен по скорости или по числу запросов за период

2) Структура данных на каждом источнике отличается

Таким образом, мы имеет уточненное задание обработки больших данных в сети Интернет, которые являются распределенными и неструктурированными, но над которыми необходимо выполнить анализ по строгим правилам расчета.

Для решения подобных задач, используют технологию Data Mining. Под этим названием обозначают совокупность методов, позволяющих извлекать полезные сведения по определенным правилам из объема собранных данных, в которых эти сведения содержатся неявно [2]. При этом используют как статистические, так и интеллектуальные методы обработки [10]. Поставленная нами задача подходит под эти условия — в большом массиве записей о продаже/покупке есть интересующие нас сведения об инвестиционных объектах, но мы не можем получить эти данные явным образом.

Для применения Data Mining к конкретной задаче, необходимо выполнить следующие подготовительные шаги:

1) Определить правила, по которым будут вычисляться интересующие нас данные;

2) Записать эти правила на искусственном языке работы с базой данных [11]

Применительно к веб-данным, потребуются два дополнительных шага:

1) Разработать надежный алгоритм для получения данных из источников данных для алгоритма;

2) Разработать модули-переходники, которые приведут различные данные к единообразному виду перед их передачей в алгоритм.

Для наглядности приведем на рис. 1 обобщенную архитектуру приложения для решения указанной задачи.

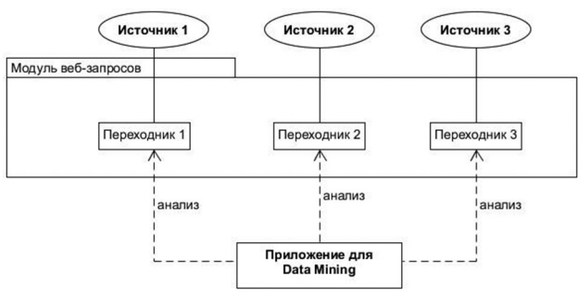

Рис. 1. Архитектура приложения для обработки веб-данных

На основе представления о больших данных, описанных сложностях обработки веб-данных и методах Data Mining, мы сформируем приёмы эффективного решения поставленной ранее задачи обработки и поиска закономерностей в большом массиве данных.

Хотя данная статья не предполагает полноценной разработки программного решения, мы опишем отдельные части системы на языке C# в виде деклараций классов, а структуру базы данных — на языке DDL. Это позволит более конкретно представить технологию решения задачи с использованием Data Mining в области больших веб-данных.

2.2. Получение веб-данных через Интернет

В соответствии со спецификой обработки больших веб-данных, мы создаем модуль, который позволит получать данные из Интернет через относительно медленный и нестабильный канал связи. Такой модуль позволит выполнять Data Mining на произвольном количестве источников, без потери качества и с максимально возможной эффективностью в заданных условиях. Опишем интерфейс класса для реализации модуля получения веб-данных.

```
public class WebGetter
{
public string getData(string ActiveSource, ref bool isSourceOk) ;
public bool testConnection() ;
}
```

где первый метод — получает данные из источника и проверяет, активен ли источник в данное время. В случае, если источник по каким-то причинам перестал отвечать на запрос, класс переключается на другой источник и таким образом, сберегает время для обработки. Второй же метод проверяет наличие Интернет-соединения и позволяет временно отключить обработку данных, если нет возможности получить новые сведения из источников.

Несмотря на то, что данный метод применим как к большим данным, так и к любым Интернет-программам, выполняющим загрузку данных, его применение в нашей задаче особенно актуально в силу большого количества данных и их источников.

2.3. Модули преобразования данных

После получения данных из разных источников, мы должны подготовить их для сохранения в постоянной или временной базе к последующей передаче в алгоритмы Data Mining. Это необходимо для обработки разнородных данных [12], которые должны быть приведены в единый формат. Приведение может быть как постоянным с сохранением в базе данных, так и временным на этапе обработки. Поскольку веб-данные гораздо сложнее получить, чем обработать, представляется целесообразным в ряде случае использовать временное интеллектуальное преобразование данных.

Здесь применяется нейронная сеть [13], которая получает входной формат строки с данными, предположительно содержащими в себе информацию об объектах недвижимости. Предварительное обучение и последующее получение данных позволит ускорить преобразование разнородных данных [14]. Класс такого интеллектуального анализа показан ниже.

```
public class ConvertString
{
public void learnParam(string source, string name, string value) ;
public Dictionary<string,string> extractData(string source) ;
}
```

где первый метод получает строку исходных веб-данных и параметр, который в ней содержится, а второй метод позволяет получать данные из строки на основе ранее выполненного обучения.

2.4. Структура базы данных

Структура базы данных, оптимизированная для Data Mining, может быть как реляционной, так и документо-ориентированной [15]. С учетом большого количества данных и малого размера одной записи в изучаемой системе, мы выбираем реляционную базу данных, однако, вводим некоторые поправки на специфику веб-данных и интеллектуальную обработку:

1) Часть данных сохраняется в том виде, в каком получена из источника — это может применяться для последующего обучения нейронной сети

2) Каждая запись имеет указание на веб-источник — это не требуется для анализа данных нашей задачи, но будет полезно для выявления других закономерностей

Ниже приведен DDL-скрипт, создающий структуру для накопления веб-данных.

```
CREATE TABLE ADVS (
    ID          INTEGER NOT NULL,
    ADDRESS       VARCHAR(255),
    DATEENTER   DATE,
    CITY        VARCHAR(32),
    ROOMTYPE    VARCHAR(32),
    OPERTYPE    VARCHAR(32),
    FLOOR       VARCHAR(3),
    FLOORALL    VARCHAR(3),
    SOURCE      VARCHAR(32),
    SQUARE      VARCHAR(5),
    PRICE       VARCHAR(9)
);
```

В заданной таблице каждое поле содержит отдельные данные. Для первичной обработки по BI-методике, огромный массив данных предварительно будет отфильтрован по простым полям, таким как этаж и площадь, а уже далее более компактные наборы будут обрабатываться интеллектуальной технологией Data Mining.

2.5. Интеллектуальная обработка данных

На уровне обработки полученных больших данных, мы будем применять описанную ранее модель BI-анализа [16], которая позволит провести предварительный отбор данных для получения результатов. Для получения инвестиционных объектов, необходимо отобрать объекты с совпадающим адресом и чередованием операций «купля-продажа» не менее двух раз, при этом все остальные параметры объекта, кроме цены, должны совпадать для дополнительной проверки.

В рассматриваемой задаче сформируем условия предварительной интеллектуальной обработки по адресу. Эту обработку должна выполнять упрощенная нейронная сеть, обучаемая на сравнении адресов.

1) Дом без квартиры считает равным дому без квартиры, если по этому адресу нет ни одного объекта с квартирой

2) Дом с квартирой считается равным дому с такой же квартирой, если нет уточняющих индексов квартиры.

3) Номера домов считаются равными, если они совпадают с точностью до дроби/литера, или если таковые отсутствуют.

Создаваемый фильтр интеллектуальной обработки состоит из двух частей — обучения и проверки. Класс AddressCompare, используемый для предварительной обработки, приведен ниже

```
public class AddressCompare
{
public void learnComparison(string address1, string address 2, bool isequal) ;
public bool doComparison(string address1, string address 2) ;
}
```

где первый метод получает два адреса и значение, равны ли они, а второй позволяет на основе обученной сети получить их сравнение.

Для наглядности выше изложенной информации обобщим все классы в единую UML-диаграмму, показанную на рис. 2.

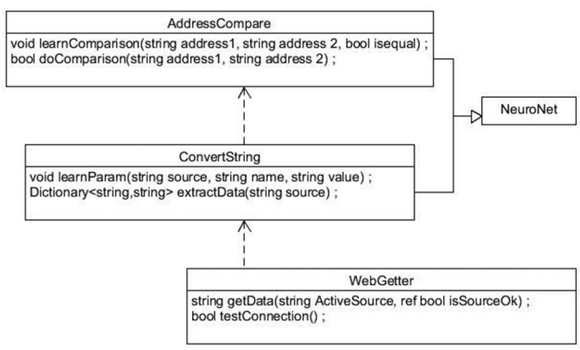

Рис. 2. UML-диаграмма системы обработки больших веб-данных

Построенная система данных и классов позволит решить основную часть проблем обработки больших веб-данных и выполнить интеллектуальную обработку по технологии Data Mining с целью решения поставленной задачи выявления определенных записей в большом массиве. Сочетание объектно-ориентированного подхода, нейронных сетей и BI-анализа для фильтрации данных позволит максимально ускорить процесс обработки и получения результата исследования

Приведенная архитектура решения позволяет выполнять эффективный анализ больших веб-данных, получаемых из разных источников. Конкретное решение может отличаться в деталях и завесить от используемой программно-аппаратной платформы. В общем случае, при реализации данной архитектуры обработки веб-данных на уровне программного кода, необходимы дополнительные уточнения задачи:

1) Какой формат данных предполагается использовать для выходных результатов анализа?

2) Будет ли приложение обрабатывать данных на десктопах, или предназначаться для серверной обработки?

3) Каковы требования к операционному окружению приложения?

Также нерассмотренным остался вопрос о масштабируемости системы. Как и большинство систем обработки больших данных, она должна быть горизонтально расширяемой, за счет включения новых узлов — это требует параллелизации обработки как на уровне получения, так и анализа данных. Отчасти эта проблема решается встроенными средствами кластеризации, которую предоставит фреймворк [17] разработки или среда выполнения.

Заключение

Область обработки больших данных всё еще является достаточно сложной с практической точки зрения и требующей дополнительного изучения в теории.

В процессе исследования:

1) Описана специфика больших данных.

2) Сформулированы проблемы обработки больших веб- данных и способы их преодоления.

3) Предложена экспериментальная задача по интеллектуальной обработке веб-данных.

4) Разработана архитектура системы, предназначенная для решения задач обработки больших веб-данных.

Полученные результаты могут использоваться как в целях первичного изучения технологий обработки больших данных, так и в качестве основы разработки уже реального приложения для анализа веб-данных. Использование нейронных сетей и созданных универсальных классов-обработчиков делает созданную архитектуру гибкой и самообучаемой, а декларации классов и DDL-структура базы существенно упростят разработку программного кода.

Литература

1. Хашковский В. В., Шкурко А. Н. Современные подходы в организации систем обработки больших объемов данных // Известия Южного федерального университета. Технические науки. 2014. № 8 (157). С. 241–250.

2. Барсегян А.А., Куприянов М.С., Степаненко В.В., Холод И.И. Технологии анализа данных. Data Mining, Visual Mining, Text Mining, OLAP. 2 изд. СПб.: БХВ-Петербург, 2007. 384 с.

3. Марц Н., Уоррен Д. Большие данные. Принципы и практика построения масштабируемых систем обработки данных в реальном времени. М.: Вильямс, 2017. 368 с.

4. Кошик А. Веб-аналитика 2.0 на практике. Тонкости и лучшие методики. М.: Вильямс, 2014. 528 с.

5. Большие Данные [Электрон. ресурс] // Толковый словарь на Академике. 2014. Режим доступа: https://dic.academic.ru/dic.nsf/ruwiki/1422719 (дата обращения: 04.04.2019).

6. Кузнецов С.Д., Посконин А.В. Распределенные горизонтально масштабируемые решения для управления данными // Труды Института системного программирования РАН. 2013. № 24. С. 327–358 .

7. Флегонтов А.В., Фомин В.В. Система интеллектуальной обработки данных // Известия Российского государственного педагогического университета им. А.И. Герцена. 2013. № 1 (154). С. 41–48.

8. Mitrovic S. Specifics of the integration of business intelligence and Big Data technologies in the processes of economic analysis // Бизнес-информатика. 2017. № 4 (42). С. 40–46.

9. Филяк П.Ю., Байларли Э.Э.О., Растворов В.В., Старченко В.И. Инструментальные средства для использования big data и data mining в целях обеспечения информационной безопасности – подходы, опыт применения // Вестник Московского финансово-юридического университета. 2017. №2. С. 210-220.

10. Data Mining: что внутри. Habr. [Электрон. ресурс] Режим доступа: https://habr.com/ru/post/95209/ (Дата обращения: 04.04.2019).

11. Кадырова Н.О., Павлова Л., В. Эффективная методика обработки многоразмерных данных большого объема // Научно-технические ведомости Санкт-Петербургского государственного политехнического университета. Информатика. Телекоммуникации. Управление . 2012. №6 (162). С. 118–124.

12. Новиков Б.А., Графеева Н.Г., Михайлова Е.Г. BIG DATA: Новые задачи и современные подходы // Компьютерные инструменты в образовании. 2014. №4. С. 10-18.

13. Лосева Е.Д., Антамошкин А.Н. Алгоритм автоматизированного формирования ансамблей нейронных сетей для решения сложных задач интеллектуального анализа данных // Известия Тульского государственного университета. Технические науки. 2017. № 4. С. 234–243.

14. Автор. 2014.

15. Клеппман М. Высоконагруженные приложения. Программирование, масштабирование, поддержка. СПб: Питер, 2018. 740 с.

16. Флегонтов А. В., Фомин В. В. Система интеллектуальной обработки данных // Известия Российского государственного педагогического университета им. А.И. Герцена. 2013. №1 (154). С. 41–48.

17. Самарев Р.С. Обзор состояния области потоковой обработки данных // Труды Института системного программирования РАН. 2017. № 1 . С. 231–260.

References

1. KHashkovskiy V.V., Shkurko A.N. Modern approaches in the organization of systems for processing large volumes of data.Izvestiya Yuzhnogo federal'nogo universiteta. Tekhnicheskiye nauki = News of the Southern Federal University. Technical science. 2014; 8 (157): 241–250. (In Russ.)

2. Barsegyan A.A., Kupriyanov M.S., Stepanenko V.V., KHolod I.I. Tekhnologii analiza dannykh. Data Mining, Visual Mining, Text Mining, OLAP. 2 izd. = Data analysis technologies. Data Mining, Visual Mining, Text Mining, OLAP. 2nd ed. SPb.: BHV-Petersburg; 2007. 384 p. (In Russ.)

3. Marts N., Uorren D. Bol'shiye dannyye. Printsipy i praktika postroyeniya masshtabiruyemykh sistem obrabotki dannykh v real'nom vremeni = Big data. Principles and practice of building scalable data processing systems in real time. Moscow: Williams; 2017. 368 p. (In Russ.)

4. Koshik A. Veb-analitika 2.0 na praktike. Tonkosti i luchshiye metodiki = Web Analytics 2.0 in practice. Subtleties and best practices. Moscow: Williams; 2014. 528 p. (In Russ.)

5. Bol'shiye Dannyye = Big Data [Internet]. Tolkovyy slovar' na Akademike = Explanatory Dictionary on Academician. 2014. URL: https://dic.academic.ru/dic.nsf/ruwiki/1422719 (Cited: 04.04.2019). (In Russ.)

6. Kuznetsov P. D., Poskonin A. V. Distributed horizontally scalable solutions for data management. Trudy Instituta sistemnogo programmirovaniya RAN = Works of the Institute for System Programming of the Russian Academy of Sciences. 2013; 24: 327–358. (In Russ.)

7. Flegontov A. V., Fomin V. V. System of intellectual data processing.Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. A.I. Gertsena = A.I. Herzen News of the Russian State Pedagogical University. 2013; 1 (154): 41–48. (In Russ.)

8. Mitrovic P. Specifics of the integration of business intelligence and Big Data technologies in

the processes of economic analysis. Biznes-informatika. 2017; 4 (42): 40–46.

9. Filyak P.Yu., Baylarli E.E.O., Rastvorov V.V., Starchenko V.I. Tools for using big data and data mining in order to ensure information security - approaches, application experience. Vestnik Moskovskogo finansovo-yuridicheskogo universiteta = Bulletin of Moscow Financial and Law University. 2017; 2: 210–220. (In Russ.)

10. Data Mining: chto vnutri. Habr. = Data mining: what's inside. Habr. [Internet] URL: https://habr.com/ru/post/95209/ (Cited: 04.04.2019). (In Russ.)

11. Kadyrova N.O., Pavlova L.V. Effective methods for processing large-sized multidimensional data.Nauchno-tekhnicheskiye vedomosti Sankt-Peterburgskogo gosudarstvennogo politekhnicheskogo universiteta. Informatika. Telekommunikatsii. Upravleniye = Scientific and Technical Gazette of the St. Petersburg State Polytechnic University. Computer science. Telecommunications. Management. 2012; 6 (162): 118–124. (In Russ.)

12. Novikov B.A., Grafeyeva N.G., Mikhaylova E.G. BIG DATA: New tasks and modern approaches. Komp'yuternyye instrumenty v obrazovanii = Computer tools in education. 2014; 4: 10–18. (In Russ.)

13. Loseva E.D., Antamoshkin A.N. Algorithm for automated formation of neural network ensembles for solving complex data mining problems. Izvestiya Tul'skogo gosudarstvennogo universiteta. Tekhnicheskiye nauki. = News of Tula State University. Technical science. 2017; 4: 234–243. (In Russ.)

14. The author. 2014. (In Russ.)

15. Kleppman M. Vysokonagruzhennyye prilozheniya. Programmirovaniye, masshtabirovaniye, podderzhka = Highly loaded applications. Programming, scaling, support. Saint Petersburg: Peter; 2018. 740 p. (In Russ.)

16. Flegontov A. V., Fomin V. V. System of intellectual data processing. Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. A.I. Gertsena = News of the Herzen Russian State Pedagogical University. 2013; 1 (154): 41–48. (In Russ.)

17. Samarev R.S. Review of the state of the stream data processing area. Trudy Instituta sistemnogo programmirovaniya RAN = Proceedings of the Institute for System Programming of the Russian Academy of Sciences. 2017; 1 : 231–260. (In Russ.)

Сведения об авторах

Ксения Валериановна Мулюкова
Аспирант, Кафедра «Систем автоматического управления»
Инженерно-технологическая академия ЮФУ, Ростов-на-Дону, Россия
Эл. почта: mu.ksusha@yandex.ru

Виктор Михайлович Курейчик
Д.т.н., профессор, Кафедра «Систем автоматического управления»
Инженерно-технологическая академия ЮФУ, Ростов-на-Дону, Россия
Эл. почта: vmkureychik@sfedu.ru

Information about the authors

Ksenia V. Mulyukova
Postgraduate Student, Department of Automatic Control Systems
Engineering-Technological Academy of SFU, Rostov-on-Don, Russia
E-mail: mu.ksusha@yandex.ru

Victor M. Kureichik
Dr. Sci. (Engineering), Professor, Department of Automatic Control Systems
Engineering-Technological Academy of SFU, Rostov-on-Don, Russia
E-mail: vmkureychik@sfedu.ru