ISSN 2686-679X

ВЕСТНИК РГГУ

Серия
«Информатика.
Информационная безопасность.
Математика»

Научный журнал

RSUH/RGGU BULLETIN

"Information Science.
Information Security. Mathematics"
Series

Academic Journal

Основан в 2018 г.
Founded in 2018

4
2022

VESTNIK RGGU. Seriya “Informatica. Informacionnaya bezopasnost. Matematica”

RSUH/RGGU BULLETIN. “Information Science. Information Security. Mathematics” Series
Academic Journal

There are 4 issues of the printed version of the journal a year.

Founder and Publisher
Russian State University for the Humanities (RSUH)

RSUH/RGGU BULLETIN. “Information Science. Information Security. Mathematics” series is included: in the Russian Science Citation Index; in the List of leading scientific magazines journals and other editions for publishing PhD research findings peer-reviewed publications fall within the following research area:

20.00.00 Informatics

81.93.29 Information security, data protection

27.00.00 Mathematics

Objectives and areas of research

RSUH/RGGU BULLETIN. “Information Science. Information Security. Mathematics” series publishes the results of research by scientists from RSUH and other universities and other Russian and foreign academic institutions. The areas covered by contributions include theoretical and applied computer science, up-to-date IT, means and technologies of information protection and information security as well as the issues of theoretical and applied mathematics including analytical and imitation models of different processes and objects. Special emphasis is put on articles and reviews covering research in indicated directions in the areas of social and humanitarian problems and also issues of personnel training for these directions.

RSUH/RGGU BULLETIN. “Information Science. Information Security. Mathematics” series is registered by Federal Service for Supervision of Communications Information Technology and Mass Media. 25.05.2018, reg. No. FS77-72977

Editorial staff office: 6, Miusskaya sq., Moscow, Russia, 125047

tel: +7 (916) 250-90-85

e-mail: grnat@rambler.ru

ВЕСТНИК РГГУ. Серия «Информатика. Информационная безопасность. Математика»
Научный журнал

Выходит 4 номера печатной версии журнала в год.

Учредитель и издатель – Российский государственный гуманитарный университет (РГГУ)

ВЕСТНИК РГГУ, серия «Информатика. Информационная безопасность. Математика», включен: в систему Российского индекса научного цитирования (РИНЦ); в Перечень рецензируемых научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени кандидата наук, на соискание ученой степени доктора наук по следующим научным специальностям и соответствующим им отраслям науки:

20.00.00 Информатика

81.93.29 Информационная безопасность, защита информации

27.00.00 Математика

Цели и область

В журнале «Вестник РГГУ», серия «Информатика. Информационная безопасность. Математика», публикуются результаты научных исследований ученых и специалистов РГГУ, а также других университетов и научных учреждений России и зарубежных стран. Направления публикаций включают теоретическую и прикладную информатику, современные информационные технологии, методы, средства и технологии защиты информации и обеспечения информационной безопасности, а также проблемы теоретической и прикладной математики, включая разработку аналитических и имитационных моделей процессов и объектов различной природы. Особое внимание уделяется статьям и обзорам, посвященным исследованиям по указанным направлениям в области социальных и гуманитарных проблем, а также вопросам подготовки кадров по соответствующим специальностям для данных направлений.

ВЕСТНИК РГГУ, серия «Информатика. Информационная безопасность. Математика», зарегистрирован Федеральной службой по надзору в сфере связи, информационных технологий и массовых коммуникаций 25.05.2018 г., регистрационный номер ПИ № ФС77-72977.

Адрес редакции: 125047, Россия, Москва, Миусская пл., 6

Тел: +7 (916) 250-90-85

электронный адрес: grnat@rambler.ru

Founder and Publisher
Russian State University for the Humanities (RSUH)

Учредитель и издатель
Российский государственный гуманитарный университет (РГГУ)

Ответственный за выпуск:

Н.В. Гришина, кандидат технических наук, доцент, Российский государственный гуманитарный университет (РГГУ)

CONTENTS

Information Science

Mariya V. Vinogradova, Alexei A. Maksakov, Alexei E. Samokhvalov, Irina A. Smolyakova
Methodology for analyzing open data of selective federal statistical observation on the topic of information security of students 8

Evgenii K. Mazaishvili, Kirill L. Tassov
Method for clustering flying objects over a video stream based on neural networks 21

Information Security

Nataliya V. Grishina
Analysis of the dynamics of personal data leakage in the context of the implementation of the program "Digital Economy of the Russian Federation" 34

Tamara M. Volosatova, Anastasiya A. Kozar'
Analysis of the application of digital holography methods for information protection 44

Mathematics

Irina V. Gadolina
Application of fuzzy data in durability assessment tasks 59

Vyacheslav Yu. Sinitsyn, Valentina S. Kashparova
Frequency properties of the lexis of scientific texts and Zipf's laws of higher orders 75

СОДЕРЖАНИЕ

Информатика

Мария В. Виноградова, Алексей А. Максаков, Алексей Э. Самохвалов, Ирина А. Смолякова
Методика анализа открытых данных выборочного федерального статистического наблюдения по теме информационной безопасности учащихся 8

Евгений К. Мазайшвили, Кирилл Л. Тассов
Метод кластеризации летящих объектов по видеопотоку на основе нейронных сетей 21

Информационная безопасность

Наталия В. Гришина
Анализ динамики утечки персональных данных в условиях реализации программы «Цифровая экономика Российской Федерации» 34

Тамара М. Волосатова, Анастасия А. Козарь
Анализ применения методов цифровой голографии для защиты информации 44

Математика

Ирина В. Гадолина
Применение нечетких данных в задачах оценки долговечности 59

Вячеслав Ю. Синицын, Валентина С. Кашпарова
Частотные свойства лексики научных текстов и законы Ципфа высших порядков 75

УДК 004.85
DOI: 10.28995/2686-679X-2022-4-8-20

Методика анализа открытых данных выборочного федерального статистического наблюдения по теме информационной безопасности учащихся

Мария В. Виноградова
Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия, vinogradova.m@bmstu.ru

Алексей А. Максаков
Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия, 256m@mail.ru

Алексей Э. Самохвалов
Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия, Академия джаза, Москва, Россия, samox@bmstu.ru

Ирина А. Смолякова
Академия джаза, Москва, Россия, u.gudi@yandex.ru

Аннотация. Авторы статьи разработали методику обработки анкет выборочного федерального статистического наблюдения по вопросам использования населением информационных технологий и информационно-телекоммуникационных сетей, в том числе и по теме обеспечения родителями информационной безопасности детей. Выделены следующие этапы процесса: парсирование xml-файла, подготовка датафреймов, верификация данных, выбор показателей для исследования и анализ информации. Программная реализация методики выполнена на языке Python с применением библиотеки Pandas.

Выявленные при верификации и стандартизации набора данных ошибки имеют практическую значимость для составления рекомендаций Росстату по организации форматно-логического контроля при загрузке результатов анкетирования. Анализ итогов статистического наблюдения за

2021 г. показал, что родители не применяют программные средства защиты компьютеров и мобильных устройств, которые блокируют попытки посещения школьниками нежелательных сайтов и страниц в социальных сетях.

На основе авторского подхода предлагается разработать единую технологию аналитической обработки открытых данных о социологических опросах, которые проводит Федеральная служба государственной статистики. Методика может использоваться не только для научных аналитических исследований, но и в образовательных целях как практическое учебное пособие по курсам «Информационная аналитика», «Интеллектуальные информационные системы» и другим.

Ключевые слова: открытые данные, Росстат, информационная безопасность, информационная аналитика, Python, Pandas

Для цитирования: *Виноградова М.В., Максаков А.А., Самохвалов А.Э., Смолякова И.А.* Методика анализа открытых данных выборочного федерального статистического наблюдения по теме информационной безопасности учащихся // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2022. № 4. С. 8–20. DOI: 10.28995/2686-679X-2022-4-8-20

Methodology for analyzing open data of selective federal statistical observation on the topic of information security of students

Mariya V. Vinogradova
Bauman Moscow State Technical University, Moscow, Russia, vinogradova.m@bmstu.ru

Alexei A. Maksakov
Bauman Moscow State Technical University, Moscow, Russia, 256m@mail.ru

Alexei E. Samokhvalov
Bauman Moscow State Technical University, Moscow, Russia, Academy of Jazz, Moscow, Russia, samox@bmstu.ru

Irina A. Smolyakova
Academy of Jazz, Moscow, Russia, u.gudi@yandex.ru

Abstract. The authors of the article have developed a methodology for processing questionnaires of selective federal statistical observation on the use of information technologies and information and telecommunication

networks by the population, including on the topic of ensuring information security of children by parents. The following stages of the process are highlighted: parsing an xml file, preparing dataframes, verifying data, selecting indicators for research and analyzing information. The software implementation of the technique is performed in Python using the Pandas library.

The errors identified during the verification and standardization of the data set are of practical importance for making recommendations to Rosstat on the organization of format-logical control when uploading survey results. Analysis of the results of statistical observation for 2021 showed that parents do not use computer and mobile device protection software that blocks attempts by schoolchildren to visit unwanted sites and pages on social networks.

Based on the author's approach, it is proposed to develop a unified technology for analytical processing of open data on sociological surveys conducted by the Federal State Statistics Service. The methodology can be used not only for scientific analytical research, but also for educational purposes as a practical textbook for the courses "Information Analytics", "Intelligent Information Systems" and others.

Keywords: open data, Rosstat, information security, information analytics, Python, Pandas

For citation: Vinogradova, M.V., Maksakov, A.A., Samokhvalov, A.E. and Smolyakova, I.A. (2022), "Methodology for analyzing open data of selective federal statistical observation on the topic of information security of students", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 4, pp. 8–20, DOI: 10.28995/2686-679X-2022-4-8-20

Введение

Проблема обеспечения информационной безопасности молодежной среды появилась много лет назад. Еще в 2009 г. опрос школьников, проведенный Фондом развития Интернет, показал, что более 75% подростков осознают опасность встретиться с негативной, агрессивной, вредоносной информацией [Желтова 2013]. Сенатор и главный организатор проекта «Единый урок безопасности в сети Интернет» Л.Н. Бокова отметила, что «ситуация с каждым годом становится лучше, но во многих аспектах есть еще проблемы. Так, дети часто становятся жертвами буллинга, интернет-фишинга, других угроз, которые несет в себе информационное пространство. Пока еще наши дети не умеют защищать свои устройства и оберегать себя в цифровой жизни» [Бокова 2017].

В 2022 г. решением Межведомственной комиссии по профилактике правонарушений при Правительстве Москвы была утверждена Памятка для родителей несовершеннолетних, обучающихся в образовательных организациях. Документ содержит практические рекомендации по применению программных средств защиты и контроля доступа к ресурсам в сети Интернет.

На сайте Федеральной службы государственной статистики опубликован открытый набор данных «Использование населением информационных технологий и информационно-телекоммуникационных сетей, 2021 г.» (URL: https://rosstat.gov.ru/opendata). Файл формата XML [Ревунков, Гапанюк 2010] состоит из двух частей: ответы респондентов на вопросы анкеты № 1-ИТ «Анкета выборочного федерального статистического наблюдения по вопросам использования населением информационных технологий и информационно-телекоммуникационных сетей»[1] и справочник указанных в ней показателей.

Проведенное авторами исследование, которое легло в основу предложенной методики, состояло из следующих этапов:

1) скачивание, распаковка и расшифровка xml-файла, преобразование его в табличные формы (датафреймы);
2) верификация данных;
3) выбор показателей для исследования;
4) анализ данных.

Цель работы – создать единую методику аналитической обработки открытых данных о социологических опросах, проводимых Федеральной службой государственной статистики. Для демонстрации ее применения выбраны разделы, посвященные информационной безопасности школьников.

Актуальность работы

В действующей Концепции открытости федеральных органов исполнительной власти закреплен принцип вовлеченности гражданского общества – «обеспечение возможности участия граждан Российской Федерации, общественных объединений и предпринимательского сообщества в разработке и реализации управленческих решений с целью учета их мнений и приоритетов, а также создания

[1] Приказ Федеральной службы государственной статистики от 30.05.2022, № 404 // Информационно-правовой портал ГАРАНТ.РУ. URL: https://base.garant.ru/404796997/ (дата обращения 20 сентября 2022).

системы постоянного информирования и диалога»[2]. Предложенная методика позволяет широкому кругу специалистов анализировать опубликованные Росстатом наборы открытых данных. Возможна ее реализация в виде подсистемы репозитория научных информационных ресурсов, разрабатываемых «для публикации, поиска и просмотра научных материалов, а также обеспечения совместной работы над проектами» [Виноградова, Черненький 2017].

Загрузка открытого набора данных

Архивный файл открытого набора данных data-20220413-structure-20220413.zip размещен на официальном сайте Федеральной службы государственной статистики по адресу: https://rosstat.gov.ru/opendata/7708234640-IKT2021-v01. После извлечения из архива файл data-20220413-structure-20220413.xml преобразуется в табличный вид. С помощью программы, написанной на языке Python [Алмазбек, Абалиева, Шамырова 2021], создаются два pkl-файла: 1) результаты выборочного федерального статистического наблюдения по вопросам использования населением информационных технологий и информационно-телекоммуникационных сетей за 2021 г.; 2) справочник показателей статистического наблюдения.

Фрагмент программы парсирования файла с набором данных выглядит следующим образом:

```
import sys
import xml.etree.ElementTree as ET
import pandas as pd
print('Parsing...')
tree = ET.parse('data-20220413-structure-20220413.xml')
root = tree.getroot()
print('Reading...')
i = 0
data = []
cols = []
pr_cols = 0
data2 = []
```

[2] Распоряжение Правительства Российской Федерации от 30.01.2014, № 93-р «Об утверждении Концепции открытости федеральных органов исполнительной власти» // СПС «КонсультантПлюс». URL: https://www.consultant.ru/document/cons_doc_LAW_158273/ (дата обращения 20 сентября 2022).

```
        cols2 = []
        pr_cols2 = 0
        for elem in root: # Reports ReportHeaderValues
            for subelem in elem: # Report ReportHeaderValue
                if elem.tag == 'Reports' and pr_cols == 0 or elem.tag ==
    'ReportHeaderValues' and pr_cols2 == 0:
                    if elem.tag == 'Reports' and pr_cols == 0:
                        cols = [subelem2.tag for subelem2 in subelem]
                        pr_cols = 1
                    if elem.tag == 'ReportHeaderValues' and pr_cols2 == 0:
                        cols2 = [subelem2.tag for subelem2 in subelem]
                        pr_cols2 = 1
                new_row = [subelem2.text for subelem2 in subelem]
                if elem.tag == 'Reports':
                    data.append(new_row)
                if elem.tag == 'ReportHeaderValues':
                    data2.append(new_row)
        df = pd.DataFrame(data,None,cols)
        df.to_pickle('data.pkl')
        df2 = pd.DataFrame(data2,None,cols2)
        df2.to_pickle('data2.pkl')
```

Файл data.pkl содержит таблицу с результатами анкетирования (см. рис. 1).

	OKRYG_SV	SETKA	POSEL	GOD	VESA_DX	VESA_SVOD	RESP	\
0	41	76	1	2021	972.3745470000	650.7625000000000	21472110	
1	41	76	1	2021	972.3745470000	922.5833335000000	21462110	
2	41	76	1	2021	972.3745470000	613.6363635000000	21922110	

	NAS_POL	BB2	NAS_VOZR	NAS_VOZ2	NAS_VOZ4	NASOBRAZ	C1	CInt1	CInt2_1	CInt2_2	\
0	2	5	42	6	6	4	1	1	1	0	
1	2	5	63	10	10	3	1	1	1	0	
2	2	5	22	2	2	4	1	1	1	0	

	CInt2_3	CInt2_8	CInt2_9	CInt2_5	CInt2_6	CInt2_7	CInt3_1	CInt3_2	CInt3_3	\
0	0	8	0	0	0	0	1	2	0	
1	0	8	0	0	0	0	1	2	0	
2	0	8	0	0	0	0	1	2	0	

	CInt5_1	CInt5_2	CInt5_8	CInt5_9	CInt5_4	CInt5_5	CInt5_6	CInt5_10	CInt5_7	m1	\
0	0	0	0	0	0	0	0	0	0	1	
1	0	0	0	0	0	0	0	0	0	1	
2	0	0	0	0	0	0	0	0	0	1	

	C2	C4_1	C4_2	C4_3	C4_4	C4_5	C4_6	C4_7	C4_8	C4_14	C4_15	C4_16	C4_9	C4_10	\
0	1	0	0	0	0	0	0	0	0	0	0	0	0	10	
1	4	0	0	0	0	0	0	0	0	0	0	0	0	0	
2	1	0	0	0	0	0	0	0	0	0	0	0	0	10	

Рис. 1. Датафрейм с результатами анкетирования

Второй файл data2.pkl содержит таблицу с описанием показателей (рис. 2).

	Name	Value
0	A62	социальное положение по мнению респондента
1	BB2	общее количество членов домохозяйства постоянно проживающих в помещении
2	C1	наличие персонального компьютера в домашнем хозяйстве
3	C1_CInt1	наличие персонального компьютера и доступа в сеть интернет в домашнем хозяйстве
4	C2	использование компьютера в последний раз дома, на работе или в любом другом ...
5	C3	места использования персонального компьютера за последние 3 месяца
6	C3_1	дома
7	C3_2	на работе
8	C3_3	по месту учебы
9	C3_4	у друзей, знакомых
10	C3_5	в других местах
11	C4	действия, связанные с работой на персональном компьютере, выполненные за пос...
12	C4_1	работа с текстовым редактором
13	C4_10	иное
14	C4_11	отправка электронной почты с прикрепленным(-и) файлом(-ами) (например, с док...
15	C4_12	копирование или перемещение файла или папки
16	C4_13	использование инструмента копирования и вставки для дублирования или перемещ...
17	C4_2	работа с электронными таблицами (например, использование таких функций работ...
18	C4_3	использование программ для редактирования фото-, видео- и аудио- файлов)

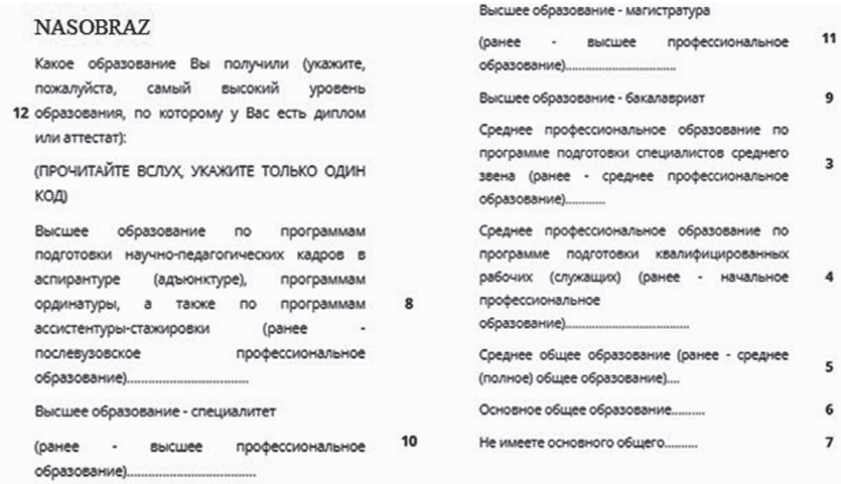
Рис. 2. Датафрейм с показателями анкетирования

Верификация данных

Процесс верификации данных авторы предложили разделить на следующие этапы:

1) проверка соответствия шифра и значений показателя (в первой части файла с открытыми данными) его типу и расшифровке (во второй части файла);
2) поиск пропущенных и некорректных значений (не соответствующих области допустимых значений реквизитов формы № 1-ИТ);
3) стандартизация данных.

При расшифровке первого файла выявились пропуски в файле показателей анкетирования, например отсутствует описание пункта анкеты “NASOBRAZ” и его допустимых значений. Для заполнения пропусков в справочнике выполнена загрузка требуемых данных из Приказа Росстата от 30.07.2021, № 457 «Об утверждении форм федерального статистического наблюдения для организации федерального статистического наблюдения за численностью, условиями и оплатой труда работников, потребностью организаций в работниках по профессиональным группам, составом кадров государственной гражданской и муниципальной службы» (рис. 3).



NASOBRAZ

12 Какое образование Вы получили (укажите, пожалуйста, самый высокий уровень образования, по которому у Вас есть диплом или аттестат):

(ПРОЧИТАЙТЕ ВСЛУХ, УКАЖИТЕ ТОЛЬКО ОДИН КОД)

Высшее образование по программам подготовки научно-педагогических кадров в аспирантуре (адъюнктуре), программам ординатуры, а также по программам ассистентуры-стажировки (ранее - послевузовское профессиональное образование)..... 8

Высшее образование - специалитет

(ранее - высшее профессиональное образование)..... 10

Высшее образование - магистратура

(ранее - высшее профессиональное образование)..... 11

Высшее образование - бакалавриат 9

Среднее профессиональное образование по программе подготовки специалистов среднего звена (ранее - среднее профессиональное образование)..... 3

Среднее профессиональное образование по программе подготовки квалифицированных рабочих (служащих) (ранее - начальное профессиональное образование)..... 4

Среднее общее образование (ранее - среднее (полное) общее образование)..... 5

Основное общее образование..... 6

Не имеете основного общего..... 7

Рис. 3. Дополнение справочника показателей анкетирования

Анализ указанных в анкетах значений выявил их неоднозначность. На вопрос «Применяли ли Вы за последние 12 месяцев... средства родительского контроля или фильтрации Интернет-ресурсов» (показатель Int7_3) записаны коды ответов: 0, 3 и 4. Согласно инструкции по заполнению формы № 1-ИТ допустимыми являются только коды 0 (нет) и 3 (да). Код ответа 4 относится к показателю Int7_4. Аналогичная неоднозначность обнаружилась при обработке вопроса «Укажите, пожалуйста, с какими проблемами (угрозами) информационной безопасности Вы сталкивались за последние 12 месяцев... Int6_4 Посещение детьми нежелательных сайтов, контакты детей с потенциально опасными людьми через сеть Интернет». Допустимыми являются только коды 0 (нет) и 4 (да). Код ответа 5 относится к показателю Int6_5.

Ошибки заполнения анкет наглядно продемонстрированы на точечной диаграмме (рис. 4).

Использование Росстатом разных целочисленных кодов в качестве значений показателей требует применения методов их стандартизации. Авторы предложили перекодировку ответов: 0 – «нет», 1 – «да».

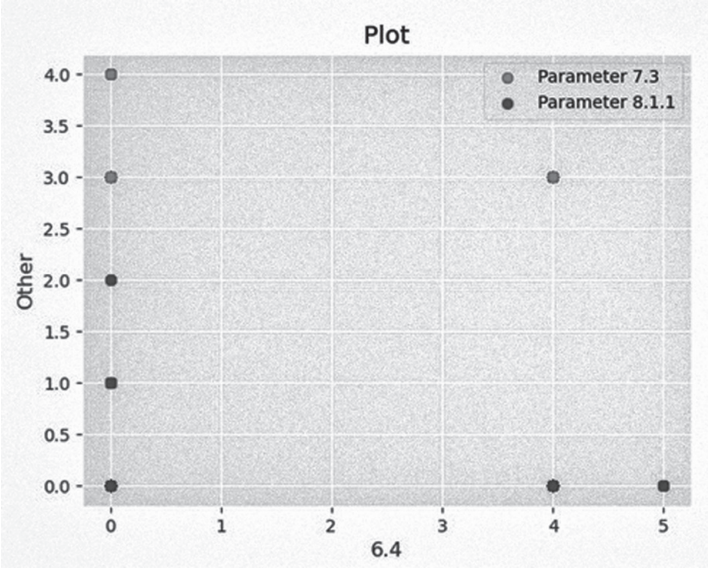


Рис. 4. Верификация показателей Int6_4 и Int7_3

Выбор показателей для исследования и анализ данных

Для исследования проблемы защиты школьников от несанкционированного контента в сети Интернет определены соответствующие показатели анкетирования из открытого набора данных. Для демонстрационного примера (для наглядности) отберем только следующие:

Int6_4 Укажите, пожалуйста, с какими проблемами (угрозами) информационной безопасности Вы сталкивались за последние 12 месяцев... Посещение детьми нежелательных сайтов, контакты детей с потенциально опасными людьми через сеть Интернет;

Int7_3 Применяли ли Вы за последние 12 месяцев следующие средства защиты информации при использовании сети Интернет? Средства родительского контроля или фильтрации Интернет-ресурсов;

Int8_1_1 Назовите причины, из-за которых Вы не пользовались сетью Интернет... Стремление ограничить доступ детей к нежелательной информации и программам.

С целью выявления зависимостей между показателями, определения их силы и характеристик авторы составили матрицу корреля-

ции. Фрагмент программы построения корреляционной матрицы выглядит следующим образом:

```
corr_matrix = df[['Int6_4','Int7_3','Int8_1_1']].astype(float).corr()
print(corr_matrix, file=file1)
```

Результат построения изображен на рис. 5.

	Int6_4	Int7_3	Int8_1_1
Int6_4	1.000000	0.172960	-0.001252
Int7_3	0.172960	1.000000	-0.002328
Int8_1_1	-0.001252	-0.002328	1.000000

Рис. 5. Корреляционная матрица

К удивлению авторов, между «показателем-угрозой» и «показателями-защитой» не выявились значимые корреляционные связи. Количественный анализ также подтвердил их отсутствие. Из 154021 опрошенных: 151135 – не сталкивались с проблемой посещения детьми нежелательных сайтов, их опасных контактов через сеть Интернет и не пользуются программными средствами защиты; 2180 – еще не сталкивались с проблемой, но установили программные средства защиты; 470 – столкнулись с проблемой, но не установили программные средства защиты; 236 – столкнулись с проблемой и установили программные средства защиты.

Таким образом, обработка итогов выборочного федерального статистического наблюдения по вопросам использования населением информационных технологий и информационно-телекоммуникационных сетей за 2021 г. показала, что родители в большинстве своем игнорируют угрозу посещения школьниками нежелательных сайтов, их контактов с потенциально опасными людьми через сеть Интернет. Этот вывод доказывает своевременность принятых Межведомственной комиссией по профилактике правонарушений при Правительстве Москвы мероприятий по информированию родителей о программных средствах защиты компьютеров и мобильных устройств учащихся.

Заключение

Предложенная авторами методика обработки открытых данных Росстата, продемонстрированная на примере анализа показателей информационной безопасности учащихся, может исполь-

зоваться для научных исследований, а также как практическое учебное пособие по дисциплинам «Информационная аналитика», «Интеллектуальные информационные системы» [Гапанюк, Ревунков, Спиридонов, Терехов, Черненький 2016] в высших учебных заведениях.

Литература

Алмазбек, Абалиева, Шамырова 2021 – *Алмазбек М., Абалиева А.Д., Шамырова Д.Р.* Принципы работы с библиотеками Python в прикладных исследованиях // Современные проблемы механики. 2021. № 43 (1). С. 120–131.

Бокова 2017 – *Бокова Л.Н.* За рубежом поражены масштабностью и эффективностью нашего Единого урока и перенимают российский опыт // Дети в информационном обществе. 2017. № 26. С. 4–9.

Виноградова, Черненький 2017 – *Виноградова М.В., Черненький М.В.* Концепция создания репозитория научных информационных ресурсов // Динамика сложных систем – XXI век. 2017. № 4. С. 38–45.

Гапанюк, Ревунков, Спиридонов, Терехов, Черненький 2016 – *Гапанюк Ю.Е., Ревунков Г.И., Спиридонов С.Б., Терехов В.И., Черненький В.М.* Концепция преподавания курсов по гибридным интеллектуальным информационным системам // Управление качеством инженерного образования. Возможности вузов и потребности промышленности: Тезисы докладов второй международной научно-практической конференции, Москва, 23–25 июня 2016 г. М.: Московский государственный технический ун-т им. Н.Э. Баумана, 2016. С. 165.

Желтова 2013 – *Желтова И.А.* Сетевая культура и информационная безопасность школьников в интернет-пространстве // Вестник Северо-Восточного государственного ун-та. 2013. № 19. С. 106–108.

Ревунков, Гапанюк 2010 – *Ревунков Г.И., Гапанюк Ю.Е.* Введение в XML-технологии. М.: МГТУ им. Н.Э. Баумана, 2010.

References

Almazbek, M., Abalieva, A.D. and Shamyrova, D.R. (2021), “Principles of Python libraries in applied research”, *Sovremennye problemy mekhaniki*, no. 43 (1), pp. 120–131.

Bokova, L.N. (2017), “Abroad they are amazed at the scale and effectiveness of our Single Lesson and are adopting Russia’s experience”, *Deti v informacionnom obshchestve*, no. 26, pp. 4–9.

Gapanyuk, Yu.E., Revunkov, G.I., Spiridonov, S.B., Terekhov, V.I. and Chernen’kii, V.M. (2016), “The concept of teaching courses on hybrid intelligent information systems”, *Upravlenie kachestvom inzhenernogo obrazovaniya. Vozmozhnosti VUZov i potrebnosti*

promyshlennosti: Tezisy dokladov vtoroi mezhdunarodnoi nauchno-prakticheskoi konferencii [Engineering Education Quality Management. Opportunities for Higher Education Institutions and Industry Needs. Abstracts of the 2nd International Scientific and Practical Conference], Moscow, 23–25 June 2016, Bauman Moscow State Technical University, Moscow, Russia, pp. 165.

Revunkov, G.I. and Gapanyuk, Yu.E. (2010), Vvedenie v XML-tekhnologii [Introduction into XML Technologies], Bauman Moscow State Technical University, Moscow, Russia.

Vinogradova, M.V. and Chernen'kii, M.V. (2017), "Concept of Scientific Information Resources Repository", *Dinamika slozhnykh sistem – XXI vek*, no. 4, pp. 38–45.

Zheltova, I.A. (2013), "Network culture and information security of schoolchildren in the Internet sphere", *Vestnik Severo-Vostochnogo gosudarstvennogo universiteta*, no. 19, pp. 106–108.

Информация об авторах

Мария В. Виноградова, кандидат технических наук, доцент, Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия; 105005, Россия, Москва, 2-я Бауманская ул., д. 5; vinogradova.m@bmstu.ru

Алексей А. Максаков, кандидат технических наук, Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия; 105005, Россия, Москва, 2-я Бауманская ул., д. 5; 256m@mail.ru

Алексей Э. Самохвалов, кандидат экономических наук, Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия; 105005, Россия, Москва, 2-я Бауманская ул., д. 5;

Академия джаза, Москва, Россия; 123022, Россия, Москва, Трёхгорный Вал, д. 2-4, стр. 1; samox@bmstu.ru

Ирина А. Смолякова, Академия джаза, Москва, Россия; 123022, Россия, Москва, Трёхгорный Вал, д. 2-4, стр. 1; u.gudi@yandex.ru

Information about the authors

Mariya V. Vinogradova, Cand. of Sci. (Computer Sciences), associate professor, Bauman Moscow State Technical University, Moscow, Russia; bld. 5, 2nd Bauman Str., Moscow, Russia, 105005; vinogradova.m@bmstu.ru

Alexei A. Maksakov, Cand. of Sci. (Computer Sciences), Bauman Moscow State Technical University, Moscow, Russia; bld. 5, 2nd Bauman Str., Moscow, Russia, 105005; 256m@mail.ru

Alexei E. Samokhvalov, Cand. of Sci. (Economics), Bauman Moscow State Technical University, Moscow, Russia; bld. 5, 2nd Bauman Str., Moscow, 105005, Russia;

Academy of Jazz, Moscow, Russia; bld. 2-4, Trekhgornyi Val Str., Moscow, Russia, 123022; samox@bmstu.ru

Irina A. Smolyakova, Academy of Jazz, Moscow, Russia; bld. 2-4, Trekhgornyi Val Str., Moscow, Russia, 123022; u.gudi@yandex.ru

УДК 004.93
DOI: 10.28995/2686-679X-2022-4-21-33

Метод кластеризации летящих объектов по видеопотоку на основе нейронных сетей

Евгений К. Мазайшвили
Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия, evgenij997@yandex.ru

Кирилл Л. Тассов
Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия, ktassov@policesoft.ru

Аннотация. В данной работе представлен метод кластеризации нейронной сетью летящих на фоне неба объектов. Данный метод предлагается использовать для задач обнаружения, отнесения к определенному кластеру, а также быстрого определения опасности объекта, попавшего в зону мониторинга. Для решения данной задачи используется нейронная сеть, состоящая из комбинации сверточной сети (определяющей статические признаки объекта, например форму и цвет) и рекуррентной сети (определяющей динамические признаки объекта, например движение лопастей пропеллера, взмахи крыльев), результаты работы которых подаются на вход карты Кохонена, где, собственно, и происходит кластеризация объекта. Показаны примеры входных и выходных данных метода, а также способы интерпретации выходных данных (карты Кохонена) для быстрого определения опасности летящего объекта. Приведены способы предварительной обработки данных, поступающих на вход нейронной сети. Также изложены основы архитектуры рекуррентных и сверточных нейросетей, используемых в представленном методе. Описан алгоритм и набор данных, использованных при обучении нейронной сети. Проведены исследования и представлены результаты работы метода. Дана оценка эффективности применения данного метода для видеоаналитики летящих объектов, взятых из видеопотока.

Ключевые слова: Кохонен, нейронные сети, машинное обучение, распознавание изображений, похожие модели, кластеризация

Для цитирования: Мазайшвили Е.К., Тассов К.Л. Метод кластеризации летящих объектов по видеопотоку на основе нейронных сетей // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2022. № 4. С. 21–33. DOI: 10.28995/2686-679X-2022-4-21-33

Method for clustering flying objects over a video stream based on neural networks

Evgenii K. Mazaishvili
Bauman Moscow State Technical University, Moscow, Russia, evgenij997@yandex.ru

Kirill L. Tassov
Bauman Moscow State Technical University, Moscow, Russia, ktassov@policesoft.ru

Abstract. The paper presents a method of using a neural network for clustering objects flying against the sky background. Such a method is proposed to be used for the tasks of the detection or assignment to a certain cluster, as well as for quickly determining the danger of an object flying through monitoring zone. To solve this problem, a neural network is used, consisting of a convolutional network (determining the static features of an object, for example, shape and color) combined with a recurrent network (determining the dynamic features of an object, for example, the movement of propeller blades, flapping wings), the results of which are fed to the Kohonen map, where the clustering of the object occurs. Examples of input and output data of the method are shown, as well as ways to interpret the output data (Kohonen map) to quickly determine the danger of a flying object. Methods for preliminary processing of input data are given. The article also outlines the basics for the architecture of recurrent and convolutional neural networks used in the presented method. A training algorithm and a set of data used in neural network training are described. Studies have been carried out and the results of the method are presented. An estimate of the effectiveness of the method for video analytics of flying objects taken from a video stream is given.

Keywords: Kohonen, neural networks, machine learning, image recognition, similar models, clustering

For citation: Mazaishvili, E.K. and Tassov, K.L. (2022), "Method for clustering flying objects over a video stream based on neural networks", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 4, pp. 21–33, DOI: 10.28995/2686-679X-2022-4-21-33

Дроны – беспилотные летательные аппараты (БЛА) могут представлять серьезную угрозу для гражданских и военных объектов. Небольшие, легкие модели дронов широко доступны [Юферов 2021] и могут использоваться противником в различных целях: сбор разведданных, промышленный шпионаж, постановка радиопомех, доставка мелких грузов, целеуказание и нанесение ударов при террористических актах.

Основным материалом БЛА является пластик, который большей частью радиопрозрачен или обладает настолько малым отражением, что не виден для радаров [Защита от дронов, которая есть только у России]. Дорогостоящие комплексы, осуществляющие радиоэлектронное обнаружение и требующие большого штата сотрудников для небольших дешевых в изготовлении дронов, легко заменимы автономными камерами, не требующими участия человека.

В настоящее время самый доступный способ обнаружения БЛА – система из обзорных и поворотных видеокамер, действующих в паре. Стационарно закрепленная обзорная камера с широкоугольным объективом просматривает значительную область неба и распознает на ней возможное движение. При получении сигнала от обзорной камеры об обнаруженном движении поворотная камера наводится на источник движения и делает увеличенную детализированную видеозапись летящего объекта.

После получения детализированной видеозаписи движущегося объекта первичной задачей остается его распознавание, а именно является ли данный объект БЛА, представляющим угрозу. Зачастую не требуется четкой классификации объекта как дрона определенной модели, достаточно определить принадлежность объекта к «дронам», тем самым «отделив» похожие на БЛА видеозаписи от видеозаписей других объектов в небе (птицы, самолеты).

Целью данной работы является кластеризация видеозаписей неизвестных летящих объектов на фоне неба и их наглядное представление на двумерной сетке.

Кластеризация изображений считается хорошо изученной областью машинного обучения с существующими решениями по кластеризации изображений и видео на основе предобученных нейросетей [Пастухов, Прокофьев 2016] [Кузнецов, Семенов, Матросова 2019]. При всем многообразии классических методов кластеризации (K-Means, DBSCAN) они плохо подходят для использования при большом входном векторе параметров, а K-Means при этом требует знания точного количества кластеров. Частично были опробованы методы Mean Shift и Гауссовой смеси, которые тоже не дали значительных результатов.

Задача кластеризации была решена с применением карты Кохонена, на которой сразу видно, насколько пойманный в поле зрения камеры летящий объект близок к «дронам», «птицам» и т. п.

Однако для выполнения кластеризации на карте Кохонена необходимо пространство входных изображений сначала перевести в вектор признаков формы и вектор признаков движения.

Для выделения из изображения признаков формы лучше всего подходят сверточные нейросети. Они получили свое название по используемой операции свертки: каждый фрагмент изображения умножается поэлементно на матрицу (ядро свертки). Результат суммируется и записывается в соответствующий фрагмент выходного изображения.

Используемая сверточная сеть является широко известной архитектурой сети для выделения признаков изображений VGG16 [Simonyan, Zisserman 2015].

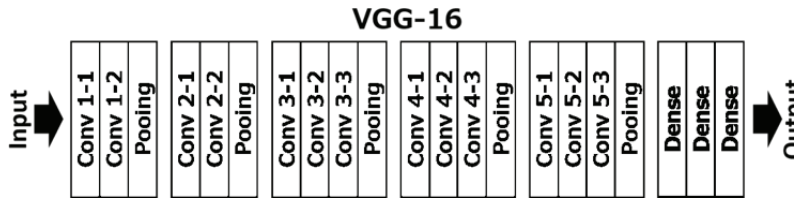


Рис. 1. Сверточная сеть VGG16

Поскольку основная задача сверточных сетей – не кластеризация, а классификация, и на выходном слое получаем класс объекта, то для получения необходимого нам вектора признаков формы объекта последний слой сверточной сети был удален. Сформированное на предпоследнем слое сверточной нейросети линейно разделимое пространство признаков формы объекта идеально подходит в качестве входного вектора признаков для сети Кохонена.

Перед тем, как подать изображение на сверточную нейросеть, программа выполняла его предобработку. Поскольку информация о цвете объекта не особо важна для отнесения его к определенному кластеру, изображение переводилось в оттенки серого, масштабировалось до размеров 200 на 200 пикселей и обрабатывалась фильтром Собеля для получения контуров объекта.

Фильтр Собеля выполняет роль «нулевого» сверточного слоя, в котором ядро задано вручную. Такая обработка позволяет избавиться от неконтрастных деталей на изображении – шума.

В операторе Собеля применяется следующая плавающая матрица:

$$P = \begin{pmatrix} P_{11} & P_{12} & P_{13} \\ P_{21} & P_{22} & P_{23} \\ P_{31} & P_{32} & P_{33} \end{pmatrix} \quad (1)$$

Данная матрица попиксельно накладывается на изображение таким образом, чтобы P_{22} поочередно совмещался со всеми пикселями изображения (там, где это возможно) [Алексанин 2017, с. 46]. Для расчета перепада яркости в пикселе с координатами (x, y) используется следующий градиентный оператор Собеля:

$$\begin{aligned} G_x &= (P_{31} + 2P_{32} + P_{33}) - (P_{11} + 2P_{12} + P_{13}) \\ G_y &= (P_{13} + 2P_{23} + P_{33}) - (P_{11} + 2P_{21} + P_{31}) \end{aligned} \quad (2)$$

На основании этих данных вычисляется перепад яркости:

$$G = \sqrt{{G_x}^2 - {G_y}^2} \quad (3)$$

Рис. 2. Пример предобработанного фильтром Собеля изображения для сверточной сети

В то время как сверточная нейронная сеть выделяет визуальные признаки изображения, для получения вектора признаков движения объекта («взмахи крыльев», «вращение пропеллеров») была использована рекуррентная нейросеть, у которой был также удален последний слой.

Рекуррентные сети работают с данными, представляющими собой последовательность элементов, упорядоченных по времени (предложения из слов, видео из кадров).

Использовалась рекуррентная сеть следующей структуры:

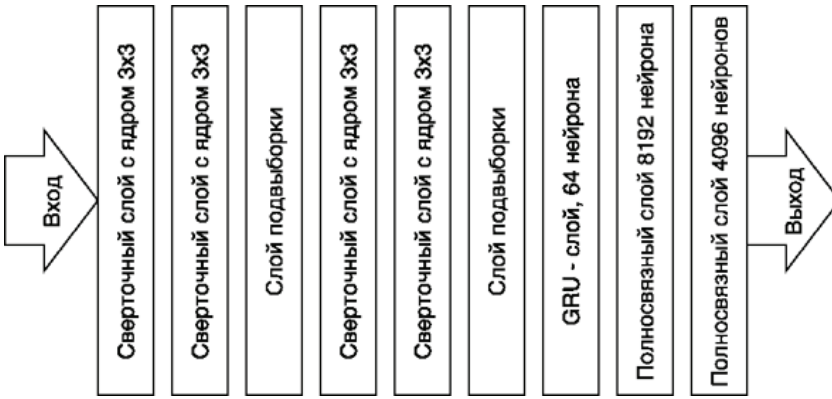


Рис. 3. Структура рекуррентной нейросети

Непосредственно рекуррентный слой использованной сети представлен GRU-нейронами.

GRU (Gated recurrent unit) – рекуррентный нейрон, пропускающий через себя последовательность элементов, поданную на вход, на каждом элементе обновляя свое внутреннее состояние [Lipton, Berkowitz, Elkan 2015].

Исходное видео разделялось на серии по 5 кадров, и выполнялась их предобработка для рекуррентной сети – каждый кадр масштабировался до размера 200 × 200, и вычислялся оптический поток относительно предыдущего кадра. Для вычисления оптического потока был использован метод Фарнебека (швед. Gunnar Farnebäck) [Farnebäck 2003]. Метод работает на предположении, что яркость двигающихся пикселей (I) константна, а пиксель сдвигается на расстояние dx, dy за время t. Метод вычисляет величину и угол (u, v) движения каждого пикселя на изображении.

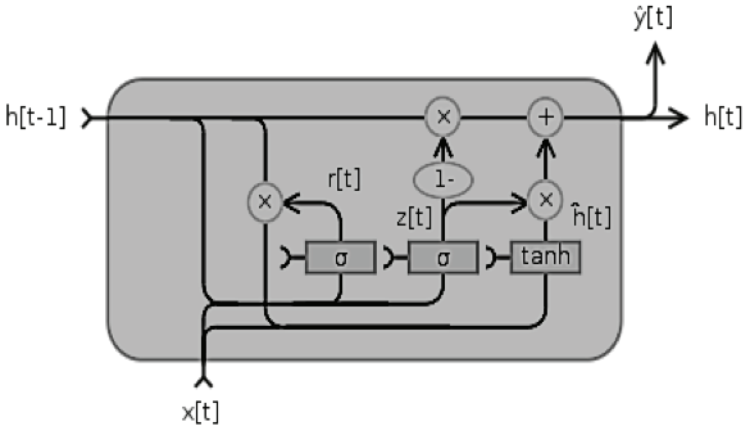


Рис. 4. Схема GRU нейрона
h[t−1], h[t] – состояние скрытого нейрона, x[t] – входной сигнал,
r[t] реализует вентиль сбрасывания, z[t] – вентиль обновления

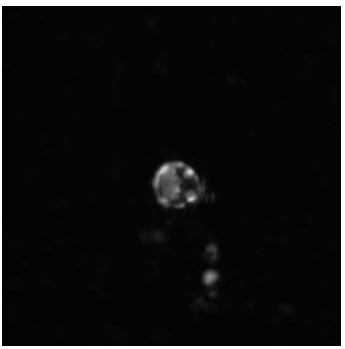


Рис. 5. Пример предобработанного изображения
для рекуррентной сети (светлые участки – наличие движения,
темные – движения нет)

В результате обработки видеозаписи сверточной и рекуррентной нейросетью получили два вектора признаков объекта: визуальных признаков формы (очертаний) объекта и признаков характера его движения. Данные векторы признаков конкатенировались вместе и становились входом сети Кохонена, непосредственно выполняющей кластеризацию и определяющей размещение объекта на двумерной карте.

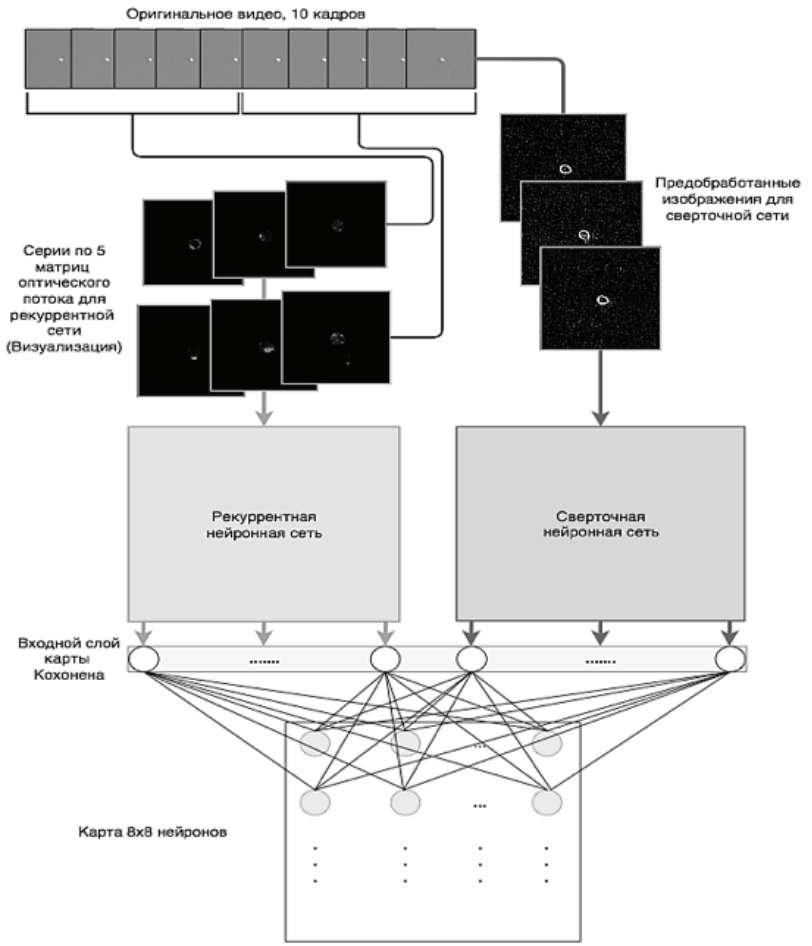


Рис. 6. Структура нейросети

Ключевой идеей данной работы было объединение картой Кохонена двух векторов признаков: вектора движения летящего объекта, полученного из рекуррентной нейросети, и вектора формы, полученного из сверточной.

Обучение нейросети проводилось в 3 этапа. Сначала сверточная нейросеть обучалась на 3-х классах изображений: дроны, птицы и воздушные шары. После обучения последний слой сети удалялся.

Рис. 7.1, 7.2, 7.3. Примеры изображений
из каждого класса (дрон, птица, воздушный шар)
для обучения сверточной и рекуррентной сетей.

Затем подобным образом обучалась рекуррентная сеть, с той разницей, что бралось не одно изображение объекта, а последовательность из 5 кадров видео. Последний слой рекуррентной сети также удалялся.

Полученные на предпоследних слоях сверточной и рекуррентной сетей векторы признаков, соответственно формы и движения, конкатенировались в один вектор, который становился входом карты Кохонена.

Самоорганизующаяся карта Кохонена – тип сетей, не требующих размеченного набора данных, которые сами сопоставляют входным данным кластеры исходя из близости признаков.

Сеть Кохонена [Юнусова, Магсумова 2019] состоит из двух слоев – входного, который не производит вычислений, а только распределяет входные данные на следующий слой, и выходного, где нейроны организованы в прямоугольную карту. Каждый нейрон входного слоя связан с каждым нейроном выходного слоя.

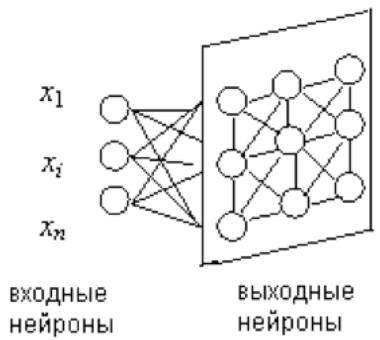


Рис. 8. Схема слоев карты Кохонена

Выходы обрабатываются по принципу «победитель забирает все», т. е. нейрон с наибольшим значением выхода получает значение 1, а все остальные нейроны – 0. Из этого принципа следует, что в результате работы сети активируется только один выходной нейрон. Причем близкие входные вектора должны активировать близкие нейроны на выходном слое.

Используемая в данной работе состоящая из 64 нейронов на выходном слое карта Кохонена (8 × 8) позволила объединить сверточную и рекуррентную сети, решающие задачу классификации, и обобщить их результаты в кластеры. В качестве входа в карту Кохонена использовался вектор длины 4096 + 64 (конкатенация 4096 выходных нейронов сверточной и 64 нейронов рекуррентной сетей).

Входом программного продукта являются видеозаписи летящих объектов. В примере на рис. 9 входом служили три видео: с воздушными шарами, с коптерами и с птицами. На выходе получаем представленные на рис. 9 изображения, где кадры входных видео (слева) либо цифры, соответствующие номерам видео (справа), распределены по сетке 8 × 8 карты Кохонена.

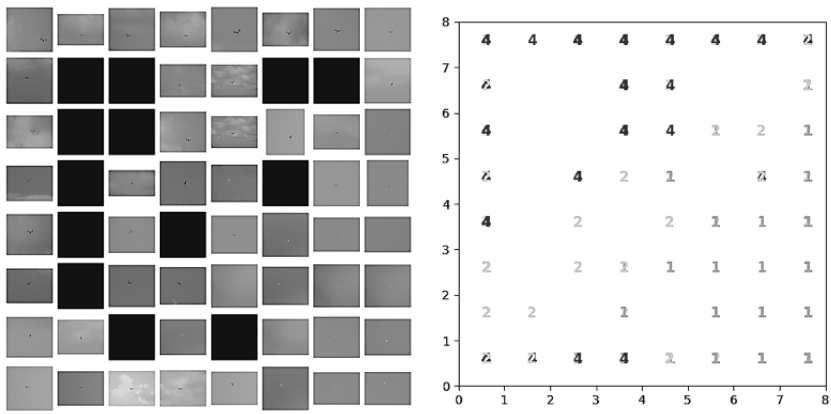


Рис. 9. Визуализация карты Кохонена
с распределением объектов по 64 нейронам.
Цифра на карте справа соответствует номеру видео.
1 – видео с воздушными шарами, 2 – видео с птицами,
4 – видео с дронами

Как можно видеть, в левом верхнем углу карты расположен кластер дронов, в нижнем правом – кластер воздушных шаров, между ними, диагонально, расположен кластер птиц.

Таким образом, готовый программный продукт был составлен из 4-х модулей: разбивка видео на кадры, предобработка отдельных кадров видео (фильтр Собеля – рис. 2, матрица оптического потока – рис. 5), обработка изображений обученной нейросетью (примеры для обучения представлены на рис. 7.1, рис. 7.2, рис. 7.3), визуализация выходных данных на карте Кохонена 8 × 8 (рис. 9). IDEF0 диаграмма программного продукта представлена на рис. 10.

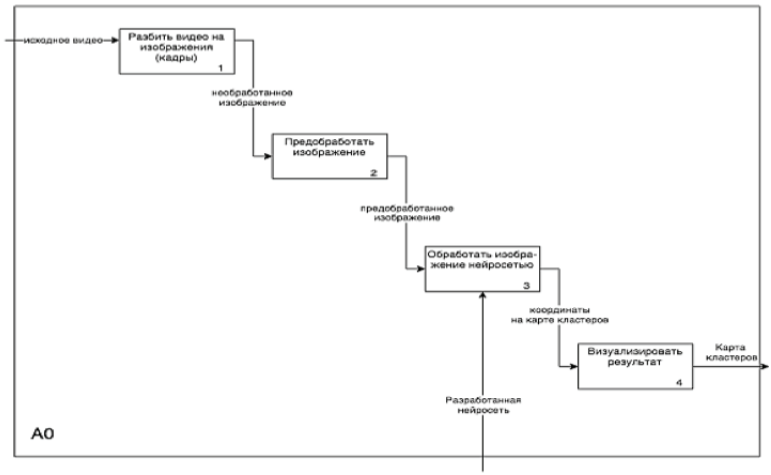


Рис. 10. IDEF0 диаграмма разработанной программы

При постоянном использовании данного ПО на двумерной карте Кохонена образуются несколько локализованных кластеров, соответствующих определенным типам летящих объектов, в данном случае дроны, птицы, воздушные шары. При загрузке в программу нового видео ПО располагает его кадры в те области карты Кохонена, которые соответствуют его кластеру. Применение карты Кохонена как выходного слоя позволяет визуально распознавать представляющие угрозу летящие объекты.

Литература

Алексанин 2017 – *Алексанин С.А.* Проектирование методов автоматизированной обработки изображений для систем дефектоскопии: Дис. ... канд. техн. наук. СПб.: С.-Петерб. нац. исслед. ун-т информац. технологий, механики и оптики, 2017. 125 с.

Защита от дронов, которая есть только у России – Защита от дронов, которая есть только у России, 2019. URL: https://zen.yandex.ru/media/survival_task/zascita-ot-dronov-kotoraia-est-tolko-u-rossii-5da6be220ce57b00ad50240b (дата обращения 10 ноября 2022).

Кузнецов, Семенов, Матросова 2019 – *Кузнецов А.С., Семенов Е.Ю., Матросова Л.Д.* Кластеризация изображений при использовании предобученных нейронных сетей // International Journal of Open Information Technologies. 2019. № 7. С. 42–47.

Пастухов, Прокофьев 2016 – *Пастухов А.А, Прокофьев А.А.* Применение самоорганизующихся карт Кохонена для формирования представительской выборки при обучении многослойного персептрона // Научно-технические ведомости СПб. ГПУ. Физико-математические науки. 2016. № 2. С. 242.

Юнусова, Магсумова 2019 – *Юнусова Л.Р., Магсумова А.Р.* Кластеризация с помощью нейронных сетей и поиск зависимостей // Наука, образование и культура. 2019. № 7 (41). С. 18–20.

Юферов 2021 – *Юферов С.* Новые российские средства борьбы с беспилотниками, 2021 // Военное обозрение. URL: https://topwar.ru/181712-novye-rossijskie-sredstva-borby-s-bespilotnikami.html (дата обращения 10 ноября 2022).

Farnebäck 2003 – *Farnebäck G.* Two-Frame Motion Estimation Based on Polynomial Expansion // 13th Scandinavian Conference on Image Analysis 2003. Heidelberg: Springer, 2003. P. 363–370. (LNCS, vol. 2749).

Lipton, Berkowitz, Elkan 2015 – *Lipton Z., Berkowitz J., Elkan C.A.* Critical Review of Recurrent Neural Networks for Sequence Learning, 2015. URL: https://arxiv.org/abs/1506.00019 (дата обращения 10 ноября 2022).

Simonyan, Zisserman 2015 – *Simonyan K., Zisserman A.* Very Deep Convolutional Networks for Large-Scale Image Recognition // Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7–9, 2015. Stroudsburg, PA: Association for Computational Linguistics, 2015.

References

Aleksanin, S.A. (2017), *Proektirovanie metodov avtomatizirovannoi obrabotki izobrazhenii dlya sistem defektoskopii* [Designing methods of automated image processing for flaw detection systems], Ph.D. Thesis, Universitet ITMO, SPb, Saint Petersburg, Russia, 125 p.

Farnebäck, G. (2003), "Two-Frame Motion Estimation Based on Polynomial Expansion", *13th Scandinavian Conference on Image Analysis 2003*, Springer, Heidelberg, Germany, p. 363–370 (LNCS, vol. 2749).

Kuznetsov, A.S., Semenov, E.Yu. and Matrosova, L.D. (2019), "Image clustering when using pre-trained neural networks", *International Journal of Open Information Technologies*, vol. 7, pp. 42–47.

Lipton, Z., Berkowitz, J., and Elkan, C.A. (2015), "Critical Review of Recurrent Neural Networks for Sequence Learning", 2015, available at: https://arxiv.org/abs/1506.00019 (Accessed 10 November 2020).

Pastukhov, A.A. and Prokof'ev, A.A. (2016), "Applying Self-Organizing Kohonen Maps to Form Representative Sampling in Multilayer Perspectron Learning", *Nauchno-tekhnicheskie vedomosti SPb GPU. Fiziko-matematicheskie nauki*, vol. 2, p. 242.

"Defense against drones that only Russia has" (2019), available at: https://zen.yandex.ru/media/survival_task/zascita-ot-dronov-kotoraia-est-tolko-u-rossii-5da-6be220ce57b00ad50240b (Accessed 10 November 2022).

Simonyan, K., and Zisserman, A. (2015), "Very Deep Convolutional Networks for Large-Scale Image Recognition", *Proceedings of the 3rd International Conference on Learning Representations, ICLR 2015*, San Diego, CA, USA, May 7–9, 2015, Association for Computational Linguistics, Stroudsburg, PA, USA.

Yuferov, S. (2021), "New Russian tools to combat drones", available at: https://topwar.ru/181712-novye-rossijskie-sredstva-borby-s-bespilotnikami.html (Accessed 17 November 2022).

Yunusova, L.R. and Magsumova, A.R. (2019), "Clustering with neural networks and dependency search", *Nauka, obrazovanie i kul'tura*, vol. 7 (41), p. 18–20.

Информация об авторах

Евгений К. Мазайшвили, Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия; 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1; evgenij997@yandex.ru

Кирилл Л. Тассов, Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия; 105005, Москва, 2-я Бауманская ул., д. 5, стр. 1; ktassov@policesoft.ru

Information about authors

Evgenii K. Mazaishvili, Bauman Moscow State Technical University, Moscow, Russia; bld. 5, 2nd Bauman Str., Moscow, Russia, 105005; evgenij997@yandex.ru

Kirill L. Tassov, Bauman Moscow State Technical University, Moscow, Russia; bld. 5, 2nd Bauman Str., Moscow, Russia, 105005; ktassov@policesoft.ru

УДК 004.056
DOI: 10.28995/2686-679X-2022-4-34-43

Анализ динамики утечки персональных данных в условиях реализации программы «Цифровая экономика Российской Федерации»

Наталия В. Гришина
Российский государственный гуманитарный университет, Москва, Россия,
Московский государственный лингвистический университет, grnat@rambler.ru

Аннотация. В результате реализации национальных проектов России в области цифровой экономики появились принципиально новые возможности практически во всех областях деятельности человека. Однако проблема обеспечения информационной безопасности является системообразующим фактором для успешной реализации всего проекта: если не обеспечить целостность циркулирующей информации, ее доступность, достоверность и, в случае необходимости, ее конфиденциальность, то вся остальная деятельность будет нецелесообразной.

Реализация программы «Цифровая экономика» позволила сформировать принципиально новые возможности для каждого гражданина. Внедрено огромное количество новых сервисов. В этих условиях важная роль отводится обеспечению защиты персональных данных. Граждане доверяют «цифре» свои документы, пользуются онлайн-банками для управления своими денежными средствами, обмениваются сообщениями с помощью различных мессенджеров и т. д. В то же время, несмотря на принятие нормативных правовых документов в сфере обеспечения защиты персональных данных и определения необходимых требований по обеспечению безопасности персональных данных, проблема не решена в полной мере.

В статье была рассмотрена динамика изменений в сфере защиты персональных данных. Отмечается, что в целом она вызывает опасения.

Ключевые слова: персональные данные, цифровая экономика, информационная безопасность, защита информации

Для цитирования: Гришина Н.В. Анализ динамики утечки персональных данных в условиях реализации программы «Цифровая экономика Российской Федерации» // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2022. № 4. С. 34–43. DOI: 10.28995/2686-679X-2022-4-34-43

Analysis of the dynamics of personal data leakage in the context of the implementation of the program "Digital Economy of the Russian Federation"

Nataliya V. Grishina
Russian State University for the Humanities, Moscow, Russia,
Moscow State Linguistic University, Moscow, Russia,
grnat@rambler.ru

Abstract. As a result of the implementation of Russia's national projects in the field of digital economy, fundamentally new opportunities have appeared in almost all areas of human activity. However, the issue of ensuring information security is a system-forming factor for the successful implementation of the entire project: if one does not ensure the integrity of the circulating information, its availability, reliability and, if necessary, its confidentiality, then all other activities will be not feasible.

The implementation of the digital economy program allowed the formation of fundamentally new opportunities for every citizen. A huge number of new services have been introduced. In these conditions, an important role is assigned to ensuring the protection of personal data. Citizens trust Digit with their documents, use online banks to manage their funds, exchange messages using various messengers, etc.

At the same time, despite the adoption of regulatory legal documents in the field of ensuring the protection of personal data and determining the necessary requirements for ensuring the security of personal data, the issue has not been fully solved. The article considered the dynamics of changes in the field of personal data protection. It is noted that in general, it raises concerns.

Keywords: personal data, digital economy, information security, information protection

For citation: Grishina, N.V. (2022), "Analysis of the dynamics of personal data leakage in the context of the implementation of the program 'Digital Economy of the Russian Federation' ", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 4, pp. 34–43, DOI: 10.28995/2686-679X-2022-4-34-43

Введение

В 2011 г. Указом Президента РФ были определены приоритетные направления развития в области науки, технологий и техники. Среди восьми приоритетных направлений особенно выделяется обеспечение информационной безопасности объектов информа-

тизации[1]. Таким образом, можно сказать, что Указ № 899 стал той основой, на которой в дальнейшем базировалось Распоряжение Правительства РФ от 28.07.2017 № 1632-р «Об утверждении программы “Цифровая экономика Российской Федерации”».

Реализация национальных проектов России в области цифровой экономики открывает новые горизонты для внедрения цифровых технологий в социальную и экономическую сферу, предоставляет условия для высокотехнологичного бизнеса. Как следствие – повышение конкурентоспособности страны, укрепление национальной безопасности и качества жизни людей.

Реализация проекта цифровой экономики включает в себя ряд направлений[2]:

– правовое регулирование цифровой среды;
– развитие цифровой инфраструктуры;
– подготовка кадров для цифровой экономики;
– обеспечение информационной безопасности;
– создание условий для развития и внедрения цифровых технологий;
– внедрение цифровых технологий в управление.

Каждое из указанных направлений вносит свою лепту в общий проект. Например, правовое регулирование предполагает не только реализацию проекта в рамках существующего законодательства, но и возможность установления юридической значимости документов, полученных в цифровой среде, наравне с их твердыми копиями.

Особое место занимает проблема обеспечения информационной безопасности. Этот фактор является системообразующим для успешной реализации всего проекта: если не обеспечить целостность циркулирующей информации, ее доступность, достовер-

---

[1] Об утверждении приоритетных направлений развития науки, технологий и техники в Российской Федерации и перечня критических технологий Российской Федерации: Указ Президента Российской Федерации от 07.07.2011 № 899 (в ред. Указа Президента Российской Федерации от 16.12.2015 № 623) // СПС «КонсультантПлюс». URL: http://www.consultant.ru/document/cons_doc_LAW_116178/ (дата обращения 9 октября 2022).

[2] Об утверждении приоритетных направлений развития науки, технологий и техники...; О национальных целях и стратегических задачах развития Российской Федерации на период до 2024 года: Указ Президента РФ от 7 мая 2018 г. № 204 // СПС «КонсультантПлюс». URL: https://mvd.consultant.ru/documents/1056500?items=1&page=1 (дата обращения 9 октября 2022).

ность и, в случае необходимости, ее конфиденциальность, то вся остальная деятельность будет нецелесообразной.

Реализация программы «Цифровая экономика» позволила сформировать принципиально новые возможности для каждого гражданина. Внедрено огромное количество новых сервисов. Если раньше для того чтобы получить рядовую справку, людям приходилось отстаивать огромные очереди, а в сложных ситуациях на это необходимо было потратить несколько дней, то теперь получить многие документы и заказать услуги можно не выходя из дома.

В этих условиях важная роль отводится обеспечению защиты персональных данных. Граждане доверяют «цифре» свои документы, пользуются онлайн-банками для управления своими денежными средствами, обмениваются сообщениями с помощью различных мессенджеров и т. д.

Еще в 2006 г. был принят Федеральный закон «О персональных данных»[3], но до сих пор проблема защиты персональных данных стоит очень остро.

В России существует «черный рынок» персональных данных. Лидерами этого рынка являются финансовые организации и интернет-провайдеры. Этот рынок растет год от года. Не секрет, что источником «утечки» персональных данных чаще всего являются собственно сотрудники компаний.

15 августа 2022 г. был опубликован очередной отчет по утечке данных, их структуре и динамике Экспертно-аналитического центра ГК InfoWatch[4].

Опираясь на данные Экспертно-аналитического центра ГК InfoWatch за последние годы, можно проследить динамику по ряду вопросов, касающихся утечки данных вообще и персональных в частности.

Если рассматривать структуру всех данных, подвергшихся утечке, то львиную долю среди них занимают именно персональные данные. Причем с каждым годом эта доля увеличивается все больше (рис. 1). Таким образом, говоря об утечке информации, можно с высокой долей вероятности утверждать, что речь идет именно о персональных данных.

---

[3] О персональных данных: Федер. закон [принят Гос. Думой 27.07.2006 № 152-ФЗ] // СПС «КонсультантПлюс». URL: http://www.consultant.ru/document/cons_doc_LAW_61801/ (дата обращения 9 октября 2022).

[4] Россия. Утечки информации ограниченного доступа в 2021 году // Экспертно-аналитический центр InfoWatch. 2022. URL: https://www.infowatch.ru/sites/default/files/analytics/files/rossiya-rost-latentnosti-intsidentov-i-vnutrennikh-utechek.pdf (дата обращения 9 октября 2022).

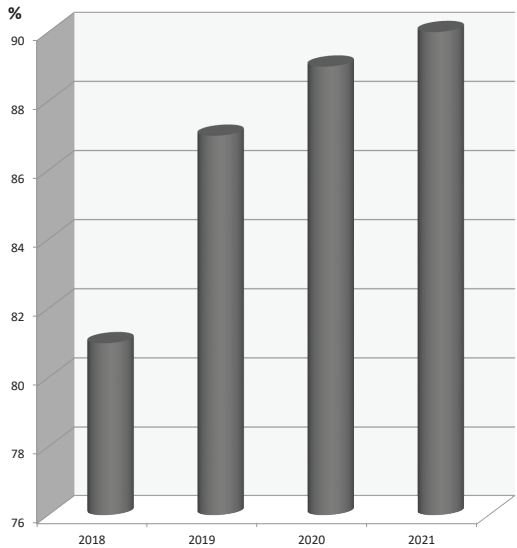


Рис. 1. Доля персональных данных среди всех данных, которые были скомпрометированы в России

Описанный факт можно достаточно легко объяснить: именно персональные данные «востребованы» в большей мере на «черном» рынке. Интересно проследить динамику виновников в утечке информации. На рис. 2. показана динамика утечки информации с точки зрения ее виновников. Все возможные участники информационного обмена разделены на две группы: сотрудники предприятия или организации и хакеры или неизвестные лица. Большая часть нарушений, связанных с утечкой информации, относится именно к действиям сотрудников организации, поэтому они и выделены в отдельную группу. За последние четыре года заметна тенденция снижения доли сотрудников предприятия и, соответственно, увеличение доли всех остальных злоумышленников.

Реализация программы развития цифровой экономики позволила выявить еще целый ряд закономерностей, связанных с утечкой информации. Например, более чем в два раза уменьшилась доля бумажных носителей в структуре похищенной информации. Если в 2018 г. доля бумажных носителей составляла около 39%, то к 2021 г. она составила 14%. Указанные изменения демонстрирует рис. 3, из которого видно, что в этом плане произошли качественные изменения.

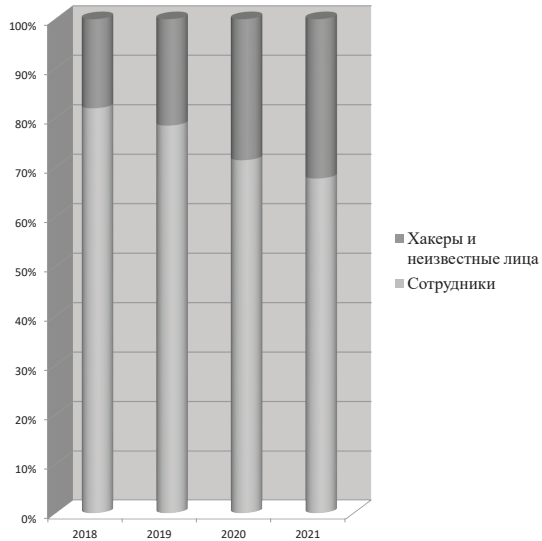


Рис. 2. Процентное соотношение доли виновных в утечке информации

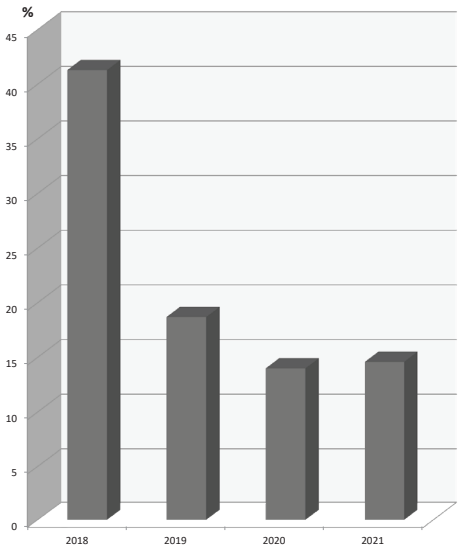


Рис. 3. Доля бумажных носителей в структуре похищенной информации

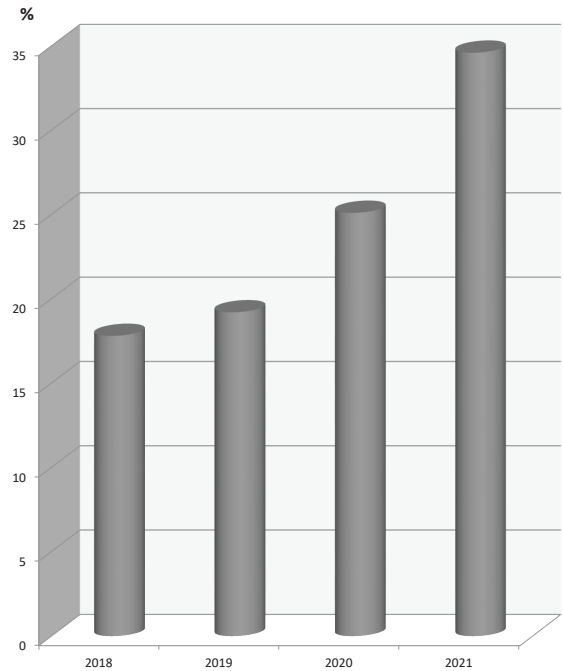


Рис. 4. Объем утечек информации
в высокотехнологичной сфере

В 2021 г. по сравнению с 2018 г. в два раза увеличился объем утечек информации в высокотехнологичной сфере (рис. 4.). Тенденция на увеличение этой доли прослеживается за весь рассматриваемый период.

Также можно отметить, что существенно (на треть) выросла доля умышленных утечек в общем потоке (рис. 5). Этот факт, на взгляд авторов, скорее положительный: раз выросла доля умышленных утечек, соответственно уменьшилась доля случайных и связанных с ошибками. А это уже может свидетельствовать о том, что сотрудники повышают свою квалификацию и совершенствуется деятельность по организации самого процесса защиты информации.

Что же получается? Почему, несмотря на то что принимаются различные нормативные акты, развиваются технологии, ведется работа по повышению грамотности населения в плане защиты персональных данных, проблема не решается в должной мере?

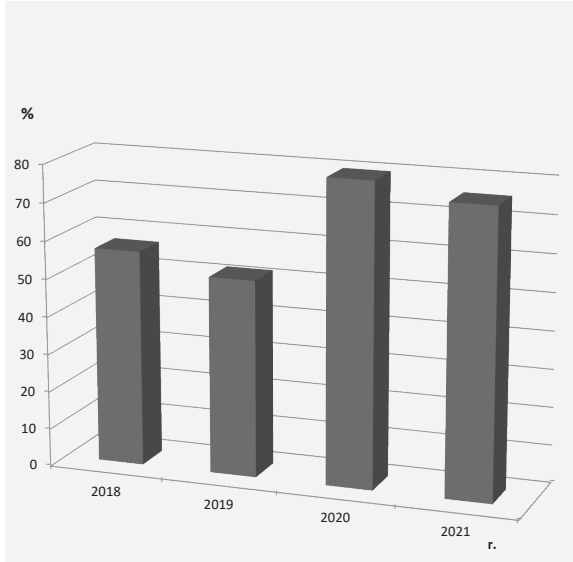


Рис. 5. Доля умышленных утечек
в общем объеме скомпрометированной информации

Можно назвать несколько причин:

– действующие нормативные и правовые документы не учитывают «нюансы» конкретных противоправных действий относительно персональных данных;
– несущественная ответственность организаций, связанная с утечкой персональных данных;
– субъекты скомпрометированных персональных данных не обращаются с исковыми заявлениями в суд;
– активная деятельность злоумышленников с привлечением методов социальной инженерии, направленная на «слабые» слои населения;
– невозможность построения полностью защищенной информационной инфраструктуры и т. п. [Скрынникова 2022].

Определить какой-либо конкретный перечень последствий утечки персональных данных практически невозможно. Этот перечень может существенно меняться в зависимости от ряда факторов:

– содержания скомпрометированных персональных данных;
– изобретательности мошенников, которые эти данные получили;
– поведения лица, чьи данные скомпрометированы.

Заключение

В статье была рассмотрена динамика изменений в сфере защиты персональных данных. Отмечается, что в целом она вызывает опасения.

В то же время у пользователей и клиентов различных сервисов и компаний отсутствует возможность оказывать какое-либо влияние на сохранность своих персональных данных. И, поскольку пользователь не может определять порядок защиты и хранения своих персональных данных, его задача состоит в том, чтобы оставлять как можно меньше персональных данных при любых обстоятельствах.

Не лишним будет отзывать согласие на обработку персональных данных при прекращении использования тех или иных услуг.

Возможна утечка персональных данных и в случае использования бесплатного WI-FI в общественных местах. Поэтому следует проявлять бдительность в случае его использования.

Отдельно хотелось бы отметить, что, несмотря на существование нормативных методических документов, связанных с обеспечением защиты персональных данных, их действие недостаточно эффективно.

Литература

Арутюнов, Авралева 2021 – *Арутюнов В.В., Авралева И.Ю.* Технология блокчейн: начало, настоящее, будущее // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2021. № 4. С. 30–46.

Благов, Митюшин, Пучков, Ремизова 2021 – *Благов В.О., Митюшин Д.А., Пучков Г.Ю., Ремизова Е.В.* Основные направления создания информационной системы для идентификации личности по фенотипическим признакам человека // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2021. № 2. С. 37–47.

Гришина, Емельянов 2006 – *Гришина Н.В., Емельянов С.А.* Деловая разведка как разновидность информационной работы // Прикладная информатика. 2006. № 3 (3). С. 34–41.

Скрынникова 2022 – *Скрынникова А.* Побочное явление цифровизации: как в России крадут и продают персональные данные // Forbes. URL: https://www.forbes.ru/tehnologii/433651-pobochnoe-yavlenie-cifrovizacii-kak-v-rossii-kradut-i-prodayut-personalnye-dannye (дата обращения 9 октября 2022).

References

Arutyunov, V.V. and Avraleva, I.Yu. (2021), "Blockchain technology. The beginning, the present, the future", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 4, pp. 30–46.

Blagov, V.O., Mityushin, D.A., Puchkov, G.Yu. and Remizova, E.V. (2021), "The main directions of creating an information system for personal identification based on phenotypic characteristics", *RSUH/RGGU Bulletin*. "*Informatics. Information security. Mathematics*" Series, no. 2, pp. 37–47.

Grishina, N.V., Emelyanov S.A. (2006), "Business intelligence as a kind of information work", *Applied Informatics*, no. 3 (3), pp. 34–41.

Skrynnikova, A. (2022), "A side effect of digitalization. How personal data is stolen and sold in Russia", available at: https://www.forbes.ru/tehnologii/433651-pobochnoe-yavlenie-cifrovizacii-kak-v-rossii-kradut-i-prodayut-personalnye-dannye (Accessed 9 October 2022).

Информация об авторе

Наталия В. Гришина, кандидат технических наук, доцент, Российский государственный гуманитарный университет, Москва, Россия; 125047, Россия, Москва, Миусская пл., д. 6;

Московский государственный лингвистический университет, Москва, Россия; 119034, Россия, Москва, ул. Остоженка д. 38 стр. 1; grnat@rambler.ru

Information about the author

Nataliya V. Grishina, Cand.of Sci. (Computer Science), associate professor, Russian State University for the Humanities, Moscow, Russia; bld. 6, Miusskaya Sq., Moscow, Russia, 125047;

Moscow State Linguistic University, Moscow, Russia; bld. 3, ed. 1, Ostozhenka Str., Moscow, Russia, 119034; grnat@rambler.ru

УДК 003.26:004.056
DOI: 10.28995/2686-679X-2022-4-44-58

Анализ применения методов цифровой голографии для защиты информации

Тамара М. Волосатова
Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия, tamaravol@gmail.com

Анастасия А. Козарь
Московский государственный технический университет имени Н.Э. Баумана, Москва, Россия, kozar.a.a.rk6@yandex.ru

Аннотация. В работе приведен обзор методов цифровой голографии как методов регистрации информации; приведена общая информация о понятии «голограмма» различных типов, проведена параллель между физической (оптической, аналоговой) и цифровой голографией; рассмотрены различия и сходства, отмечены причины преобладания цифровой голографии для современных целей. Кратко рассмотрены принципы и некоторые способы синтеза цифровых голограмм, приведены их преимущества и недостатки. Составлена краткая историческая справка о создании самого понятия, рассмотрены текущие достижения в области исследований как в России, так и за рубежом. Рассмотрены также различные области для практического применения цифровых голограмм, в частности дан анализ проблемы защиты информации как перспективной области для применения цифровых голограмм, в том числе в комбинации их с различными методами сокрытия информации. Выделены преимущества и недостатки подобного способа защиты информации. Рассмотрено отличие криптографии от стеганографии, кратко рассмотрены некоторые стеганографические методы сокрытия информации, указаны их недостатки и преимущества, а также основной принцип, используемый указанными методами для сокрытия информации; сделаны выводы на основе указанной информации.

Ключевые слова: цифровая голография, голограмма, стеганография, криптография, защита информации

Для цитирования: Волосатова Т.М., Козарь А.А. Анализ применения методов цифровой голографии для защиты информации // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2022. № 4. С. 44–58. DOI: 10.28995/2686-679X-2022-4-44-58

Analysis of the application of digital holography methods for information protection

Tamara M. Volosatova
Bauman Moscow State Technical University, Moscow, Russia,
tamaravol@gmail.com

Anastasiya A. Kozar'
Bauman Moscow State Technical University, Moscow, Russia,
kozar.a.a.rk6@yandex.ru

Abstract. The paper provides an overview of digital holography as a method of recording information, provides general information about the concept of "hologram" of different types, draws a parallel between physical (optical, analog**o**us) and digital holography, considers differences and similarities, and points out the reasons for the predominance of digital holography for modern purposes. The principles and some methods of synthesizing digital holograms are briefly considered, their advantages and disadvantages are given. The article also considers a brief historical background on the creation of the concept itself is given, current achievements in the field of research both in Russia and abroad. Besides it as well reviewed various areas for the practical application of digital holograms, in particular, the issue of information security – as a promising area for the use of digital holograms including in combination with various methods of hiding information. The advantages and disadvantages of such a method of information protection are highlighted. The distinction of cryptography from steganography is studied, and some steganographic methods of hiding information are briefly outlined, their disadvantages and advantages are indicated, as well as the basic principle used by those methods to hide information. Conclusions are drawn on the basis of the specified information.

Keywords: digital holography, hologram, steganography, cryptography, data protection

For citation: Volosatova, T.M. and Kozar', A.A. (2022), "Analysis of the application of digital holography methods for information protection", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 4, pp. 44–58, DOI: 10.28995/2686-679X-2022-4-44-58

Введение

На сегодняшний день голография – широко известное понятие, использующееся в совершенно различных областях и имеющее сотни применений. Это и искусство, и хранение данных, и защита

информации, микроскопия, микробиология, медицина, электромеханика – список может продолжаться бесконечно.

Способность голографии регистрировать данные об объемных предметах в практически двухмерной среде хотя и является поистине удивительной, но в настоящее время уже не удивляет так сильно – сейчас каждый человек так или иначе встречается с голографией в повседневной жизни на постоянной основе. В свое же время это открытие перевернуло представление о регистрации изображений.

В 1947 г. Деннис Габор (рис. 1), профессор государственного колледжа в Лондоне, синтезировал первую голограмму в процессе экспериментов, направленных на увеличение разрешающей силы электронных микроскопов [Myung 2010].

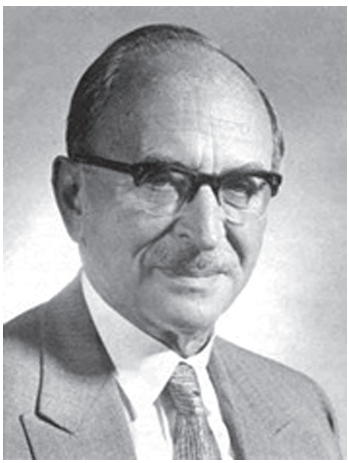


Рис. 1. Деннис Габор, создатель первой голограммы

К сожалению, технических средств того времени было недостаточно для того, чтобы по заслугам оценить значимость этого открытия. Большое распространение голография получила только после создания первого лазера в 1960 г., что дало толчок к развитию оптической голографии. Уже через 2 года была синтезирована первая объемная пропускающая голограмма с использованием двулучевой схемы. Советский ученый Денисюк Ю.Н. первый в мире удачно сочетал метод Липпмана с особенностями современной голографии [Денисюк 1962], [Денисюк 1963]. Однако подобная схема обладала существенным недостатком – для восстановления

необходим был именно лазер, который на то время еще не получил большого распространения и был достаточно сложным прибором для массового воспроизведения. Тем не менее именно в то время начали проводиться обширные исследования предметной области, велись работы по улучшению качества голограмм, а также стали выделяться методы практического применения голографии в различных областях, как в науке, так и, например, в искусстве.

Понятие цифровой голографии было предложено в конце 1960-х годов. Было предложено заменить часть дорогостоящих и технически сложных этапов синтеза голограммы моделированием на ЭВМ, однако подобная идея не получила должного признания, поскольку на тот момент не существовало ни регистрирующих устройств высокого разрешения, ни компьютеров достаточной мощности [Денисюк 1979]. Именно поэтому все исследования в этом направлении сводились к математическим моделям, формулам и теоретическим выкладкам, без решения реальных практических проблем. Медленно, но верно развитие технологий позволило не только продолжать и углублять исследования данной предметной области, но и повсеместно применять результаты исследований.

История возникновения оптической голографии. Метод Габора

История голографии началась с Денниса Габора (05.06.1900–09.02.1979) – венгерского физика, работавшего над идеей улучшения разрешающей способности электронного микроскопа и предложившего двухступенчатый метод получения оптического изображения. Первоначально объект освещается когерентной или световой волной, при этом считается, что большая часть волны проникает в объект без возмущения (предметная волна) [Андреева 2008]. Когерентная предметному пучку света, вторая волна (опорная) направляется на специальным образом подготовленную фотопластинку. В результате интерференции предметной волны с опорной возникает интерференционная картина, регистрирующая образ не по интенсивности излучения, а по образцу волнового фронта, фиксируясь в виде объемного объекта на фотопластинке. Восстановление исходного объекта требует лишь освещения голограммы аналогичной опорной волной [Беркхфрт, Кольер, Лин 1973] (рис. 2, где S – экран, B – исходный объект, H – синтезированная голограмма, L – линза, B’ – действительное изображение объекта, S’ – наблюдатель.

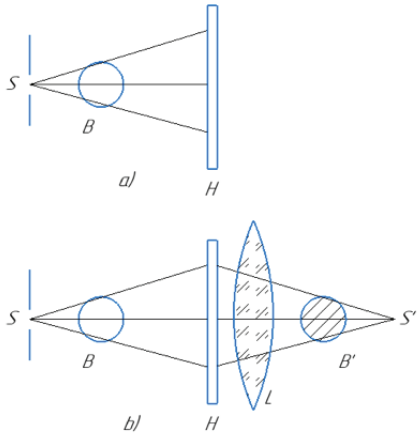


Рис. 2. Синтез (a) голограммы
и восстановление изображения (b) по методу Габора

Важно отметить, что запись на фотопленку производится без формирующих оптических систем, вследствие чего в каждой точке поверхности находится свет от всех точек исходного предмета. Таким образом, можно отметить важную особенность голограммы: каждая ее точка одновременно содержит информацию обо всех точках исходного объекта, а это значит, что по любой точке голограммы можно восстановить изображение исходного объекта. Любая часть голографической пленки, независимо от степени повреждения, позволит восстановить полное изображение исходного объекта. При увеличении степени повреждений качество воспроизведения будет ухудшаться, однако на целостности изображения это не скажется. Свое открытие Габор назвал «голограммой», что дословно означает «вижу все». Таким образом, Габор подчеркивал – его метод позволяет регистрировать полную информацию об объекте.

Для синтеза голограмм Габор использовал собственный источник когерентных волн – ртутную лампу, которая, при определенных условиях, обеспечивала подачу светового сигнала с узким спектром. Итоговое изображение было плохого качества (рис. 3), и открытие голографического принципа не вызвало должного интереса вплоть до 1960-х годов, когда были разработаны лазерные источники света, которые и позволили получать оптические голограммы высокого качества, и интерес к голограммам появился вновь [Федоров 1988].

Рис. 3. Пример результатов исследований Габора (исходный объект, синтезированная на его основании голограмма и восстановленное изображение)

Понятие цифровой голографии и ее преимущества

В большинстве источников цифровой голографией называется процесс синтеза и восстановления голограмм с применением компьютерных средств, при этом используются числовые модели волн, а не сами волны. В качестве устройства, регистрирующего интерференционные полосы, может выступать ПЗС-матрица – специальная интегральная микросхема, состоящая из светочувствительных элементов.

Это направление получило активное развитие в 1967–1980 гг. С этими исследованиями в первую очередь упоминают имена зарубежных ученых А. Ломана и В. Ли, а в нашей стране – Б. Федорова и Р. Эльмана. Активное содействие в исследованиях оказывал Ю. Денисюк, являвшийся основоположником оптической голографии в Советском Союзе и проводивший исследования параллельно с Д. Габором, но несколько позже (1958 г.) [Денисюк, Суханов 1970].

Интерес к подобному методу был вполне оправдан: успешный голографический опыт требует наличия сложного оборудования, а качество голограммы напрямую зависит от множества факторов, не всегда зависящих от экспериментаторов, например сторонние воздействия, вибрации установки или ее составных частей, дефекты оборудования, неравномерность светового потока, точность расчетов и т. д. Цифровая же голография предлагала заменить трудозатратный, нестабильный в своем результате физический опыт на математическое моделирование, тем самым позволяя избежать множества сложностей. Именно такое моделирование называется цифровой голографией.

Сейчас, в период глобальной компьютеризации, активного развития технологий и стремительного повышения производительности вычислительной техники, все больше ученых обращают свое внимание именно на цифровую голографию. Современные компьютеры позволяют добиться качественной обработки изображений, способны моделировать процесс с момента синтеза до момента восстановления исходного изображения со всеми промежуточными этапами реального физического эксперимента. Это стало возможно благодаря детально описанному математическому аппарату.

Дополнительно имеется ряд преимуществ цифровой голографии относительно оптической, например тот факт, что геометрические размеры голограммы не ограничиваются такими физическими факторами, как когерентность волн, сторонние вибрации или влияние воздушных масс. Еще одним значительным фактором является возможность создания оптического фронта для физически несуществующего объекта.

Практическое применение цифровой голографии

Цифровая голография нашла свое применение в различных сферах, и исследованиями в этой области в настоящее время занято несколько крупных университетов и компаний как в России, так и за рубежом. Так, например, НИЯУ МИФИ реализовал систему динамической записи, передачи и демонстрации в реальном времени голограмм с разрешением не менее 2 млн пикселей. Это позволяет воспроизводить сцены, записанные в различных диапазонах (оптическом и инфракрасном), что может быть применено для получения информации из агрессивной среды [Cheremkhin, Krasnov, Molodtsov, Rodin 2018].

Не менее важное применение голография нашла в таких сферах, как решение проблемы повышения качества оптических 3D-изображений, в сфере распознавания объектов по форме и спектральным признакам, а также в сфере хранения и защиты информации.

Защита информации с использованием голографии

В настоящее время оптическая голография прочно укоренилась в области защиты информации. Всем известные голографические наклейки на ряде товаров, а также на денежных знаках являются

примером защиты от фальсификации с использованием голографических методов и применяются повсеместно. Подобная пленка может быть создана как с помощью оптической, так и с помощью цифровой голографии, однако цифровая голография имеет ряд преимуществ: возможность внедрения элементов высокого разрешения, смена изображения в зависимости от угла зрения, микро- и нанотексты (высота шрифта может быть уменьшена до 5 мкм), скрытые изображения, различные кинематические эффекты и т. д. Помимо этого, в одной голограмме успешно сочетаются оптический и/или логические и цифровые способы защиты с целью повышения эффективности [Одиноков, Грейсух, Левин 2020].

Однако проверка подлинности – не единственное возможное применение голографии в области защиты информации. Перспективным видится также использование голографии для сокрытия графической информации непосредственно в самой голограмме. Действительно, принимая в расчет то, что для восстановления голограммы необходимы вычислительные мощности, а также известный математический аппарат, хранение и передача информации в виде синтезированных голограмм является неплохим вариантом сохранения информации от компрометации. Помимо этого, голограмма несет в каждом своем участке одну и ту же информацию, что позволяет восстановить исходные данные при повреждении целостности голограммы. Наконец, имеется возможность записать на одну и ту же «фотопластинку» большой объем различной информации, меняя параметры волн. Таким образом, применение цифровых голографических методов позволяет синтезировать голограммы изображений (например, чертежей или другой проектной документации), мультиплексные голограммы нескольких изображений, трехмерных объектов или объектов с различными эффектами [Борискевич, Ероховец, Ткаченко 2017].

Комбинированные методы защиты информации. Голография и стеганография

Обратимся к современным способам защиты информации. К ним принято относить такие понятия, как криптография и стеганография.

Криптографией называют науку сохранения информации. В широком смысле это наука о шифровании данных. Информация преобразовывается в ключ (код), который подлежит расшифровке только с использованием подходящего ответного ключа [Ященко 2001].

Стеганография же – способ передачи или хранения данных с помощью сокрытия самого факта передачи информации от стороннего наблюдателя [Волосатова, Денисов, Чичварин 2012]. Следовательно, преимущество стеганографии состоит в том, что сокрытое сообщение не вызывает подозрений у предполагаемого злоумышленника.

Несмотря на то что голограмма сама по себе не несет в себе никакой полезной информации об исходном объекте, на ней зачастую отображена характерная интерференционная картина (рис. 4), позволяющая заподозрить в графическом фрагменте голограмму, что даст потенциальному злоумышленнику стимул предпринять попытку восстановления информации [Грибунин, Оков, Туринцев 2002].

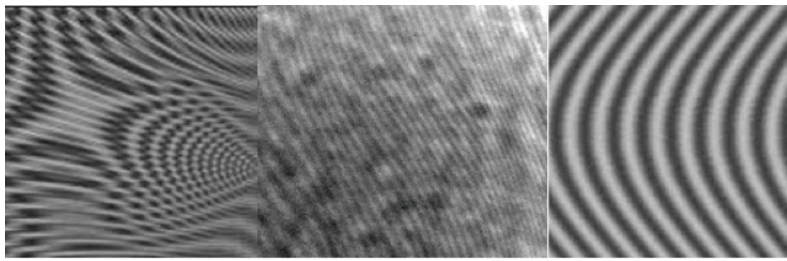


Рис. 4. Несколько примеров
интерференционной картины на голограмме

Следовательно, комбинация стеганографических методов с голографическими представляется наиболее интересной с точки зрения надежности сокрытия, обеспечивая большую надежность [Волосатова, Спасенов 2014].

При этом возможных комбинаций существует большое множество, как и последовательностей сокрытия. Имеют место быть варианты использования голографии как контейнера для стегометодов, но, как было отмечено выше, это не имеет большого смысла с точки зрения надежности сокрытия.

Наиболее продуктивной, на первый взгляд, представляется следующая последовательность действий, основанная на общей структуре стегосистемы (рис. 5):

– синтез голограммы на основе скрываемого изображения;
– встраивание полученной голограммы в стегоконтейнер;
– передача полученного стегоконтейнера получателю;
– восстановление голограммы получателем из стегоконтейнера по известному защищенному ключу;
– восстановление исходного изображения из голограммы.

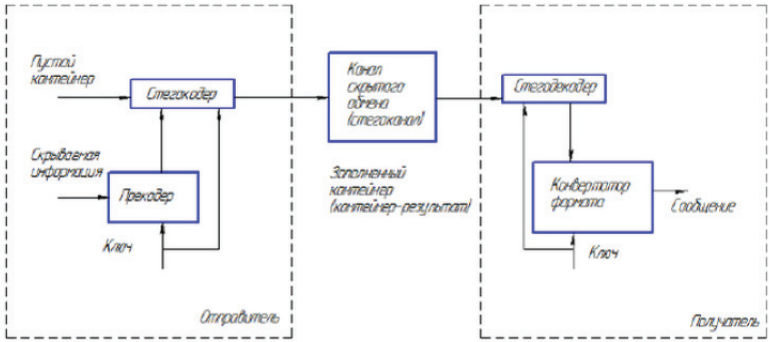


Рис. 5. Общая структура стегосистемы

При этом следует принимать во внимание важное очевидное требование к стеганографическому методу: он не должен быть хрупким, а также должен быть максимально устойчив к помехам и искажениям. Метод является хрупким, если при модификации контейнера его содержимое каким-либо образом меняется или разрушается. Понятно, что оба этих требования важны для сохранения голограммы внутри контейнера в максимально первоначальном виде, чтобы обеспечить последующее восстановление изображения. Впрочем, стегосистема может проектироваться хрупкой или неустойчивой специально, для повышения надежности, или иных целей.

Некоторые стеганографические методы

Из наиболее известных стеганографических методов сокрытия данных можно для примера рассмотреть следующие:

– использование младших бит (LSB метод);
– эхо-методы;
– фазовое кодирование.

LSB метод – популярный метод, суть которого заключается в замене последних «наименее значимых» в массе информации бит на биты скрываемого сообщения. Метод прост в реализации, но неустойчив к любым видам атак и искажений контейнера, а также имеет небольшую емкость (соотношение объема контейнера к объему скрываемого сообщения) [Ross 1998].

Эхо-метод – суть метода во встраивании информации за счет времени задержки эхо-сигнала в аудиофайле. Несущими параметрами являются начальная амплитуда, время спада и сдвиг между исходным и эхо-сигналом. Человеческое ухо не способно различить два сигнала при уменьшении сдвига, таким образом, сигналы смешиваются и воспринимаются единым целым. Эхо-методы устойчивы к ряду атак, уязвимы к атакам по времени. Также в качестве существенного минуса можно выделить емкость контейнеров, которую необходимо удерживать для успешного сокрытия [Волосатова, Чичварин 2015].

Фазовое кодирование: сокрытие сообщения производится путем встраивания вместо исходного звукового элемента относительной фазы, которая и является скрываемым сообщением, при этом подряд идущие элементы должны сохранять относительную фазу между исходными. Человеческое ухо не различает значение фазы, а фиксирует только разность, что и позволяет пользоваться подобным методом. Метод является эффективным способом сокрытия информации, однако его пропускная способность невелика [Конахович, Пузыренко 2006].

Помимо этого, можно отметить методы компьютерной стеганографии, используемые для сокрытия особенности компьютерных систем. Например, использование зарезервированных полей формата файлов, сокрытие информации в неиспользуемом пространстве на жестком диске, особенность хранения файлов в некоторых файловых системах. Все эти методы имеют общий недостаток – легкость обнаружения и малая вместимость [Ремизов, Филлипов 2012].

Заключение

Цифровая голография повсеместно используется для совершенно различных целей, и одна из самых интересных на данный момент – задача защиты проектной документации и иных данных. Перспективным представляется использование голографического метода вкупе со стеганографическими методами сокрытия информации, что позволит скрыть факт передачи информации и дополнительно усилить уровень безопасности сокрытия при компрометации канала передачи и раскрытия потенциальным злоумышленником факта передачи зашифрованной информации. Перечисленное позволяет утверждать, что комбинированные методы на основе синтеза голограмм и стеганографических методов могут дать существенное повышение уровня защиты информации

и обеспечить скрытую передачу данных с применением дополнительного уровня безопасности, однако результативность и эффективность защиты данных сильно зависит от выбранного стеганографического метода. Необходимы дальнейшие исследования для выделения критериев методов, выбора оптимального метода и проведение численных экспериментов.

Литература

Андреева 2008 – *Андреева О.В.* Прикладная голография. СПб.: ИТМО, 2008. 184 с.

Беркхфрт, Кольер, Лин 1973 – *Беркхфрт К., Кольер Р., Лин Л.* Оптическая голография. М.: Мир, 1973. 698 с.

Борискевич, Ероховец, Ткаченко 2017 – *Борискевич А.А., Ероховец В.К., Ткаченко В.В.* Синтез и восстановление квантованных голограмм Фурье и Френеля для защиты цифровых изображений // Труды БГТУ. 2017. Серия 4. № 2. С. 29–36

Волосатова, Денисов, Чичварин 2012 – *Волосатова Т.М., Денисов А.В., Чичварин Н.В.* Комбинированные методы защиты данных в САПР // Информационные технологии. 2012. № 5. С. 1–32.

Волосатова, Спасенов 2014 – *Волосатова Т.М., Спасенов А.Ю.* Разработка программной реализации комбинированного метода для сокрытия информации на основе цифровой голограммы Френеля и алгоритма LSB. URL: http://ainsnt.ru/doc/730990.html (дата обращения 5 октября 2022).

Волосатова, Чичварин 2015 – *Волосатова Т.М., Чичварин Н.В.* Метод сокрытия данных в стереофонических аудиофайлах. URL: http://ainjournal.ru/doc/792392.html (дата обращения 5 октября 2022).

Грибунин, Оков, Туринцев 2002 – *Грибунин В.Г., Оков И.Н., Туринцев И.В.* Цифровая стеганография. М.: СОЛОН-Пресс, 2002. 265 с.

Денисюк 1962 – *Денисюк Ю.Н.* Об отражении оптических свойств объекта в волновом поле рассеянного им излучения. М.: ДАН СССР, 1962. 1275 с.

Денисюк 1963 – *Денисюк Ю.Н.* Об отражении оптических свойств объекта в волновом поле рассеянного им излучения // Оптика и стереоскопия. 1963. Т. 15. С. 522.

Денисюк 1979 – *Денисюк Ю.Н.* Принципы голографии. Л.: ГОИ, 1979. 125 с.

Денисюк, Суханов 1970 – *Денисюк Ю.Н., Суханов В.И.* Голограмма с записью в трехмерной среде как наиболее совершенная форма изображения // Успехи физических наук. 1970. № 6. С. 337–345.

Козарь 2016 – *Козарь А.А.* Исследования применимости метода LSB для сокрытия текстовой и графической документации в аудиосреде. URL: http://ainsnt.ru/doc/839394.html (дата обращения 7 октября 2022).

Конахович, Пузыренко 2006 – *Конахович Г.Ф., Пузыренко А.Ю.* Компьютерная стеганография. Теория и практика. М.: МК-Пресс, 2006. 288 с.

Одиноков, Грейсух, Левин 2020 – *Одиноков С.Б., Грейсух Г.И., Левин Г.Г.* Цифровая голография – от математической идеи до реальных приложений компьютерной оптики // VI Международная конференция и молодежная школа «Информационные технологии и нанотехнологии». Самара: Самарский национальный исследовательский ун-т, 2020. С. 758–769.

Ремизов, Филлипов 2012 – *Ремизов А.В., Филлипов М.В.* Оценка необнаруживаемости стеганографических алгоритмов. URL: https://cyberleninka.ru/article/n/otsenka-neobnaruzhimosti-steganograficheskih-algoritmov/viewer (дата обращения 7 октября 2022).

Федоров 1988 – *Федоров Б.Ф.* Лазеры. Основы устройства и применение. М.: ДОСААФ, 1988. 190 с.

Ященко 2001 – *Ященко В.В.* Введение в криптографию. СПб.: Питер, 2001. 289 с.

Cheremkhin, Krasnov, Molodtsov, Rodin 2018 – *Cheremkhin P.A., Krasnov V.V., Molodtsov D.Yu., Rodin V.G.* Recognition of objects radiating with broad spectrum in dispersive holographic correlator // Optics Communications. 2018. № 8. С. 73–78.

Myung 2010 – *Myung K.* Principles and techniques of digital holographic microscopy // SPIE Reviews. 2010. Vol. 1, issue 1. DOI: https://doi.org/10.1117/6.0000006.

Ross 1998 – *Ross J.* Stretching the limits of steganography // IEEE Journal on Selected Areas in Communications. 1998. № 3. P. 39–48.

References

Andreeva, O.V. (2008), *Prikladnaya golografiya* [Applied holography], ITMO, Saint Petersburg, Russia.

Boriskevich, A.A., Erokhovets, V.K. and Tkachenko, V.V. (2017), "Synthesis and restoration of quantized Fourier and Fresnel holograms for digital image protection", *Proceedings of BSTU, 4 Series,* vol. 2, pp. 29–36.

Cheremkhin, P.A., Krasnov, V.V., Molodtsov, D.Yu. and Rodin, V.G. (2018), "Recognition of objects radiating with broad spectrum in dispersive holographic correlator", *Optics Communications*, vol. 8, pp. 73–78.

Denisyuk, Yu.N. (1962), *Ob otrazhenii opticheskikh svoistv ob"ekta v volnovom nole rasseyannogo im izlucheniya* [About the reflection of the optical properties of an object in the wave field of radiation scattered by it], DAN SSSR, Moscow, Russia.

Denisyuk, Yu.N. (1963), "About the reflection of the optical properties of an object in the wave field of radiation scattered by it", *Optika i stereoskopiya*, vol. 15, p. 522.

Denisyuk, Yu.N. (1979), *Printsipy golografii* [Principles of holography], GOI, Leningrad, Soviet Union.

Denisyuk, Yu.N. and Sukhanov, V.I. (1970), "A hologram with a record in a three-dimensional environment as the most perfect form of an image", *Sov. Phys. Usp.*, vol. 6, pp. 337–345.

Fedorov, B.F. (1988), *Lazery. Osnovy ustroistva i primenenie* [Lasers. Basics of the device and application], DOSAAF, Moscow, Russia.

Gribunin, V.G., Okov, I.N. and Turintsev, I.V. (2002), *Tsifrovaya steganografiya* [Digital steganography], SOLON-Press, Moscow, Russia.

Kozar', A.A. (2016), "Research on the applicability of the LSB method for hiding text and graphic documentation in an audio environment", available at: http://ainsnt.ru/doc/839394.html (Accessed 7 October 2022).

Konakhovich, G.F. and Puzyrenko, A.Yu. (2006), *Komp'yuternaya steganografiya. Teoriya i praktika* [Computer steganography. Theory and practice], MK-Press, Moscow, Russia.

Myung, K. (2010), "Principles and techniques of digital holographic microscopy", *SPIE Reviews*, vol. 1, issue 1, DOI: https://doi.org/10.1117/6.0000006.

Odinokov, S.B., Greisukh, G.I. and Levin, G.G. (2020), "Digital holography – from a mathematical idea to real applications of computer optics", *VI Medzhunarodnaya konferentsiya i molodezhnaya shkola "Informatsionnye tekhnologii i nanotekhnologii"* [VI International Conference and Youth School "Information Technologies and Nanotechnologies"], Samara National Research University, Samara, Russia, pp. 758–769.

Remizov, A.V. and Fillipov, M.V. (2012), "Estimating the undetectableness of steganographic algorithms", available at: https://cyberleninka.ru/article/n/otsenka-neobnaruzhimosti-steganograficheskih-algoritmov/viewer (Accessed 7 October 2022).

Ross, J. (1998), "Stretching the limits of steganography", *IEEE Journal on Selected Areas in Communications*, vol. 3, pp. 39–48.

Volosatova, T.M., Denisov, A.V. and Chichvarin, N.V. (2012), "Combined data protection methods in CAD", *Informatsionnye tekhnologii*, vol. S5, pp. 1–32.

Volosatova, T.M. and Spasenov, A.Yu. (2012), "Development of a software implementation of a combined method for hiding information based on a digital Fresnel hologram and the LSB algorithm", available at: http://ainsnt.ru/doc/730990.html (Accessed 5 October 2022).

Volosatova, T.M. and Chichvarin, N.V. (2015), "Method for hiding Data in stereo audio files", available at: http://ainjournal.ru/doc/792392.html (Accessed 5 October 2022).

Yashchenko, V.V. (2001), *Vvedenie v kriptografiyu* [Introduction to cryptography], Piter, Saint Petersburg, Russia.

Информация об авторах

Тамара М. Волосатова, кандидат технических наук, доцент, Московский государственный технический университет им. Н.Э. Баумана, Москва, Россия; 105005, Россия, Москва, 2-я Бауманская ул., д. 5; tamara-vol@gmail.com

Анастасия А. Козарь, аспирант, Московский государственный технический университет им. Н.Э. Баумана, Москва, Россия; 105005, Россия, Москва, 2-я Бауманская ул., д. 5; kozar.a.a.rk6@yandex.ru

Information about the authors

Tamara M. Volosatova, Cand. of Sci. (Computer Engineering), associate professor, Bauman Moscow State Technical University, Moscow, Russia; bld. 5, 2nd Bauman Str., Moscow, Russia, 105005; tamaravol@gmail.com

Anastasiya A. Kozar', postgraduate student, Bauman Moscow State Technical University, Moscow, Russia; bld. 5, 2nd Bauman Str., Moscow, Russia, 105005; kozar.a.a.rk6@yandex.ru

УДК 621
DOI: 10.28995/2686-679X-2022-4-59-74

Применение нечетких данных в задачах оценки долговечности

Ирина В. Гадолина
Институт машиноведения им. А.А. Благонравова Российской академии наук (ИМАШ РАН), Москва, Россия, gadolina@mail.ru

Аннотация. Рассмотрено создание научно обоснованного блока нагружения, который учитывал бы возможные режимы работы в правильной пропорции, с учетом вариабельности и неопределенности (размытости, нечеткости). Это связано с тем, что усталостное повреждение накапливается в течение всего срока работы машины и должно быть научно оценено для адекватной оценки в вероятностном аспекте. Так как эксплуатация некоторой конкретной детали точно не определена (и не может быть определена по логике случайного использования машин), рассматриваются проекции случайных нечетких распределений. Удалось научно обоснованно учесть конечное множество эксплуатационных режимов в их разумной пропорции. На примере анализа нагружения ответственной детали подвижного состава построены распределения и оценено возможное распределение ресурса детали. Полученные на основе разработанного метода результаты позволят оценить риски эксплуатации и спрогнозировать потребное число запчастей. Рассмотрение цензурированных элементов выборки при построении кривой усталости позволит сделать оценку параметров кривой усталости более состоятельной. Применение нечетких множеств может оказаться весьма полезным при рассмотрении кривой усталости и при оценке вариации долговечности. Показаны примеры применения предложенного подхода.

Ключевые слова: нечеткие множества, оценка долговечности, режимы нагружения, кластерный анализ, временные ряды

Для цитирования: *Гадолина И.В.* Применение нечетких данных в задачах оценки долговечности // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2022. № 4. С. 59–74. DOI: 10.28995/2686-679X-2022-4-59-74

Application of fuzzy data in durability assessment tasks

Irina V. Gadolina
Mechanical Engineering Research Institute of the Russian Academy of Sciences (IMASH RAN), Moscow, Russia, gadolina@mail.ru

Abstract. The author considers the creation of a scientifically based loading unit which would take into account the possible modes of operation in the correct proportion, with due regard for variability and unclearness (fuzziness). It is due to the fact that the fatigue damage accumulates throughout the life of the machine and must be scientifically evaluated for an adequate assessment in the probabilistic aspect. Since the operation of some specific part is not precisely defined (and cannot be determined by the logic of random use of machines during exploitation), projections of random fuzzy distributions are considered. It worked out possible to scientifically take into account a finite set of operating modes in their reasonable proportion. Analyzing the loading of a critical performance part of the rolling stock served as example for construction of distributions and in particular for estimation of the part's resource distribution. The results obtained on the basis of the developed method will allow assessing the risks of operation and predicting the required number of spare parts. Consideration of the censored sample elements in the construction of the fatigue curve will make the estimation of the parameters of the fatigue curve more consistent. The use of fuzzy sets can be very useful when considering the fatigue curve and when evaluating the durability variation. Examples of application of the proposed approach are shown.

Keywords: fuzzy sets, durability assessment, loading modes, cluster analysis, time series

For citation: Gadolina, I.V. (2022), "Application of fuzzy data in durability assessment tasks", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 4, pp. 59–74, DOI: 10.28995/2686-679X-2022-4-59-74

Введение

Одной из парадигм современной статистики [Орлов 2014] является статистическое исследование объектов нечисловой природы. Оно исследуется наряду с непараметрической статистикой. Нечеткие множества в процессе работы с непараметрической ста-

тистикой и анализе объектов нечисловой природы [Орлов 2021] играют немаловажную роль.

При рассмотрении задачи оценки ресурса деталей машин в вероятностной постановке [Когаев 1993] одной из основных инженерных задач является задача построения научно обоснованного обобщенного блока (ОБ). ОБ – это некоторый массив нагрузок, который в дальнейшем применяется для оценки долговечности совместно с гипотезами суммирования усталостных повреждений. Особенно важно иметь научно обоснованный инструмент создания этого блока. В свете применения цифровой трансформации в промышленности важно отрабатывать специфические информационные технологии, которые будут призваны в дальнейшем усовершенствовать MES (Manufacturing Execution System) – производственная исполнительная система. Специализированные программные комплексы MES, предназначенные для решения задач оперативного планирования и управления производством, базируются на научных методах оценки необходимого числа запчастей, что, в свою очередь, основывается на усовершенствованных методах оценки ресурса деталей. Все возможные режимы эксплуатации должны быть отражены в ОБ, причем с учетом актуальной пропорции конкретного режима в общем времени (или километраже) использования. ОБ относится ко категорированным данным, поэтому используются понятия объектов нечисловой природы. После определения конкретных режимов, входящих в ОБ с помощью кластерного анализа [Гадолина, Петрова 2021] оценок экспертов или с применением нечетких множеств, дальнейшие исследования осуществляются с вещественными числами.

В настоящее время обобщенный блок строится следующим образом. Расчетные характеристики нагруженности деталей основываются на замене реального случайного процесса, записанного путем тензометрирования, некоторым схематизированным процессом. Смысл схематизации заключается в том, что схематизированный процесс по уровню вносимого усталостного повреждения предполагается эквивалентным исходному [Гадолина, Козлов, Монахова, Серебрякова 2019]. При этом некоторый режим может оказаться весьма незначительным по нагрузкам, но его относительную продолжительность эксплуатации необходимо учесть в ОБ при расчете ресурса.

Был проанализирован возможный процесс нагружения в детали несущей конструкции рамы зерноуборочного комбайна. При этом было замечено, что наблюдаются участки с большей интенсивностью колебаний – «А» (соответствующие движению комбайна по проселочной грунтовой дороге) и режимы работы

в поле – участки «Б». Участки «Б» являются щадящими с точки зрения усталостного повреждения для исследуемой детали. Анализ истории нагружения детали привел к выводу о возможности проведения ускоренных ресурсных испытаний лишь на режимах «А». Окончательный вывод о расчетном ресурсе должен базироваться на информации о процентном соотношении времени эксплуатации в режимах «А» и «Б». Таким образом, они оба должны присутствовать в обобщенном блоке ОБ.

Вопрос выбора и обоснования ОБ неоднократно затрагивался в литературе. В области автомобилестроения ученые из ФРГ разработали систему анализа процесса нагружения с применением современных цифровых инструментов: GSP (Global Positioning System – глобальная система позиционирования), big data (большие данные) [Burger, Dreßler, Speckert 2021]. Проблемы вероятностного распределения режимов по нагрузкам исследуются также в [Dressler, Speckert, Müller, Weber 2009] [Bellec 2021]. Авторы анализировали вариабельность нагрузок в зависимости от места приписки машины, при этом анализировали вероятность события попадания машины в определенную категорию использования по логистике. В работе [Stadele, Mundl, Rap 2020] был разработан метод для надежного представления конкретных требований заказчика при производстве новых транспортных средств. В этом методе типы дорог могут сравниваться с точки зрения важных параметров, таких как динамика продольной и поперечной нагрузки. Типы дорог могут быть также сопоставлены по вычисленным усталостным повреждениям и применению гипотез накопления повреждений [Когаев 1993].

Ранее была рассмотрена задача выделения режимов с использованием кластерного анализа характерных показателей нагрузки [Гадолина, Петрова 2021]. Для анализа временных зависимостей при кластеризации был использован метод изучения временных последовательностей параметров выборки. Обзор данных методов содержится в [Aghabozorgi, Shirkhorshidi 2015]. Кластерный анализ, основанный на биологических образцах, описан в [Плющенко, Шахматов, Родин 2020]. На рис. 1 показана визуализация трех факторов, которые имеют решающее значение для оценки долговечности машин. На самом деле было использовано большее количество факторов, здесь число три было выбрано для большей наглядности. Данные, представленные на рис. 1, соответствуют задаче кластерного анализа комбинированного режима SAE [Tucker, Bussa 1977].

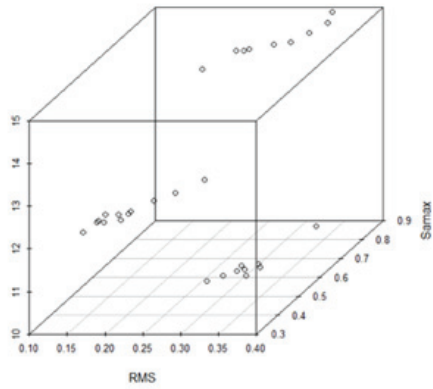


Рис. 1. Три фактора (эффективная частота, СКО, максимальная амплитуда в блоке) (пример автомобилей – модельная выборка SAE [Tucker, Bussa 1977])

Согласно предложенному ранее алгоритму [Гадолина, Петрова 2021] в качестве первого шага предлагалось создать список всех возможных режимов по предложениям экспертов для данной конкретной машины. Далее на каждом режиме проводится тензометрирование напряжений на интересующей детали с последующей обработкой записи по методу дождя. Полученные распределения следует просуммировать с учетом их весов (при умножении на веса целые числа частостей в гистограммах превратятся в вещественные – см. формулу (1)) и длительности записи на конкретном режиме. После применения процедур дискретизации, выделения экстремумов и схематизации по методу дождя [Гадолина, Козлов, Монахова, Серебрякова 2019] для каждого из различных вариантов использования механизма (режимов, регламентированных условий работы) получают распределение амплитуд нагружения, или спектр нагружения, характеризующий данный вариант (например, езда автомобиля по проселочной дороге или движение по асфальту). Назовем его частной функцией или спектром.

Всего может быть рассмотрено k спектров в соответствии с условиями эксплуатации. Для оценки долговечности необходимо располагать обобщенным спектром, который формируется путем сложения частных спектров, соответствующих различным вариантам использования механизма с указанием доли этих вариантов в общей продолжительности эксплуатации. Проблема при этом может заключаться в том, что границы интервалов разбивки непрерывной

физической величины напряжений (так называемые «карманы») для подсчета числа попадающих в них чисел циклов для разных режимов могут быть различными [Петрова, Гадолина 2018].

Возможным вариантом решений данной проблемы может явиться аппроксимация гистограммы непрерывным распределением с помощью непараметрического ядерного сглаживания. Гистограмма вместо отдельных «кубиков» в этом случае будет строиться из мини-распределений заданной формы: треугольников, нормальных распределений и т. п. Суммируя все составляющие такой «гистограммы», мы получим сглаженную кривую. В программном комплексе R для этой цели использовалась функция “density” с ядром, установленным по умолчанию, “gaussian”. Сглаженные кривые для нескольких режимов эксплуатации суммируются с учетом распределения режимов.

Обобщенный блок подобен гистограмме, только число повторений событий в карманах может представлять собой положительную вещественную величину. Это связано с применением формулы (1) для расчета распределения ОБ. Коэффициенты для пересчета с целью последующего суммирования приведенных частных распределений z_i вычисляются по формуле:

$$z_i = \frac{3600}{l_i} p_i, \qquad (1)$$

здесь l_i –продолжительность реализации, зафиксированной на i – том режиме, p_i – доля режима в эксплуатации. Таким образом, для каждого значения амплитуд напряжений в режиме i был определен коэффициент z_i, единый для данной реализации.

Альтернативой описанному методу является использование непрерывного распределения, полученного ядерным сглаживанием гистограмм. В этом случае появляется возможность численного суммирования псевдо-непрерывных распределений. В программном комплексе R имеется возможность произвести сглаживание с помощью различных ядер: гауссовского, Епанченкова, прямоугольный, треугольный, косинус. Ядра масштабируются таким образом, что его стандартное отклонение соответствует ядру сглаживания. Алгоритм, используемый в программе R по оценке плотности по умолчанию, распределяет массу эмпирической функции распределения по регулярной сетке из не менее 512 точек, а затем использует быстрое преобразование Фурье для свертки этой аппроксимации с дискретизированной версией ядра. Следующим шагом является линейная аппроксимация для оценки плотности в указанных точках.

Примем сглаживание с помощью гауссовского ядра, предлагаемого программой R по умолчанию. Форма ядра Гаусса описывается формулой:

$$G(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{x^2+y^2}{2\sigma^2}}, \quad (2)$$

здесь $G(x, y)$ – значение, рассчитанное по формуле ядра Гаусса. Это значение является частью ядра, представляющего один элемент; σ – пороговое значение или значение фактора, указанное пользователем; x, y – переменные, обозначенные как x и y, относятся к координатам пикселей, в изображении y представляет вертикальное смещение или строку, а x представляет горизонтальное смещение или столбец.

На примере гистограмм амплитуд полных циклов в детали подвижного состава для пяти скоростей движения применим сглаживание гауссовским ядром. Сглаженные гистограммы для пяти режимов показаны на рис. 2.

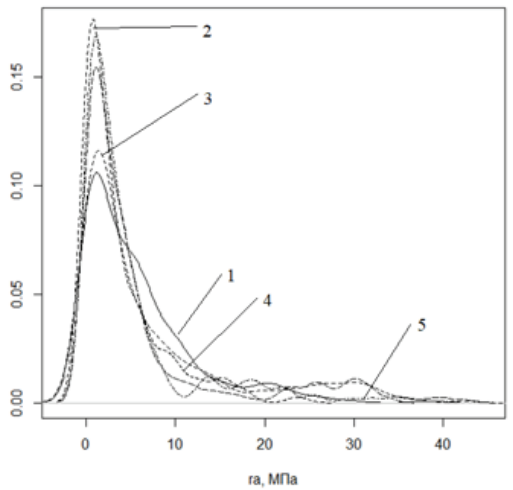


Рис. 2. Сглаженные гистограммы
для распределений амплитуд полных циклов Ra
для напряжений в детали несущей конструкции
при движении состава с различными скоростями:
1 – V = 45 км/ч; 2 – V = 54 км/ч; 3 – V = 63 км/ч;
4 – V = 90 км/ч; 5 – V = 99 км/ч.

Для создания обобщенного блока псевдо-непрерывные значения сглаженных гистограмм умножаются на индивидуальные коэффициенты, рассчитываемые по формуле, аналогичной (1), при этом учитывается также продолжительность реализации, которая подвергалась схематизации по методу дождя [Гадолина, Козлов, Монахова, Серебрякова 2019]. Далее будет показано, что для оценки вариабельности вычисленного ресурса хорошо подходит аппарат нечетких множеств. На следующем этапе приведенные частные гистограммы суммируются для получения результирующего ОБ. В силу сделанных предположений о нечетком характере множества распределений режимов ОБ также является нечетким.

Метод

Помимо задачи с отчетливо выраженными режимами, которые можно дистанцировать непосредственно по визуальному анализу записей либо с применением кластерного анализа, описанного в [Гадолина, Петрова 2021], существует обширный класс задач, где такое разделение является нечетким, размытым. Применим концепцию нечетких множеств. В математике нечеткие множества (они же неопределенные множества) – это множества, элементы которых имеют разные степени принадлежности.

На рис. 3 схематически показан пример функции принадлежности нечетких множеств. Здесь схематически показана нечеткая принадлежность индивидуума к множествам людей: А (молодой), Б (средний) и В (старый). Рассматривая функцию принадлежности, можно увидеть некоторую аналогию с вероятностями.

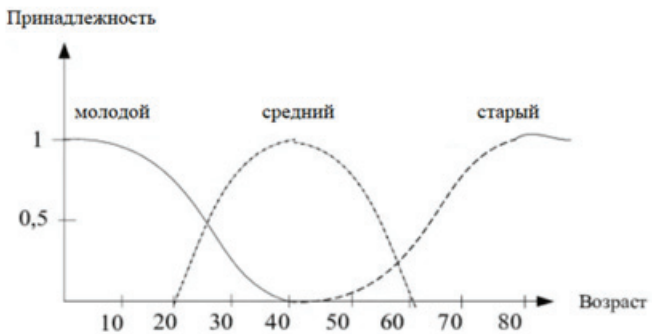


Рис. 3. Схема функций
принадлежности трех нечетких множеств

В [Орлов 2014] авторы проводят параллели между теорией вероятности и нечеткостью. Они полагают, что связь между нечеткостью и вероятностью позволит применить в нечеткой теории методы и результаты, накопленные в теории случайных множеств. И наоборот, это позволит перенести концепции и формулирование задач из первой теории на вторую, что послужит прогрессом у обеих. Подобно тому, как решению задачи о кластерах предшествует предварительное решение о числе кластеров, в задаче оценивания с применением нечетких множеств задача разбивается на две: оценивание структуры моделей и оценивание параметров при заданной структуре [Орлов 2014]. При этом структура – это объект нечисловой природы.

Спецификация нечеткого объекта A представляет собой вектор вида

$$P(A) = (p_1(A),...,p_n(A)), \quad (3)$$

где $p_i(A)$, $i = 1, n$, – количественные или качественные свойства объекта A.

Нечетким свойством объекта A является кортеж (т. е. набор данных фиксированной длины) – это нечеткое множество, описывающее количественные значения свойства p(A) и качественные u(p(A)). Данные параметры являются единицами измерения объекта. Совпадение объектов определяется совпадением всех параметров $i = 1, n$ [Aghabozorgi, Shirkhorshidi 2015]. Другими словами, спецификация нечеткого объекта может состоять из четко определенных (crisp) и нечетких (fuzzy) свойств. Качественные свойства являются частными случаями нечетких количественных свойств.

Рассмотрим данные на кривой усталости как нечеткие объекты типа (3). Как известно [Когаев 1993], на кривой усталости содержатся элементы двух типов: разрушенные и цензурированные (приостановленные, снятые с испытаний) образцы. У обоих перечисленных объектов имеется количественная характеристика – наработка t (размерность [мин], [циклы] или другая мера), а также качественная мера I, определяющая статус наблюдения:

$$I \in [0,1]. \quad (4)$$

В простейшем случае при отсутствии уточняющей информации о статусе цензурированного образца для сломавшихся образцов $I = 1$ и для цензурированных $I = 0$. В настоящее время нами ведутся исследования по введению уточняющих параметров для I,

которые могут быть получены дополнительными измерениями физических величин (томография, измерение модуля упругости, микрофотографирование). В общем случае данная характеристика представляет нечеткую характеристику, подобную изображенной на схеме (рис. 3).

Автор утверждает:

> Теория нечетких множеств сводится к теории случайных множеств с использованием понятия «проекция случайного множества». Каждое случайное множество может быть связано с некоторой функцией – вероятностью того, что элемент принадлежит к множеству. Эта функция имеет все свойства функции принадлежности нечеткого множества. Соответствующее нечеткое множество называется проекцией исходного случайного множества. Верно и обратное – для любого нечеткого множества можно выбрать случайное множество так, чтобы вероятность принадлежности элемента к случайному множеству повсюду совпадала с функцией принадлежности данного нечеткого множества. Такое соответствие можно установить так, чтобы результаты операций над множествами тоже соответствовали друг другу [Орлов 2014].

Классификация относится к разделению набора элементов на классы – группы элементов, похожих друг на друга. В четкой классификации каждый элемент относится к одному определенному классу. А в нечеткой классификации задана функция принадлежности элемента к разным классам. Нечеткая классификация обычно больше соответствует действительности, чем строгая [Terletskyi, Provotar 2015].

В основе современной математики лежит понятие множества. Для того чтобы установить тот или иной конкретный набор режимов эксплуатации, необходимо уметь ответить по каждому режиму на вопрос: «Принадлежит этот режим этому множеству или не принадлежит?» Но границы понятий обычно размыты, так что однозначный ответ на такой вопрос возможен далеко не всегда. Это означает, что размытие нужно описывать в терминах множества, которое немного отличается и шире, чем обычно.

В работе [Stadele, Mundl, Rap 2020] отмечено, что распределение по режимам для грузовых автомобилей меняется от региона к региону. Рассмотрим некоторое распределение продолжительности работы по режимам для рассматриваемого примера грузовых вагонов. Допустим, имеются некоторые возможные подвижки по долям времени машин по различным режимам. Схематически эта ситуация показана на диаграммах табл. 1. В строке под изображе-

ниями круговых распределений показаны вероятности p_{il} появления режимов i в j-той комбинации:

$$\sum p_i = 1, \text{ для всех } j. \quad (5)$$

Результаты

Данные вероятности являются проекциями нечетких множеств на пространство событий.

Далее проанализируем нечеткое распределение по комбинациям j. С учетом того, что распределения по p_i также нечеткое, задача может быть сформулирована как дважды размытая. Поскольку рассматриваются вероятности нахождения объекта в некотором состоянии (режиме) i, проанализируем проекцию [Орлов 2014] нечетких множеств.

Пусть X – случайное множество – измеримое отображение {K} семейства элементарных исходов произвольного вероятностного пространства { {Q}, {A}, {P} } в некоторое пространство{M}, элементами которого являются множества распределений (пример в табл. 1).

Таблица 1

Представление режимов нагружения детали железнодорожного состава как нечеткого множества

I	II	III	IV	V
А: 4.7%	3.3%	4.4%	6.8%	3.9%
Б: 7.0%	6.6%	6.8%	7.2%	8.7%
В: 16.7%	14.0%	19.5%	14.9%	13.9%
Г: 34.4%	38.8%	35.1%	31.7%	28.1%
Д: 37.2%	37.2%	34.1%	39.4%	45.5%
$V_{m-6} = 0.433$	$V_{m-6} = 0.476$	$V_{m-6} = 0.491$	$V_{m-6} = 0.518$	$V_{m-6} = 0.503$

Вместо вероятностей p_i поставим меру нечеткости μ_i из теории нечетких множеств [Meyer, Dimitriadou, Hornik, Weingessel, Leisch 2017]. Тот факт, что в отличие от теории вероятности, где справедливо соотношение (5) и сумма функций принадлежности не равна единице:

$$\sum \mu_i \neq 1. \quad (6)$$

Данный факт не нарушает общности рассуждений, но дает возможность оценить расстояние между нечеткими множествами A и B множества $X = \{x_1, x_2 .... x_k\}$ [Орлов 2014] как

$$d(A, B) = \sum_{j=1}^{k} |\mu_A(x_j) - \mu_B(x_j)|, \quad (7)$$

где $\mu_A(x_j)$ – функция принадлежности нечеткого множества A и где $\mu_B(x_j)$ – функция принадлежности нечеткого множества B.

Выражение (4) приведено в [Орлов 2014] для двух нечетких множеств. Как видно на примере рис. 3, этих множеств может быть больше в случае конечного k числа эксплуатационных режимов, диагностированных кластерами.

Обсуждение результатов

Используя проекции нечетких множеств на пространство событий, оценим нечеткое событие как случайный ресурс детали. Для оценки рассеивания используется мера коэффициента полноты спектра V [Савкин 2017] вычисляется по формуле (8) для разных распределений режимов, показанных в таблице:

$$V = \sqrt[m]{\frac{1}{n}\sum h_i\left(\frac{\sigma_{ai}}{\hat{\sigma}_a}\right)^m}. \quad (8)$$

Величина V является безразмерной. В формуле (8) m – коэффициент угла наклона кривой усталости; n – суммарное число циклов в блоке; h_i – число циклов на i-той ступени; σ_{ai} – текущее значение амплитуды напряжений; $\hat{\sigma}_a$ – максимальная амплитуда в блоке. Видно, что V зависит не только от формы блока, но и от m. Примем $m = 6$ и вычисленные значения занесем в последнюю строку таблицы.

Распределение случайной величины V, полученной на основании анализа нечетких кластеров, показано на нормальной вероятностной бумаге на рис. 4. График построен в R. По вертикали отложены выборочные квантили, по горизонтали – теоретические. Характер вариации данной случайной величины подобен вариации вычисленного ресурса исследуемой детали.

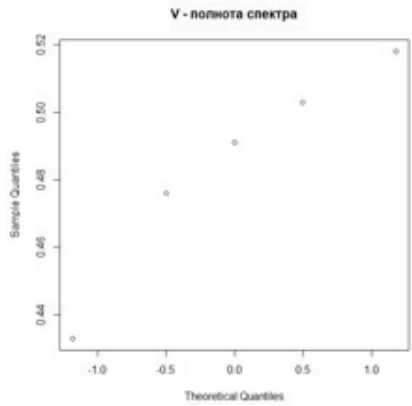


Рис. 4. Выборочные и теоретические квантили для V

Выводы

Рассмотрено применение теории нечетких множеств к задачам усталости. Проанализировано два типа задач: 1) построение кривой усталости с учетом цензурирований; 2) задача нечеткой кластеризации режимов нагружения. В задаче о построении кривой усталости предложено рассматривать цензурирования как нечеткие элементы с функцией принадлежности, лежащей в интервале [0,1], где 0 – это чистое цензурирование, а 1 – безусловный отказ (четкие данные). Промежуточные варианты предлагается оценивать с использованием перспективных физических методов исследования. Распределение общего времени эксплуатации по режимам также является нечетким в силу до конца не определенных условий эксплуатации. Рассмотрение данной задачи с точки зрения нечетких данных позволило предварительно оценить расчетное рассеивание ресурса детали.

Литература

Гадолина, Козлов, Монахова, Серебрякова 2019 – *Гадолина И.В., Козлов А.Д., Монахова А.А., Серебрякова И.Л.* Оптимальный способ ЦОС в задачах оценки долговечности // Вестник РГГУ. Серия: Информатика. Информационная безопасность. Математика. 2019. № 1. С. 78–93.

Гадолина, Петрова 2021 – *Гадолина И.В., Петрова И.М.* Решение проблемы построения представительного блока нагружения с использованием аппарата

кластеризации // Вестник РГГУ. Серия: Информатика. Информационная безопасность. Математика. 2021. № 4. С. 69–79.

Когаев 1993 – *Когаев В.П.* Расчеты на прочность при напряжениях, переменных во времени. М.: Машиностроение, 1993. 364 с.

Орлов 2014 – *Орлов А.И.* Нечисловая статистика. М.: МЗ-Пресс, 2014. 513 с.

Орлов 2021 – *Орлов А.И.* Смена парадигм в прикладной статистике // Заводская лаборатория. Диагностика материалов. 2021. № 87 (7). С. 6–7. DOI: https://doi.org/10.26896/1028-6861-2021-87-7-6-7.

Петрова, Гадолина 2018 – *Петрова И.М., Гадолина И.В.* Создание обобщенного спектра нагружения при различных вариантах нагружения в эксплуатации // Технологическое оборудование для горной и нефтегазовой промышленности: Сб. трудов XVI Международной научно-технической конференции в рамках Уральской декады / Под ред. Ю.А. Лагуновой. Екатеринбург: Уральский государственный горный ун-т, 2018. С. 318–321.

Плющенко, Шахматов, Родин 2020 – *Плющенко В.П., Шахматов Д.Г., Родин И.А.* Алгоритм сочетания хромато-спектрометрического ненаправленного профилирования и многомерного анализа для выявления веществ-маркеров в образцах сложного состава // Заводская лаборатория. Диагностика материалов. 2020. Т. 86. № 7. С. 12–19.

Савкин 2017 – *Савкин А.Н.* Компьютерное моделирование и анализ прочности конструкций при переменном нагружении: монография. Волгоград: ВолгГТУ, 2017. 228 с.

Aghabozorgi, Shirkhorshidi 2015 – *Aghabozorgi S., Shirkhorshidi A.S.* Time-series clustering – a decade review // Information Systems. 2015. Vol. 53. P. 16–38.

Bellec 2021 – *Bellec E. at al.* Modelling and identification of fatigue load spectra: Application in the automotive industry // International Journal of Fatigue. 2021. Vol. 149. P. 106–222.

Burger, Dreßler, Speckert 2021 – *Burger M., Dreßler K., Speckert M.* Load assumption process for durability design using new data sources and data analytics // International Journal of Fatigue. 2021. Vol. 145. P. 106–116.

Dressler, Speckert, Müller, Weber 2009 – *Dressler K., Speckert M., Müller R. Weber C.* Customer loads correlation in truck engineering. Conference // World automotive Congress 2008. Kaiserslautern: Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, 2009.

Meyer, Dimitriadou, Hornik, Weingessel, Leisch 2017 – *Meyer D., Dimitriadou E., Hornik K., Weingessel A., Leisch F.* E1071: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien, 2017. URL: https://CRAN.R-project.org/package=e1071 (дата обращения 21 октября 2022).

Stadele, Mundl, Rap 2020 – *Stadele M., Mundl B., Rap A.* Derivation and categorization of road quality in view of operation loads to consider customer requirements for customer-focused sales // Proceedings of the Fourth International Conference on Material and Component Performance under Variable Amplitude Loading (VAL4), March 30 – April 1 2020, Darmstadt, Germany. Berlin: DVM, 2020. P. 175–183.

Terletskyi, Provotar 2015 – *Terletskyi D.A., Provotar A.I.* Fuzzy Object-Oriented Dynamic Networks. I // Cybern. Syst. Anal. 2015. № 51. P. 34–40. DOI: https://doi.org/10.1007/s10559-015-9694.

Tucker, Bussa 1977 – *Tucker L., Bussa S.* The SAE Cumulative Fatigue Damage Test Program // Fatigue under Complex Loading / R.M. Wetzel (ed.). Warrendale, PA: Soc. Autom. Eng., 1977. P. 1–44.

References

Aghabozorgi, S. and Shirkhorshidi, A.S. (2015), "Time-series clustering – a decade review", *Information Systems*, vol. 53, pp. 16–38.

Bellec, E.at al. (2021), "Modelling and identification of fatigue load spectra: Application in the automotive industry", *International Journal of Fatigue*, vol. 149, pp. 106–222.

Burger, M., Dreßler, K. and Speckert, M. (2021) "Load assumption process for durability design using new data sources and data analytics", *International Journal of Fatigue*, vol. 145, pp. 106–116.

Dressler, K., Speckert, M., Müller, R. and Weber, C. (2009), "Customer loads correlation in truck engineering", *World automotive Congress 2008*, München, Fraunhofer-Institut für Techno- und Wirtschaftsmathematik, Kaiserslautern, Germany.

Gadolina, I.V. and Petrova, I.M. (2021), "Solving the issue of constructing a representative loading block using the clustering apparatus", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 4, pp. 69–79.

Gadolina, I.V., Kozlov, A.D., Monakhova, A.A. and Serebryakova, I.L. (2019), "Optimal discretization method in problems of durability assessment", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series*, no. 1, pp. 78–93.

Kogaev, V.P. (1993), *Raschety na prochnost' pri napryazheniyakh, peremennykh vo vremeni* [Calculations for strength under stresses, variables in time], Moscow, Russia, 364 p.

Meyer, D., Dimitriadou, E., Hornik, K., Weingessel, A., and Leisch, F. (2017), "E1071: Misc Functions of the Department of Statistics, Probability Theory Group, TU Wien", available at: https://CRAN.R-project.org/package=e1071 (Accessed 21 October 2022).

Orlov, A.I. (2014), *Nechislovaya statistika* [Non-numerical statistics], MZ–Press, Moscow, Russia, 513 p.

Orlov, A.I. (2021), "Paradigm shifts in applied statistics", *Zavodskaya Laboratoriya, Diagnostika Materialov*, vol. 87 (7), pp. 6–7. DOI: https://doi.org/10.26896/1028-6861-2021-87-7-6-7

Petrova, I.M. and Gadolina, I.V. (2018), "Creation of a generalized loading spectrum for various loading variants in operation, Technological equipment for the mining and oil and gas industry", in Lagunova, Yu.A., (ed.), *Proceedings of the XI Scientific and Technical International Conference within the framework of the Ural Decade*, Ural State Mining University, Ekaterinburg, Russia, pp. 318–321.

Pluschenko, V.P., Shakhmatov, D.G. and Rodin, I.A. (2020), "Algorithm for combining chromatographic-spectrometric non-directional profiling and multidimensional analysis for the detection of marker substances in samples of complex composition", *Zavodskaya Laboratoriya, Diagnostika Materialov*, vol. 86, no. 7, pp. 12–19.

Savkin, A.N. (2017), *Komp'yuternoe modelirovanie i analiz prochnosti konstruktsii pri peremennom nagruzhenii: monografiya* [Computer modeling and analysis of structural strength under variable loading. Monograph], Volgograd State Technical University, Volgograd, Russia.

Stadele, M., Mundl, B. and Rap, A. (2020), "Derivation and categorization of road quality in view of operation loads to consider customer requirements for customer-focused sales", *Proceedings of the Fourth International Conference on Material and Component Performance under Variable Amplitude Loading (VAL4)*, March 30 – April 1 2020, Darmstadt, Germany, DVM, Berlin, Germany, pp. 175–183.

Terletskyi, D.A. and Provotar, A.I. (2015), "Fuzzy Object-Oriented Dynamic Networks. I", *Cybern. Syst. Anal.*, vol. 51, pp. 34–40. DOI: https://doi.org/10.1007/s10559-015-9694.

Tucker, L. and Bussa, S. (1977), "The SAE Cumulative Fatigue Damage Test Program", in Wetzel, R.M. (ed.), *Fatigue under Complex Loading*, Soc. Autom. Eng., Warrendale, PA, USA, pp. 1–44.

Информация об авторе

Ирина В. Гадолина, кандидат технических наук, Институт машиноведения им. А.А. Благонравова Российской академии наук (ИМАШ РАН), Москва, Россия; 101000, Россия, Москва, Малый Харитониевский пер., д. 4; gadolina@mail.ru

Information about the author

Irina V. Gadolina, Cand. of Sci. (Engineering), Mechanical Engineering Research Institute of the Russian Academy of Sciences (IMASH RAN), Moscow, Russia; bld. 4, Malyi Kharitonievskii lane, Moscow, Russia, 101000; gadolina@mail.ru

УДК 519.2
DOI: 10.28995/2686-679X-2022-4-75-91

Частотные свойства лексики научных текстов и законы Ципфа высших порядков

Вячеслав Ю. Синицын
Российский государственный гуманитарный университет, Москва, Россия, fpmrggu@yandex.ru

Валентина С. Кашпарова
Московский педагогический государственный университет, Москва, Россия, vs.kashparova@mpgu.su

Аннотация. Статистические свойства лексики текстов на естественных языках за последние сто лет привлекали к себе пристальное внимание многих математиков и лингвистов. Основной закон о зависимости частоты слова от его ранга, который известен под именем закона Ципфа, состоит в том, что произведение частоты слова на его ранг приблизительно постоянно и является маркером языка текста. Следует отметить, что важное прикладное значение для описания функционирования различных социотехнических систем имеют многочисленные аналоги закона Ципфа в других предметных областях: закон Ауэрбаха о распределении городов по численности населения, закон Парето о распределении материальных благ в обществе, закон Бредфорда о распределении ученых по продуктивности, закон Лотке о распределении публикаций в библиографических источниках.

В данной работе рассматриваются латентные частотные характеристики лексики научных текстов различной тематики на массиве статей мультидисциплинарного журнала открытого доступа «Молодой ученый». В работе сформулировано понятие таблицы частот высшего порядка и эмпирически исследована зависимость частот в таких таблицах от их рангов. Для всех статей из рассматриваемого корпуса текстов построены степенные, гиперболические и другие двухпараметрические модели зависимости частот от рангов в таблицах частот высших порядков. Построенные модели являются обобщениями широко известных моделей Ципфа и имеют высокие показатели качества. В работе получены новые предикторы, которые могут быть полезны для решения задач классификации научных текстов.

Ключевые слова: закон Ципфа, обобщенная финитная модель Ципфа, автоматическая классификация текстовых документов

Для цитирования: Синицын В.Ю., Кашпарова В.С. Частотные свойства лексики научных текстов и законы Ципфа высших порядков // Вестник РГГУ. Серия «Информатика. Информационная безопасность. Математика». 2022. № 4. С. 75–91. DOI: 10.28995/2686-679X-2022-4-75-91

Frequency properties of the lexis of scientific texts and Zipf's laws of higher orders

Vyacheslav Yu. Sinitsyn
Russian State University for the Humanities, Moscow, Russia,
fpmrggu@yandex.ru

Valentina S. Kashparova
Moscow Pedagogical State University, Moscow, Russia,
vs.kashparova@mpgu.su

Abstract. The statistical properties of the lexis of texts in natural languages have attracted the close attention of many mathematicians and linguists over the past hundred years. The basic law of the dependence of the word frequency on its rank, which is known as Zipf's law, is that the product of the frequency of a word and its rank is approximately constant and is a marker of the text language. It should be noted that numerous analogs of Zipf's law in other subject areas have great practical significance for describing the functioning of various sociotechnical systems: Auerbach's law of the distribution of cities by population, Pareto's law of the distribution of material goods in society, Bradford's law of the distribution of scientists by productivity, Lotka law of the distribution of publications in bibliographic sources.

The article considers the latent frequency characteristics of the lexis of scientific texts on various topics by the array of articles in the multidisciplinary open access journal "Young Scientist". The paper formulates the concept of a higher-order frequency table and empirically investigates the dependence of frequencies in such tables on their ranks. Power, hyperbolic, and other two-parameter models of the dependence of frequencies on ranks in tables of frequencies of higher orders were constructed for all articles from the corpus of texts under consideration. The constructed models are generalizations of the well-known Zipf models and have high quality indicators. In the paper, new predictors are obtained, which can be useful for solving problems of classifying scientific texts.

Keywords: Zipf's law, generalized finite Zipf model, automatic classification of text documents

For citation: Sinitsyn, V.Yu. and Kashparova, V.S. (2022), "Frequency properties of the lexis of scientific texts and Zipf's laws of higher orders", *RSUH/RGGU Bulletin. "Information Science. Information Security. Mathematics" Series,* no. 4, pp. 75–91, DOI: 10.28995/2686-679X-2022-4-75-91

Введение

Согласно широко известному результату Ципфа, который он опубликовал в середине прошлого века, частоты слов в текстах на естественном языке приблизительно обратно пропорционально зависят от рангов этих частот при упорядочивании частот по убыванию [Zipf 1949]. Модель зависимости частот слов от рангов была обнаружена Ципфом эмпирически и является простой однопараметрической моделью. Обоснованию и уточнению модели Ципфа посвящено большое число работ [Шрейдер 1967] [Мандельброт 1973] [Маслов, Маслова 2006] и другие. Качество прогнозирования частот при помощи модели Ципфа значительно различается для малых, средних и больших рангов. Константа Ципфа классической модели, как хорошо известно, является маркером языка текста, но использование ее в качестве единственного предиктора недостаточно для того, чтобы прогнозировать тематику текста или предсказывать авторство текста.

В 2019 г. был подготовлен корпус научных текстов журнала «Молодой ученый» для частотного анализа и уточнения закономерностей, найденных Ципфом. В работе [Агеев, Кашпарова, Синицын 2019] построены финитные модели Ципфа, которые описывают зависимость частот от рангов для различных величин рангов. Построенные финитные модели Ципфа для научных текстов различной тематики позволили решить некоторые задачи автоматической классификации статей по рубрикам мультидисциплинарного журнала «Молодой ученый».

В 2020 г. были рассмотрены обобщенные финитные модели Ципфа для сравнительного исследования частотных свойств научных текстов различной тематики. В качестве уточнения классической зависимости Ципфа частот слов от их рангов были выбраны четыре типа двухпараметрических моделей: степенная модель, гиперболическая модель, модель Ципфа с квадратичной поправкой и модель Ципфа со свободным членом [Гламаздин, Гордин, Синицын, Кашпарова 2020]. В результате вычислений для каждой научной статьи построено 767 обобщенных финитных моделей Ципфа и получено 2630 предикторов, среди которых три общих частотных предиктора: число слов в статье, число уникальных слов, число уникальных рангов.

Для исследования полезности предикторов были использованы критерий Уилкоксона и критерий Колмогорова–Смирнова. Предикторы были разделены на три группы, к первой из них относятся наиболее полезные предикторы, способные распознать более 80% пар секций, ко второй группе относятся предикторы, способные распознать от 50 до 80% пар секций, оставшиеся предикторы были отнесены к третьей группе.

Некоторые секции научного журнала оказались достаточно близки с точки зрения их частотных свойств, и при бинарной классификации отличить такие секции одну от другой достаточно сложно. Так, пару секций Сельское хозяйство – Химия различают всего 40 предикторов, а пару секций Государство и право – Биология – 198 предикторов. Также близкими являются пары секций Биология – Химия, Социология – География, Медицина – Химия.

В работе [Гламаздин, Синицын 2021] для каждой пары секций журнала «Молодой ученый» были построены классификаторы для статей с использованием методов деревьев решений, случайного леса, бэггинга и бустинга деревьев решений. Следует отметить, что задача парной классификации не была решена полностью, так как из 231 пары секций журнала 75 пар секций не могут распознаваться адекватно ни одним из построенных классификаторов.

В работе [Гордин, Синицын 2021] для каждой пары секций мультидисциплинарного журнала «Молодой ученый» построены классификаторы, основанные на логистической регрессии, которые позволяют выполнять соотнесение статей журнала с соответствующими секциями без использования информации о семантике текстов. Для большинства пар секций журнала построенные классификаторы имеют хорошие показатели качества, но для некоторых пар секций требуется повышение качества классификации.

Целью данной работы является поиск и исследование новых эмпирических закономерностей для частотных свойств научных текстов различной тематики, а также построение математических моделей, которые могут быть полезны для решения задач автоматического соотнесения научных статей рубрикам мультидисциплинарного журнала без использования информации о семантике текстов.

Таблицы частот высших порядков

Фундаментальное значение для исследования частотных свойств текста имеет таблица частот слов, содержащихся в этом тексте. Таблица частот слов состоит из двух столбцов. В первом

столбце перечислены все уникальные слова из текста, а во втором – их частоты в порядке убывания. Частоты слов текста ранжируются по следующим общим правилам: наибольшей частоте присваивается ранг 1, меньшей частоте назначается больший ранг, одинаковые частоты имеют одинаковые ранги. Основные сводные характеристики частотных свойств текста: количество слов в тексте, количество уникальных слов в тексте, количество уникальных рангов частот слов. Эти характеристики легко найти из таблицы частот слов. Сумма всех частот во втором столбце таблицы частот равна количеству слов в тексте, число строк в таблице частот равно количеству уникальных слов в тексте, а количество разных частот во втором столбце таблицы частот равно количеству уникальных рангов. В табл. 1 приведен фрагмент таблицы частот слов для статьи с кодом «10281» из рассматриваемого корпуса научных текстов. Количество слов в этой статье – 1113, из них уникальных слов – 455, а уникальных рангов (различных частот) всего 20. Хорошо известно, что меньшие частоты в таблице частот обычно встречаются чаще. Например, из таблицы частот статьи «10281» легко найти, что частоты 56, 52, 45 встречаются по одному разу, частота 15 встречается трижды, а частоты 3, 2 и 1 встречаются 31, 73 и 293 раза соответственно. Очевидно, что частоты уникальных частот (уникальных рангов) в таблице частот слов содержат важную информацию о частотных свойствах текста. Для удобства использования этой информации можно составить таблицу частот 2-го порядка, которую мы понимаем как таблицу частот для уникальных частот из исходной таблицы частот слов. Таблица частот 2-го порядка имеет тот же формат, что и исходная таблица частот слов. В первом столбце таблицы частот 2-го порядка представлены все уникальные частоты из исходной таблицы частот, а во втором столбце – частоты уникальных частот в порядке убывания. «Слова» в таблице частот 2-го порядка представляют собой уникальные частоты из исходной таблицы частот и поэтому кодируют целые области лексики исходного текста с одинаковой частотностью. «Частоты» в таблице частот 2-го порядка показывают объемы различных областей лексики с одинаковой частотностью.

В табл. 1 приведен пример таблицы частот 2-го порядка для статьи «10281». Легко заметить, что в ней всего 20 строк, так как имеется 20 уникальных частот в исходной таблице частот статьи «10281». Сумма всех частот в таблице частот 2-го порядка статьи «10281» равна 455, что совпадает с количеством уникальных слов в исходной таблице частот. Из определения таблицы частот 2-го порядка в общем случае следуют два равенства: количество уникальных слов для исходной таблицы равно количеству слов

для таблицы частот второго порядка; количество уникальных частот для исходной таблицы равно количеству уникальных слов для таблицы частот 2-го порядка. Рассматривая таблицу частот 2-го порядка как исходную, можно для нее построить таблицу частот уникальных частот, которую будем называть таблицей частот 3-го порядка. Продолжая процесс, можно определить таблицу частот (n+1)-го порядка как таблицу частот уникальных частот таблицы частот n-го порядка.

Таблица 1

Примеры таблиц частот высших порядков

Фрагмент таблицы частот слов для статьи «10281»			Таблица частот 2-го порядка для статьи «10281»		
		word frequency			word frequency
1	воображение	56	1	1	293
2	в	52	2	2	73
3	и	45	3	3	31
4	с	25	4	4	16
5	что	22	5	6	9
6	деятельность	19	6	5	7
7	образ	16	7	7	6
8	как	15	8	10	3
9	человек	15	9	11	3
10	творческий	15	10	15	3
11	не	13	11	13	2
12	он	13	12	8	1
.....	.....	.....	13	9	1
449	смешение	1	14	16	1
450	психолого	1	15	19	1
451	развивающийся	1	16	22	1
452	игнатьев	1	17	25	1
453	разнообразие	1	18	45	1
454	понимать	1	19	52	1
455	между	1	20	56	1
Количество слов: 1113			Количество слов: 455		
Количество уникальных слов: 455			Количество уникальных слов: 20		
Количество уникальных рангов: 20			Количество уникальных рангов: 10		

Окончание табл. 1

Таблица частот 3-го порядка для статьи «10281»	Таблица частот 4-го порядка для статьи «10281»
word frequency 1 1 9 2 3 3 3 2 1 4 6 1 5 7 1 6 9 1 7 16 1 8 31 1 9 73 1 10 293 1 Количество слов: 20 Количество уникальных слов: 10 Количество уникальных рангов: 3	word frequency 1 1 8 2 3 1 3 9 1 --- Таблица частот 5-го порядка для статьи «10281» --- word frequency 1 1 2 2 8 1 --- Таблица частот 6-го порядка для статьи «10281» --- word frequency 1 1 1 2 2 1

В табл. 1 показаны таблицы частот высших порядков от 2-го до 6-го порядка включительно для статьи «10281». Необходимо отметить, что количество строк и количество уникальных рангов в таблицах частот высших порядков быстро убывает в зависимости от порядка таблицы. Начиная с некоторого порядка n, таблицы частот вырождаются. Вырожденными таблицами будем считать такие, для которых количество уникальных рангов меньше 3. В табл. 1 для статьи «10281» вырожденными являются все таблицы частот, начиная с 4-го порядка. При этом все таблицы частот, начиная с 7-го порядка, состоят из одной строки, а таблицы, начиная с 9-го порядка, все одинаковые: слово «1» встречается с частотой 1. Вырожденные таблицы частот могут оказаться мало полезными при решении прикладных задач об автоматической классификации научных текстов по их частотным характеристикам.

Построение обобщенных моделей Ципфа высших порядков

В данной работе рассматривался корпус научных текстов из журнала «Молодой ученый», который содержит 12 120 научных статей на русском языке из 22 секций журнала [Агеев, Кашпарова, Синицын 2019]. В исходных таблицах частот (таблицах частот первого порядка) количество уникальных рангов варьируется от 15 до 45. В данном исследовании для всех 12120 статей были построены невырожденные таблицы частот 2-го порядка. Количество уникальных рангов, обнаруженных для таблиц частот 2-го порядка, было от 7 до 19 включительно. При этом в большинстве случаев таблицы частот 2-го порядка имели 10, 11 или 12 уникальных рангов. Для всех статей из рассматриваемого корпуса текстов были построены также таблицы частот высших порядков до 8-го порядка включительно. Из таблиц частот 3-го порядка 366 оказались вырожденными, они имели 2 уникальных ранга каждая. Остальные таблицы частот 3-го порядка имели от 3 до 6 уникальных рангов (УР). 3 УР было для 5426 статей, 4 УР было для 5588 статей, 5 УР – для 729 статей и 6 УР всего для 11 статей. Больше половины таблиц частот 4-го порядка оказались вырожденными (6952 статьи с 2 УР). 3 УР имели 5144 статьи, а 4 УР имели всего 24 статьи. Все таблицы частот 5-го порядка вырожденные. Из них 9153 таблицы – с 2 УР, а 2967 таблиц – с 1 УР. В дальнейшем все таблицы частот высших порядков вырождаются к одной таблице: слово «1» встречается 1 раз.

В целях построения и исследования невырожденных таблиц частот более высокого порядка таблицы частот отдельных статей были объединены в 22 большие таблицы частот слов по секциям мультидисциплинарного журнала. Сводная информация о количестве слов и количестве уникальных рангов для таблиц частот по секциям журнала до 6-го порядка включительно представлена в табл. 2. Из табл. 2 видно, что все таблицы частот слов по секциям до 4-го порядка невырожденные, а 6-го порядка все вырожденные с 2 УР. Таблицы частот 5-го порядка для 13 секций журнала невырожденные, а для 9 секций вырожденные.

В 2020 г. в работе [Гламаздин, Гордин, Синицын, Кашпарова 2020] для всех статей из рассматриваемого корпуса текстов были построены обобщенные финитные модели Ципфа: степенная модель, гиперболическая модель, модель Ципфа с квадратичной поправкой, модель Ципфа со свободным членом. На их основе удалось более подробно описать частотные свойства научных текстов и выделить предикторы для автоматической классификации статей по секциям журнала. В данной работе для всех найденных

невырожденных таблиц частот высших порядков построены модели Ципфа и обобщенные модели Ципфа (нефинитные) тех же типов, что и в работе [Гламаздин, Гордин, Синицын, Кашпарова 2020], а также проанализированы характеристики построенных моделей, и выделены новые предикторы для решения задач классификации.

Таблица 2

Количество слов (n_word)
и количество уникальных рангов (ur1, ur2,..., ur6)
для таблиц частот до 6-го порядка

Section	n_word	ur1	ur2	ur3	ur4	ur5	ur6
1 Педагогика	3 305 114	1268	87	16	7	3	2
2 Психология	766 482	623	67	15	5	2	2
3 Технические науки	1 350 115	823	82	15	5	3	2
4 Экономика и управление	4 346 015	1487	92	19	5	4	2
5 Медицина	869 991	645	75	15	5	3	2
6 Государство и право	1 989 788	1013	81	15	5	3	2
7 Информатика	397 264	452	60	12	5	3	2
8 История	980 221	632	86	15	4	3	2
9 Математика	143 722	259	49	10	5	3	2
10 Социология	340 800	396	56	13	4	3	2
11 Биология	130 390	220	48	11	4	2	2
12 География	126 265	236	46	10	4	2	2
13 Искусствоведение	265 883	318	56	10	5	2	2
14 Культурология	236 186	298	55	10	5	2	2
15 Политология	293 582	369	57	12	5	2	2
16 Сельское хозяйство	158 160	266	49	10	5	3	2
17 Физика	138 533	256	45	11	3	2	2
18 Физическая культура	243 401	348	53	11	5	3	2
19 Филология	1 514 455	811	85	15	6	3	2
20 Философия	501 878	490	64	13	4	3	2
21 Химия	98 473	202	47	10	4	2	2
22 Экология	215 599	303	55	12	5	2	2

В табл. 3 представлена общая информация о моделях и о количестве полученных новых предикторов. Обобщенными моделями Ципфа высших порядков мы называем обобщенные модели Ципфа для таблиц частот высших порядков.

Таблица 3

Обобщенные модели Ципфа высших порядков

Тип модели n-го порядка (n > 1)	Формула модели n-го порядка (n > 1)	Построено моделей для каждой статьи	Получено новых предикторов
Классическая модель Ципфа	W = C / R	4	12
Степенная модель	W = B / R^D	1	3
Гиперболическая модель	W = B / (R+D)	1	3
Модель Ципфа с квадратичной поправкой	W = B / R+D / R^2	3	12
Модель Ципфа со свободным членом	W = B / R+D	3	12

В табл. 3 использованы обозначения: R – ранг слова в таблице частот, W – относительная частота слова, C – постоянный коэффициент Ципфа, B – первый параметр обобщенной модели Ципфа (аналогичный константе Ципфа), D – второй параметр для двухпараметрических моделей. Важно отметить, что формулы моделей в табл. 3 не различаются для разных порядков моделей, но значения параметров, которые получаются по таблицам частот n-го порядка, зависят от порядка модели. Нелинейные степенные и гиперболические модели 2-го порядка были построены для всех 12 120 статей. Каждая такая модель дала три предиктора для решения задач классификации. Такими предикторами являются параметры модели B и D, а также характеристика качества модели – сумма квадратов остатков (RSS). Отметим, что не удалось построить степенные и гиперболические модели 3-го порядка для некоторых статей, так как они имеют вырожденные таблицы частот 3-го порядка. Кроме того, нелинейные модели оценивались итеративно, и в сложных случаях было превышено стандартное количество итераций 50.

Таблица 4

Характеристики моделей высших порядков
для статьи «10281»

Характеристики моделей Ципфа высших порядков

	n_word	n_uniq_word	n_uniq_rank	coeff	R2	RSS
1	1113	455	20	0.073	0.8271	0.00186
2	455	20	10	0.4823	0.8280	0.07694
3	20	10	3	0.4075	0.9756	0.00598

Связь с обозначениями в табл. 3: C=coeff.

Характеристики степенных моделей n-го порядка

	n_word	n_uniq_word	n_uniq_rank	Power_b1	Power_b2	Power_RSS
1	1113	455	20	0.060	0.70	0.00037
2	455	20	10	0.644	2.04	0.00005

Связь с обозначениями в табл. 3: B=Power_b1, D=Power_b2.

Характеристики гиперболических моделей n-го порядка

	n_word	n_uniq_word	n_uniq_rank	Hyper_b1	Hyper_b2	Hyper_RSS
1	1113	455	20	0.171	2.225	0.00025
2	455	20	10	0.156	–0.758	0.00317

Связь с обозначениями в табл. 3: B=Hyper_b1, D=Hyper_b2.

Характеристики моделей Ципфа с квадратичной поправкой

	n_word	n_uniq_word	n_uniq_rank	Zipf2_b1	Zipf2_b2	Zipf2_R2	Zipf2_RSS
1	1113	455	20	0.139	–0.089	0.976	0.00026
2	455	20	10	–0.014	0.658	1.000	0.00003
3	20	10	3	0.257	0.188	0.996	0.00096

Связь с обозначениями в табл. 3: B=Zipf2_b1, D=Zipf2_b2.

Характеристики моделей Ципфа со свободным членом

	n_word	n_uniq_word	n_uniq_rank	Zipf0_b0	Zipf0_b1	Zipf0_R2	Zipf0_RSS
1	1113	455	20	0.001	0.069	0.855	0.00124
2	455	20	10	–0.061	0.619	0.917	0.03280
3	20	10	3	–0.023	0.452	0.977	0.00334

Связь с обозначениями в табл. 3: B=Zipf0_b1, D=Zipf0_b0.

Линейные обобщенные модели Ципфа с квадратичной поправкой и со свободным членом были построены до 4-го порядка включительно для всех статей, включая случаи с вырожденными таблицами частот высших порядков. Каждая такая модель дала четыре предиктора для решения задач классификации. Предикторами являются параметры модели B и D, а также коэффициент детерминации модели и RSS.

Классические однопараметрические модели Ципфа являются линейными и были построены до 5-го порядка для всех статей,

включая случаи с вырожденными таблицами частот высших порядков. Каждая модель Ципфа высшего порядка дала три предиктора, которыми являются коэффициент Ципфа, коэффициент детерминации модели и RSS.

Из табл. 3 видно, что благодаря построенным обобщенным моделям Ципфа высших порядков получено 42 новых предиктора для решения задач автоматической классификации научных текстов. Имеются также общие частотные свойства текстов: число слов, число уникальных слов и число уникальных рангов для таблиц частот до 5-го порядка. Эти общие частотные свойства можно также рассматривать как предикторы для решения задач классификации.

В табл. 4 приведены основные характеристики невырожденных обобщенных моделей Ципфа высших порядков для статьи «10281».

Обсуждение

Анализ характеристик всех построенных моделей показывает, что примеры, приведенные в табл. 4, типичны для рассматриваемого корпуса текстов. Обнаружено, что качество однопараметрических моделей Ципфа высших порядков для всех статей не хуже, чем качество классической модели Ципфа, так как коэффициент детерминации стабилен вплоть до «предвырожденной» таблицы частот высшего порядка. Имеет место эффект резкого увеличения коэффициента детерминации модели для порядка n, предшествующего вырождению таблицы частот.

Нелинейные двухпараметрические степенные и гиперболические модели высших порядков значительно превосходят по качеству однопараметрические модели Ципфа того же порядка. Об этом свидетельствует тот факт, что для нелинейных моделей RSS в десятки и даже иногда в сотни раз меньше, чем для соответствующих моделей Ципфа. Графики зависимости относительной частоты от ранга для статьи «10281», полученные на основе модели Ципфа и степенной модели, представлены на рис. 1. Аналогичное сравнение для моделей 2-го порядка продемонстрировано на рис. 2.

Слева на рис. 1 хорошо заметен эффект, при котором для средних рангов модель Ципфа дает ощутимо заниженный прогноз для относительной частоты. Это характерно для всех научных статей из нашего корпуса текстов независимо от их тематики. Если по закону Ципфа для статьи «10281» сгенерировать случайную таблицу частот, соответствующую тому же закону Ципфа, и такую, что модель Ципфа для нее имеет коэффициент детерминации значительно больший, например, 0.95 вместо реального 0.827 для статьи «10281», то в этом

случае эффекта занижения прогноза относительной частоты при средних рангах не будет. Таким образом, классическая однопараметрическая модель Ципфа хорошо, но недостаточно точно описывает зависимость относительной частоты от ранга. Этого недостатка лишена степенная модель, как можно видеть на рис. 1 справа. Прогноз степенной модели близок к идеальному. При этом показатель степени D = 0.7, что статистически значимо отличается от 1.

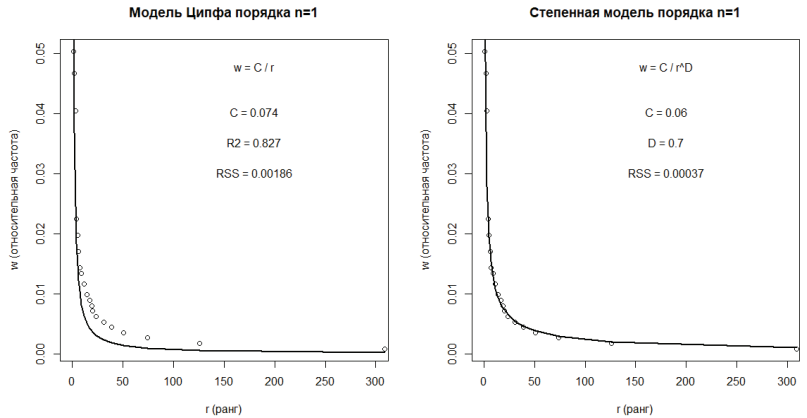


Рис. 1. Сравнение модели Ципфа и степенной модели для статьи «10281»

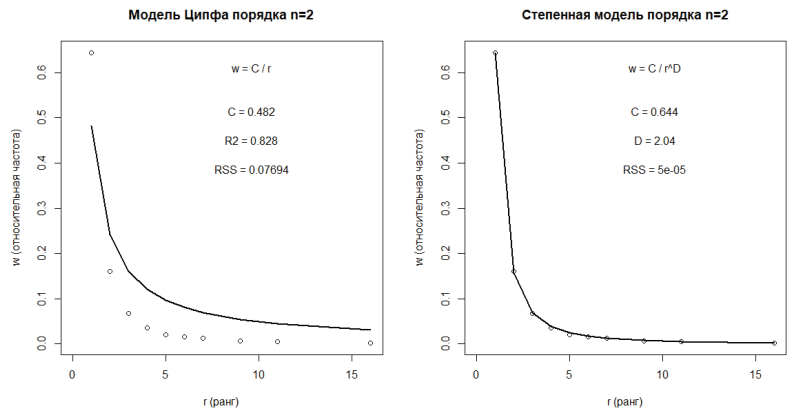


Рис. 2. Сравнение моделей 2-го порядка для статьи «10281»

Слева на рис. 2 можно видеть типичный эффект, при котором модель Ципфа 2-го порядка дает завышенный прогноз относительной частоты для всех рангов кроме случая R = 1, при котором прогноз сильно занижен. Степенная модель 2-го порядка почти идеально прогнозирует относительную частоту, как это видно на рис. 2 справа. Степенная модель в данном случае говорит о том, что относительная частота убывает обратно пропорционально (примерно) квадрату ранга, так как параметр D = 2.04 для статьи «10281».

Обобщенные модели Ципфа высших порядков с квадратичной поправкой составляют достойную конкуренцию нелинейным моделям, не уступая им по показателю RSS. Необходимо отметить, однако, что против обобщенных моделей Ципфа с квадратичной поправкой в некоторых случаях могут быть содержательные возражения. Иногда оценка параметра B для таких моделей получается отрицательной, а это означает, что для достаточно больших значений ранга R прогноз для относительной частоты W будет получаться отрицательным, затрудняя интерпретацию результатов моделирования.

Обобщенные модели Ципфа высших порядков со свободным членом превосходят немного классические однопараметрические модели Ципфа по коэффициенту детерминации, но значительно уступают другим рассмотренным двухпараметрическим моделям по показателю RSS. Кроме того, прогноз относительной частоты при больших значениях ранга может быть иногда отрицательным, так как оценка параметра D в некоторых случаях меньше нуля.

Графики зависимости относительной частоты от ранга для таблицы частот 2-го порядка статьи «10281», полученные на основе четырех обобщенных моделей Ципфа, представлены на рис. 3.

Из рис. 3 видно, что обобщенная модель Ципфа 2-го порядка с квадратичной поправкой (Zipf2) имеет коэффициент детерминации, равный 1, и самую маленькую величину RSS среди всех двухпараметрических моделей. Однако формально безупречная модель, на наш взгляд, не может быть признана лучшей из четырех представленных моделей, так как она содержательно неадекватна. Параметр, подобный константе Ципфа, отрицательный – C = –0.014. По этой причине при рангах R > 47 прогноз относительной частоты отрицательный, а это трудно интерпретировать. Обобщенная модель Ципфа со свободным членом (Zipf0) 2-го порядка имеет аналогичные проблемы с возможным отрицательным прогнозом относительной частоты. Гиперболическая модель 2-го порядка имеет довольно большое значение RSS, то есть не очень точна в своих прогнозах. Учитывая все вышесказанное, наилучшей из четырех представленных на рис. 3 моделей приходится признать степенную модель 2-го порядка.

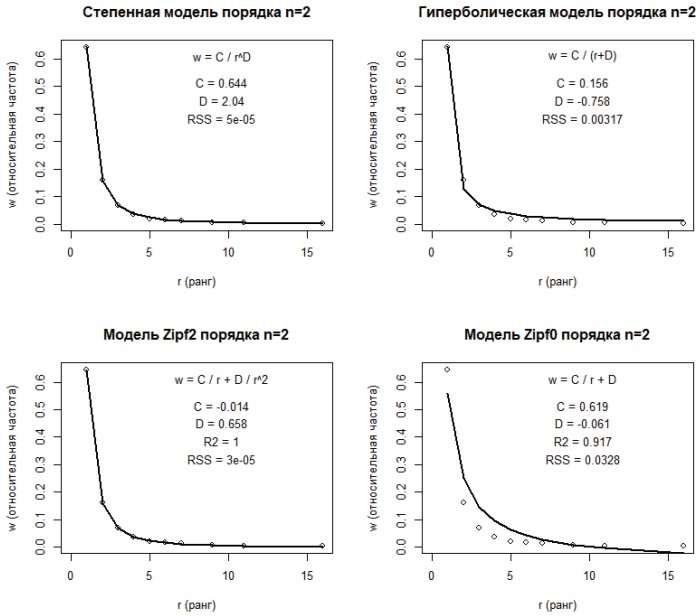


Рис. 3. Сравнение обобщенных моделей 2-го порядка для статьи «10281»

Заключение

Итак, на рассматриваемом корпусе текстов для всех невырожденных таблиц частот высших порядков выполняется закон Ципфа и обобщенные законы Ципфа о зависимости относительной частоты от ранга, которые выражаются формулами из табл. 3. В данной работе построены обобщенные модели Ципфа высших порядков, которые имеют хорошие показатели качества и могут быть полезны для решения задач автоматического соотнесения научных статей рубрикам мультидисциплинарного журнала без использования информации о семантике текстов.

Литература

Агеев, Кашпарова, Синицын 2019 – *Агеев А.В., Кашпарова В.С., Синицын В.Ю.* Применение закона Ципфа для частотного анализа текстов на естественном языке и их автоматической классификации // Международный гуманитар-

ный научный форум «Гуманитарные чтения РГГУ–2019». М.: Янус-К, 2019. С. 42–49.

Гламаздин, Гордин, Синицын, Кашпарова 2020 – *Гламаздин В.С., Гордин Р.Р., Синицын В.Ю., Кашпарова В.С.* Обобщенные финитные модели Ципфа для анализа частотных свойств научных текстов // Международный гуманитарный научный форум «Гуманитарные чтения РГГУ–2020». М.: Янус-К, 2020. С. 5–11.

Гламаздин, Синицын 2021 – *Гламаздин В.С., Синицын В.Ю.* Методы машинного обучения для автоматической классификации научных текстов различной тематики по их частотным характеристикам // Международный гуманитарный научный форум «Гуманитарные чтения РГГУ–2021». М.: Янус-К, 2021. С. 39–44.

Гордин, Синицын 2021 – *Гордин Р.Р., Синицын В.Ю.* Построение ансамблей обобщенных финитных моделей Ципфа для автоматической классификации научных текстов // Международный гуманитарный научный форум «Гуманитарные чтения РГГУ–2021». М.: Янус-К, 2021. С. 45–50.

Мандельброт 1973 – *Мандельброт Б.* Теория информации и психолингвистическая теория частот слов // Математические методы в социальных науках. М.: Прогресс, 1973. С. 316–337.

Маслов, Маслова 2006 – *Маслов В.П., Маслова Т.В.* О законе Ципфа и ранговых распределениях в лингвистике и семиотике // Математические заметки. 2006. Т. 80. Вып. 5. С. 718–732.

Шрейдер 1967 – *Шрейдер Ю.А.* О возможности теоретического вывода статистических закономерностей текста (к обоснованию закона Ципфа) // Проблемы передачи информации. 1967. Т. 3. Вып. 1. С. 57–63.

Zipf 1949 – *Zipf G.K.* Human behavior and the principle of least effort. Cambridge: Addison-Wesley Press, 1949.

References

Ageev, A.V., Kashparova, V.S. and Sinitsyn, V.Yu. (2019), “Application of Zipf’s law for frequency analysis of natural language texts and their automatic classification”, *International Humanitarian Scientific Forum “Humanitarian Conference of RSUH 2019”*, Yanus-K, Moscow, Russia, pp. 42–49.

Glamazdin, V.S., Gordin, R.R., Sinitsyn, V.Yu. and Kashparova, V.S. (2020), “Generalized finite Zipf models for the analysis of frequency properties of scientific texts”, *International Humanitarian Scientific Forum “Humanitarian Conference of RSUH 2020”*, Yanus-K, Moscow, Russia, pp. 5–11.

Glamazdin, V.S. and Sinitsyn, V.Yu. (2021), “Machine learning methods for automatic classification of scientific texts on various subjects by their frequency characteristics”, *International Humanitarian Scientific Forum “Humanitarian Conference of RSUH 2021”*, Yanus-K, Moscow, Russia, pp. 39–44.

Gordin, R.R. and Sinitsyn, V.Yu. (2021), "Construction of ensembles of generalized finite Zipf models for automatic classification of scientific texts", *International Humanitarian Scientific Forum "Humanitarian Conference of RSUH 2021"*, Yanus-K, Moscow, Russia, pp. 45–50.

Mandelbrot, B. (1973), "Information theory and psycholinguistic theory of word frequencies", *Matematicheskie metody v sotsial'nykh naukakh* [Mathematical methods in social sciences], Progress, Moscow, Russia, pp. 316–337.

Maslov, V.P. and Maslova, T.V. (2006) "On Zipf's law and rank distributions in linguistics and semiotics", *Matematicheskie zametki*, vol. 80, no. 5, pp. 718–732.

Schrader, Yu.A. (1967) "On the possibility of theoretical deduction of statistical regularities of the text (on Zipf's law substantiation)", *Problemy peredachi informatsii,* vol. 3, no. 1, pp. 57–63.

Zipf, G.K. (1949), *Human behavior and the principle of least effort*, Addison-Wesley Press, Cambridge, UK.

Информация об авторах

Вячеслав Ю. Синицын, кандидат физико-математических наук, доцент, Российский государственный гуманитарный университет, Москва, Россия; 125047, Россия, Москва, Миусская пл., д. 6; fpmrggu@yandex.ru

Валентина С. Кашпарова, кандидат филологических наук, доцент, Московский педагогический государственный университет, Москва, Россия; 119991, Россия, Москва, ул. Малая Пироговская, д. 1, стр. 1; vs.kashparova@mpgu.su

Information about the authors

Vyacheslav Yu. Sinitsyn, Cand. of Sci. (Physics and Mathematics), associate professor, Russian State University for the Humanities, Moscow, Russia; bld. 6, Miusskaya Sq., Moscow, Russia, 125047; fpmrggu@yandex.ru

Valentina S. Kashparova, Cand. of Sci. (Philology), associate professor, Moscow Pedagogical State University, Moscow, Russia; bld. 1/1, Malaya Pirogovskaya Str., Moscow, Russia, 119991; vs.kashparova@mpgu.su

Дизайн обложки
Е.В. Амосова

Корректор
А.А. Леонтьева

Компьютерная верстка
Н.В. Москвина

Подписано в печать 26.12.2022.
Формат 60×90 1/16.
Уч.-изд. л. 5,7. Усл. печ. л. 5,8.
Тираж 1050 экз. Заказ № 1644

Издательский центр
Российского государственного
гуманитарного университета
125047, Москва, Миусская пл., 6
www.rsuh.ru