Научно-технический журнал

H&ES RESEARCH

Scientific and Technical Journal

НАУКОЕМКИЕ ТЕХНОЛОГИИ В КОСМИЧЕСКИХ ИССЛЕДОВАНИЯХ ЗЕМЛИ

HIGH TECHNOLOGIES IN EARTH SPACE RESEARCH

Журнал **H&ES Research** издается с 2009 года, освещает достижения и проблемы российских инфокоммуникаций, внедрение последних достижений отрасли в автоматизированных системах управления, развитие технологий в информационной безопасности, исследования космоса, развитие спутникового телевидения и навигации, исследование Арктики. Особое место в издании уделено результатам научных исследований молодых ученых в области создания новых средств и технологий космических исследований Земли.

Журнал H&ES Research входит в перечень изданий, публикации в которых учитываются Высшей аттестационной комиссиейРоссии (ВАК РФ), в систему российского индекса научного цитирования (РИНЦ), а также включен в Международный классификатор периодических изданий.

Тематика публикуемых статей в соответствии с перечнем группспециальностей научных работников по Номенклатуре специальностей:

- 2.2.15 Системы, сети и устройства телекоммуникаций (техн. науки)
- 2.3.1 Системный анализ, управление и обработка информации (техн. науки)
- 2.3.5 Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей (техн. науки)
- 2.3.6 Методы и системы защиты информации, информационная безопасность (техн. науки)
- 2.5.13 Проектирование, конструкция и производство летательныъх аппаратов (техн. науки)
- 2.5.16 Динамика, баллистика, управление движением летательных аппаратов (техн. науки)

ИНДЕКСИРОВАНИЕ ЖУРНАЛА H&ES RESEARCH

H&ES Research is published since 2009. The journal covers achievements and problems of the Russian infocommunication, introduction of the last achievements of branch in automated control systems, development of technologies in information security, space researches, development of satellite television and navigation, research of the Arctic. The special place in the edition is given to results of scientific researches of young scientists in the field of creation of new means and technologies of space researches of Earth.

The journal H&ES Research is included in the list of scientific publications, recommended Higher Attestation Commission Russian Ministry of Education for the publication of scientific works, which reflect the basic scientific content of candidate and doctoral theses. IF of the Russian Science Citation Index.

Subject of published articles according to the list of branches of science and groups of scientific specialties in accordance with the specialties:

- 2.2.15 Telecommunication systems, networks and devices
- 2.3.1 System analysis, management and information processing
- 2.3.5 Mathematical and software support for computing systems, complexes and computer networks
- 2.3.6 Methods and systems of information security
- 2.5.13 Design, construction and production of aircraft
- 2.5.16 Dynamics, ballistics, aircraft motion control

JOURNAL H&ES RESEARCH INDEXING

• NEICON • CyberLenika (Open Science) • Google Scholar • OCLC WorldCat • Ulrich's Periodicals Directory • • Bielefeld Academic Search Engine (BASE) • eLIBRARY.RU • Registry of Open Access Repositories (ROAR)

Все номера журнала находятся в свободном доступе на сайте журнала **www.hes.ru** и библиотеке **elibrary.ru.**

Всем авторам, желающим разместить научную статью в журнале, необходимо оформить ее согласно требованиям и направить материалы на электронную почту: **HT-ESResearch@yandex.ru.**
С требованиями можно ознакомиться на сайте: **www.H-ES.ru.**

Язык публикаций: русский, английский.
Периодичность выхода – 6 номеров в год.
Свидетельство о регистрации СМИ ПИ № ФС 77-86203 от 27.10.2023
Территория распространения: Российская Федерация, зарубежные страны

Тираж 1000 экз. Цена 1000 руб.
Плата с аспирантов за публикацию рукописи не взимается.

© ООО “ИД Медиа Паблишер”, 2024

All issues of the journal are in a free access on a site of the journal **www.hes.ru** and **elibrary.ru.**

All authors wishing to post a scientific article in the journal, you must register it according to the requirements and send the materials to your email: **HT-ESResearch@yandex.ru.**
The requirements are available on the website: **www.H-ES.ru.**

Language of publications: Russian, English.
Periodicity – 6 issues per year.
Media Registration Cetrificate PI No. FS 77-86203 Date of issue: 27.10.2023
Distribution Territory: Russian Federation, foreign countries

Circulation of 1000 copies. Price of 1000 Rur.
Postgraduate students for publication of the manuscript will not be charged

© “Media Publisher”, LLC, 2024

ISSN 2409-5419 (Print) ISSN 2412-1363 (Online) WWW.H-ES.RU Дата выхода: 22 февраля 2024

Т.16 1 2024

Учредитель:
ООО "ИД Медиа Паблишер"

Издатель:
ДЫМКОВА С.С.

Главный редактор:
ЛЕГКОВ К.Е.

Редакционная коллегия:
БОБРОВСКИЙ В.И., д.т.н., доцент;
БОРИСОВ В.В., д.т.н., профессор, Действительный член академии военных наук РФ;
БУДКО П.А., д.т.н., профессор;
БУДНИКОВ С.А., д.т.н., доцент, Действительный член Академии информатизации образования;
ВЕРХОВА Г.В., д.т.н., профессор;
ГОНЧАРЕВСКИЙ В.С., д.т.н., профессор, заслуженный деятель науки и техники РФ;
КОМАШИНСКИЙ В.И., д.т.н., профессор;
КИРПАНЕВ А.В., д.т.н., доцент;
КУРНОСОВ В.И., д.т.н., профессор, академик Международной академии информатизации, Действительный член Российской академии естественных наук;
МОРОЗОВ А.В., д.т.н., профессор, Действительный член Академии военных наук РФ;
МОШАК Н.Н., д.т.н., доцент;
ПАВЛОВ А.Н., д.т.н., профессор;
ПРОРОК В.Я., д.т.н., профессор;
СЕМЕНОВ С.С., д.т.н., доцент;
СИНИЦЫН Е.А., д.т.н., профессор;
ШАТРАКОВ Ю.Г., д.т.н., профессор, заслуженный деятель науки РФ.

Адрес издателя:
111024, Россия, Москва,
ул. Авиамоторная, д. 8, корп. 1, офис 323.

Адрес редакции:
194044, Россия, Санкт-Петербург,
Лесной Проспект, 34-36, к. 1,
Тел.: +7(911) 194-12-42.

Адрес типографии:
Россия, Москва, ул. Складочная, д. 3, кор. 6.

Мнения авторов не всегда совпадают с точкой зрения редакции.
За содержание рекламных материалов редакция ответственности не несет.
Материалы, опубликованные в журнале – собственность ООО "ИД Медиа Паблишер".
Перепечатка, цитирование, дублирование на сайтах допускаются только с разрешения издателя.

СОДЕРЖАНИЕ

РАДИОТЕХНИКА И СВЯЗЬ

Чипсанова Е.В., Елагин В.С.
Методы распределения ресурсов концепции мобильных граничных вычислений **4**

Сидоренко А.А.
Построение комбинированного кода на основе турбокода с дополнительным кодированием бит **14**

ИНФОРМАТИКА, ВЫЧИСЛИТЕЛЬНАЯ ТЕХНИКА И УПРАВЛЕНИЕ

Маркевич Д.В., Хомоненко А.Д., Ермаков С.Г.
От Foxpro к PostgreSQL: оптимизация, эффективное управление данными и генерация отчетов **21**

Гусеница Я.Н., Донченко А.А., Ляскин А.С.
Концепция поддержки принятия решений задач сопровождения жизненного цикла сложных технических систем **31**

Галеев Д.Т., Титов Д.В., Волков Д.А.
Нейросетевая модель в информационной вопросно-ответной системе **39**

Хомоненко А.Д., Никитин А.Б., Кириенко А.Б.
Доступ к геоинформационным системам на основе кватернионного шифрования и нейронных сетей **47**

Переволоцкий В.С., Кравчук Д.А.
Использование автоматизированных систем для создания онтологических баз знаний **54**

CONTENTS

RF TECHNOLOGY AND COMMUNICATION

Chipsanova E.V., Elagin V.S.
Resource allocation methods of Mobile Edge Computing concept 4

Sidorenko A.A.
Building a combined code based on turbo code with additional bit encoding 14

INFORMATICS, COMPUTER ENGINEERING AND CONTROL

Markevich D.V., Khomonenko A.D., Ermakov S.G.
Foxpro to PostgreSQL: optimization, effective data management, and report generation 21

Gusenitza Ya.N., Donchenko A.A., Lyaskin A.S.
Decision support concept for lifecycle support problems of complex technical systems 31

Galeev T.D., Titov D.V., Volkov D.A.
Neural network model in the information question-answering system 39

Khomonenko A.D., Nikitin A.B., Kirienko A.B.
Access to geographic information systems based on quaternion encryption and neural networks 47

Perevolotsky V.S., Kravchuk D.A.
Using automated systems to create ontological knowledge bases 54

Founder:
"Media Publisher", LLC

Publisher:
DYMKOVA S.S.

Editor in chief:
LEGKOV K.E.

Editorial board:
BOBROWSKY V.I., PhD, Docent;
BORISOV V.V., PhD, Full Professor;
BUDKO P.A., PhD, Full Professor;
BUDNIKOV S.A., PhD, Docent, Actual Member of the Academy of Education Informatization;
VERHOVA G.V., PhD, Full Professor;
GONCHAREVSKY V.S., PhD, Full Professor, Honored Worker of Science and Technology of the Russian Federation;
KOMASHINSKIY V.I., PhD, Full Professor;
KIRPANEEV A.V., PhD, Docent;
KURNOSOV V.I., PhD, Full Professor, Academician of theInternational Academy of Informatization, law and order, Member of the Academy of Natural Sciences;
MOROZOV A.V., PhD, Full Professor, Actual Member of the Academy of Military Sciences;
MOSHAK N.N.,PhD, Docent;
PAVLOV A.N., PhD, Full Professor;
PROROK V.Y., PhD, Full Professor;
SEMENOV S.S., PhD, Docent;
SINICYN E.A., PhD, Full Professor;
SHATRAKOV Y.G., PhD, Full Professor;
Honored Worker of Science of the Russian Federation.

Address of publisher:
111024, Russia, Moscow,
st. Aviamotornaya, 8, bild. 1, office 323

Address of edition:
194044, Russia, St. Petersburg,
Lesnoy av., 34-36, h.1,
Phone: +7 (911) 194-12-42.

Address of printing house:
Russia, Moscow, st. Skladochnaya, 3, h. 6

The opinions of the authors don't always coincide with the point of view of the publisher. For the content of ads, the editorial Board is not responsible. All artocles and illustrations are copyryght. All rights reserved.No reproduction is permitted in whole or part without the express consent of Media Publisher Joint-Stock company.

doi: 10.36724/2409-5419-2024-16-1-4-13

МЕТОДЫ РАСПРЕДЕЛЕНИЯ РЕСУРСОВ КОНЦЕПЦИИ МОБИЛЬНЫХ ГРАНИЧНЫХ ВЫЧИСЛЕНИЙ

ЧИПСАНОВА
Елена Валерьевна[1]

ЕЛАГИН
Василий Сергеевич[2]

АННОТАЦИЯ

Введение: в настоящее время существует большое количество приложений, которые требуют больших вычислений, а также им необходимо обладать минимальной задержкой. Эти приложения могут превосходить вычислительные возможности мобильных устройств. Многообещающей технологией для поддержки таких приложений являются мобильные граничные вычисления (MEC). В данной концепции граничные серверы размещаются рядом с мобильными устройствами, что позволяет обеспечить разгрузку вычислений. MEC может существенно увеличить вычислительные мощности мобильных устройств. Однако разгрузка вычислений требует передачи данных с мобильных устройств на пограничные серверы. Это может приводить к дополнительной задержке передачи данных и большему потреблению энергии. В связи с этим встаёт вопрос эффективного распределения ресурсов для концепции мобильных граничных вычислений в телекоммуникационных сетях. **Цель работы:** главной задачей является ознакомление с актуальными методами, которые позволяют улучшить распределение ресурсов для концепции MEC. Также, среди рассмотренных методов, необходимо выявить наиболее актуальный для современных телекоммуникационных сетей. Был выбран подход анализа, так как он является наиболее приемлемым для поставленных задач. Результат: рассмотрены методы для распределения ресурсов граничных мобильных вычислений, а также выявлен наиболее актуальный метод из представленных.

Сведения об авторах:

[1] Аспирант, Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, Санкт-Петербург, Россия, lenchip@mail.ru

[2] к.т.н., Заведующий кафедрой ИКС, Санкт-Петербургский государственный университет телекоммуникаций им. проф. М. А. Бонч-Бруевича, Санкт-Петербург, Россия, v.elagin@sut.ru

КЛЮЧЕВЫЕ СЛОВА: *MEC, распределение ресурсов, DRL, JCC, SCA, игра Штакельберга.*

Для цитирования: *Чипсанова Е.В., Елагин В.С.* Методы распределения ресурсов концепции мобильных граничных вычислений // Наукоемкие технологии в космических исследованиях Земли. 2023. Т. 16. № 1. С. 4-13. doi: 10.36724/2409-5419-2024-16-1-4-13

Введение

Стремительное развитие индустрии мобильной информации привело к появлению различных мобильных приложений во многих областях [1], например: промышленная автоматизация, здравоохранение, транспорт. Эти приложения часто требуют тяжелых вычислений с требованиями к задержке, которые могут превосходить вычислительные возможности мобильных устройств. Мобильные приложения такого типа, как правило, требовательны к ресурсам, требуют интенсивных вычислений и реагирования в реальном времени [2].

Многообещающей технологией для поддержки таких приложений являются мобильные граничные вычисления (MEC) [3], которые размещают граничные серверы рядом с мобильными устройствами, чтобы обеспечить разгрузку вычислений. MEC может существенно увеличить вычислительные мощности мобильных устройств. Однако разгрузка вычислений требует передачи входных данных с мобильных устройств на граничные серверы, что приводит к дополнительной задержке передачи и потреблению энергии [4].

Ограниченная вычислительная мощность на граничных серверах должна быть разделена между несколькими пользователями, что, в свою очередь, может привести к значительному увеличению времени вычислений. Более того, влияние доступных вычислительных и коммуникационных ресурсов в системах MEC связано, и поэтому важно совместное распределение этих двух ресурсов [5].

В данной обзорной статье заглавной будет считаться проблема эффективности распределения ресурсов концепта MEC. Будут рассмотрены методы применимые для различных сценариев и технологий, которые позволят решить вышеперечисленные проблемы [6].

Методы распределения ресурсов

DRL

Рассмотрим реализацию MEC-среды на основе SDN, а именно метод совместного управления коммуникационными и вычислительными ресурсами (Joint Management Method of Communication and Computing Resources) на основе DRL [7].

Коммуникационные и вычислительные ресурсы жизненно важны для загрузки и выполнения вычислительных задач граничными пользователями. Первый тип ресурсов определяет скорость передачи данных и энергопотребление при передаче, последний – ограничивает время вычислений и энергию выполнения задач, загружаемых на сервер MEC. Правильное распределение ресурсов связи и вычислительных ресурсы в MEC на основе SDN имеет большое значение для обеспечения качества обслуживания (QoS) услуг мобильных терминалов. Это может снизить нагрузку на сеть и задержку передачи по каналам связи, улучшить возможности распределения услуг и повысить эффективность использования ресурсов. Цель – разумным образом распределить коммуникационные и вычислительные ресурсы между мобильными устройствами, чтобы минимизировать затраты на выполнение задач всех пользователей в системе. Схема совместного распределения ресурсов связи и вычислительных ресурсов на основе DRL показана на рисунке 1.

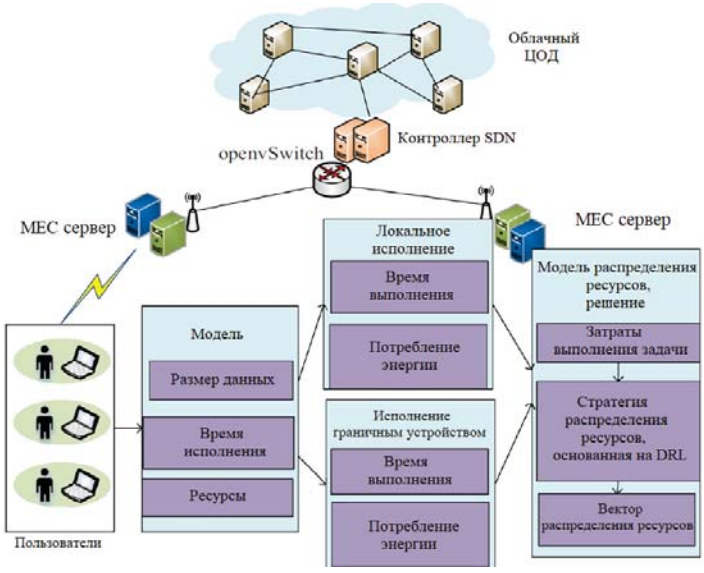


Рис. 1. Схема совместного распределения ресурсов связи и вычислительных ресурсов на основе DRL

Таблица 1

Параметры, используемые для метода DRL

Параметр	Описание
D_k	Вычислительный ресурс, необходимый для выполнения задачи
f_k^l	Локальная вычислительная мощность пользователя
z_k	Потребление энергии каждого цикла ЦП (центрального процессора), завершающего задачу
$T_{k,t}^m$	Задержка передачи пользователя для передачи входных данных на сервер MEC
$T_{k,p}^m$	Задержка обработки сервера MEC при выполнении задачи
$E_{k,t}^m$	Энергопотребление передачи пользователем, передающим входные данные на сервер MEC
$E_{k,p}^m$	Энергопотребление пользователя в процессе работы сервера MEC
$C_{cost}(s,a)$	Цена текущего состояния
$C_{cost}(local)$	Цена рассчитывается локально
$SucR$	Количество успешно выполненных задач
TRE_i	Время выполнения 1-ой задачи
TRS_i	Время начала выполнения 1-ой задачи

В таблице 1 указаны параметры, используемые для метода DRL.

Локальная задержка выполнения задачи:

$$T_k^l = \frac{D_k}{f_k^l}$$

Локальное потребление энергии для выполнения задачи:

$$E_k^l = D_k z_k$$

Задержка пользователя для загрузки задачи на MEC-сервер:

$$T_k^m = T_{k,t}^m + T_{k,p}^m$$

Энергопотребление пользователя, загружающего задачу на MEC-сервер:

$$E_k^m = E_{k,t}^m + E_{k,p}^m$$

Целью DRL является получение наибольшего вознаграждения. Вознаграждение $R_{immediate}$ определяется следующим образом:

$$R_{immediate} = 1 - \frac{C_{cos\,t}(s,a)}{C_{cos\,t}(local)}$$

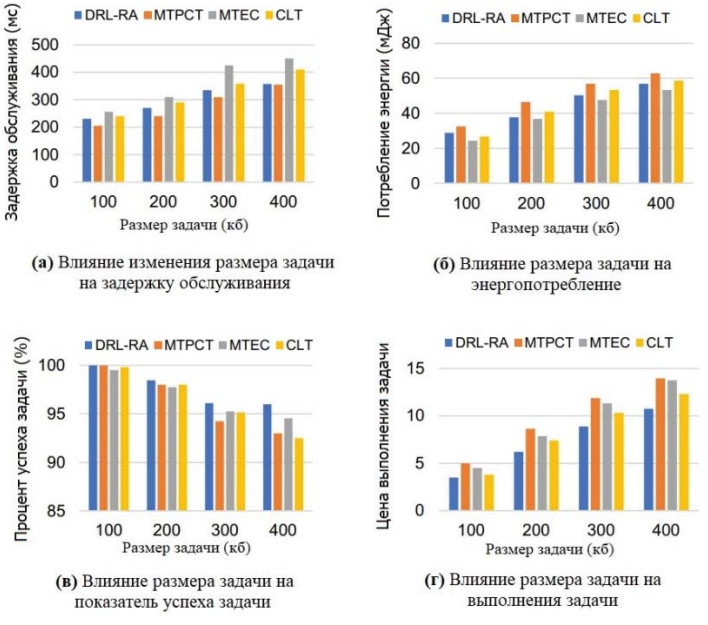
Для пользователя, если локальное вычисление не может удовлетворять предельному значению задержки, тогда его задача должна быть выполнена на сервере MEC.

Среднее время обслуживания – это средняя длительность всех задач в системе от момента генерации до конечного результата выполнения, которая рассчитывается следующим образом:

$$AST = \frac{\sum_{i=1}^{SucR}(TRE_i - TRS_i)}{SucR}$$

Энергопотребление терминала – энергопотребление мобильного терминала от момента генерации до конечного результата выполнения всех задач в системе. Сюда также входит потребление энергии при локальном выполнении и потребление энергии при передаче данных на граничный сервер.

Для оценки возможности реализации схемы совместного распределения коммуникационных и вычислительных ресурсов на основе DRL в качестве показателей оценки эффективности алгоритма выбрано среднее время обслуживания, энергопотребление терминала, скорость выполнения задачи и цена выполнения задачи. Метод совместного распределения коммуникационных и вычислительных ресурсов на основе DRL (DRL-LA), рассматриваемый в этой статье, сравнивается с MTPCT, MTEC и CLT.



(а) Влияние изменения размера задачи на задержку обслуживания

(б) Влияние размера задачи на энергопотребление

(в) Влияние размера задачи на показатель успеха задачи

(г) Влияние размера задачи на выполнения задачи

Рис. 2. Влияние размера задачи на результаты эксперимента

На рисунках 2 (а-г) соответственно показаны изменения средней задержки обслуживания, энергопотребления терминала, скорости выполнения задачи и стоимости выполнения задачи в зависимости от размера вычислительной задачи. При размере задачи 400 Кб DRL-RA по задержке обслуживания ниже, чем MTEC и CLT, потребление энергии меньше, чем у MTPCT и CLT, и больше, чем MTEC, а процент успеха задачи выше, чем у других сравниваемых методов. DRL-RA снижает цену выполнения задачи по сравнению с MTPCT и MTEC.

Результаты экспериментов показали, что данный метод, основанный на SDN, имеет более высокий уровень выполнения задач и более низкую цену выполнения задач. Метод совместного управления коммуникационными и вычислительными ресурсами на основе DRL сокращает время обслуживания и балансирует распределение ресурсов.

JCC

Рассмотрим ресурсоэффективную разгрузку вычислений каждого отдельного пользователя, определив оптимальный профиль совместных коммуникационных и вычислительных (joint communication and computation) ресурсов пользователя [8] для экономически выгодных аукционов, чтобы минимизировать занятость ресурсов.

Существует набор беспроводных базовых станций (БС) (например, макросоты, фемтосоты), а также набор мобильных граничных облаков (сервера MEC), что изображено на рисунке 3. Для БС имеются ортогональные беспроводные подканалы, выделенные для поддержки серверов MEC по требованию. БС напрямую подключена к ближайшему серверу MEC. Модель позволяет нескольким близлежащим БС использовать один и тот же сервер MEC в непосредственной близости.

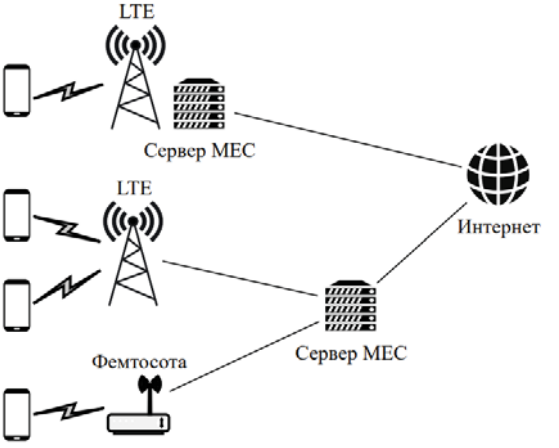


Рис. 3. Модель мобильных граничных вычислений

Набор мобильных пользователей желает приобрести сервер MEC у оператора сети для выполнения ресурсоемких задач. В зависимости от требований приложения к QoS пользователь также имеет максимально допустимое время завершения для своей задачи и обладает ценой, т.е. максимальной суммой, которую пользователь готов заплатить за приобретение сервера MEC для выполнения своей задачи.

Таблица 2

Параметры, используемые для метода JCC

Параметр	Описание
C_i	Общее количество циклов ЦП (центрального процессора) для выполнения компонента задачи
F_n^0	Вычислительная мощность (циклы ЦП в секунду) устройства пользователя
$F_{s_n}^c$	Вычислительная мощность ВМ
D_{ij}	Объем данных, поступающих из узла i в узел j
R_n	Скорость передачи данных для разгрузки вычислений пользователя
$M_{k(n)}$	Общее количество доступных подканалов
$M_{l(n)}$	Вычислительная мощность на связанной с пользователем БС и сервере MEC

В таблице 2 указаны параметры, используемые для метода JCC.

Пользователь может выполнить задачу либо локально на своем собственном устройстве, либо удаленно на сервере MEC. Время для выполнения компонента задачи локально:

$$T_{i,n}^m = \frac{C_i}{F_n^0}$$

При использовании облачных вычислений виртуальная машина (ВМ) связывается с пользователем для выполнения задач в сервере MEC. Чтобы обеспечить гибкость в предоставлении облачных услуг на границе, считаем, что существует набор из различных типов ВМ.

Вычисление времени с помощью облачных вычислений для выполнения компонента задачи на ВМ:

$$T_{i,n}^c = \frac{C_i}{F_{s_n}^c}$$

Предположим, что два компонента задачи выполняются в разных местах (один на устройстве пользователя, а другой в облаке мобильной границы). Затем общий объем данных необходимо передать между этими двумя компонентами через беспроводное соединение.

Общее время передачи данных для разгрузки вычислений между компонентами задачи:

$$T_{ij,n}^{tx} = \frac{D_{ij}}{R_n}$$

Функция занятости ресурса для пользователя – сумма нормализованных коэффициентов занятости пользователя как для коммуникационных, так и для вычислительных ресурсов:

$$\Phi_n(q_n, s_n) = \frac{q_n}{M_{k(n)}} + \frac{F_{s_n}^c}{B_{l(n)}}$$

Оценка производительности предлагаемого метода распределения ресурсов JCC по требованию. Для вычислительной задачи рассматриваются два типа приложений — распознавание лиц и распознавание QR-кода, также в них используются графы задач.

Решения, по которым проводится оценка контроля допуска при распределении ресурсов JCC:

1. Оптимальное решение (Optimal Solution) [9]: вычисляется оптимальное решение проблемы контроля допуска, используя метод перекрестной энтропии (Cross Entropy). Оптимальное решение, полученное методом перекрестной энтропии, используется в качестве основы для нормализации системной полезности других решений;

2. «Грубое жадное» решение (Coarse Greedy solution);

3. Случайный выбор (Random Selection): последовательно и случайным образом выбирается новый пользователь для добавления в набор пользователей до тех пор, пока не произойдёт ограничения емкости.

Проведено моделирование с различным количеством пользователей и серверами MEC, как показано на рисунках 4 и 5 соответственно. В обоих случаях видно, что предлагаемое решение JCC (JCC Solution) обеспечивает повышение производительности как минимум на 34,7% и 88,3% по сравнению с решениями Coarse Greedy и Random Selection соответственно. По сравнению с Optimal Solution видно, что потеря производительности решения JCC составляет не более 14,3%. Это демонстрирует эффективность предлагаемого решения JCC.

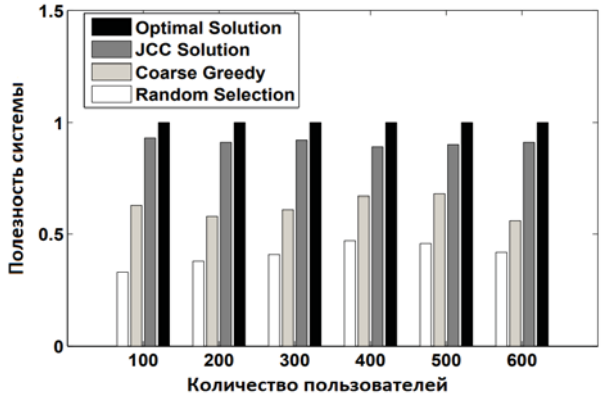


Рис. 4. Производительность системы при разном количестве пользователей

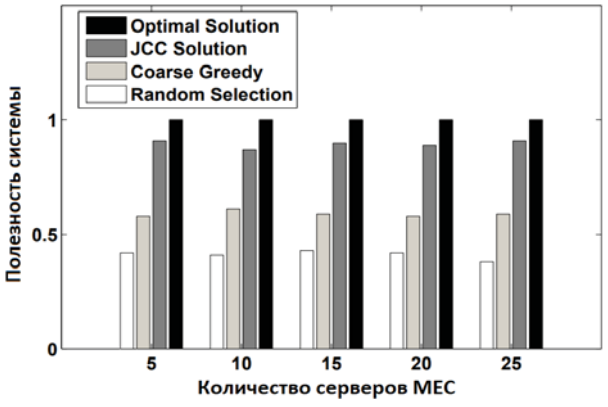


Рис. 5. Производительность системы с разным количеством серверов MEC

Было исследовано время работы контроля допуска. Результаты показаны на рисунке 6. Видно, что JCC Solution является вычислительно эффективным (со временем работы в несколько миллисекунд) с ускорением до 1000 раз по сравнению с Optimal Solution метода перекрестной энтропии.

Предложенный алгоритм контроля допуска полезен для практической реализации.

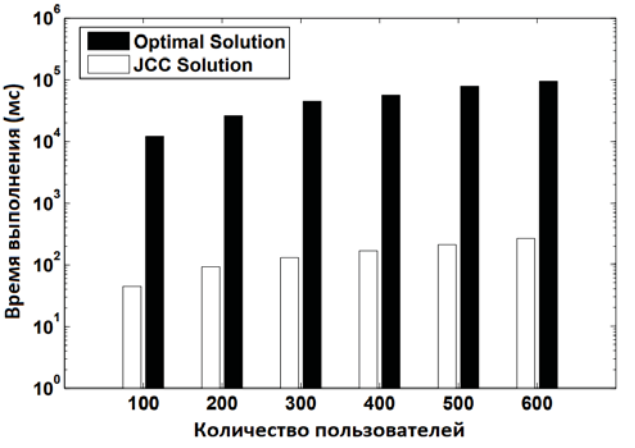


Рис. 6. Сравнение времени выполнения Optimal Solution и JCC Solution

Был рассмотрен метод совместного распределения ресурсов связи и вычислений (JCC) для оператора сети. Получен оптимальный профиль требований к коммуникационным и вычислительным ресурсам для пользователя, который минимизирует занятость ресурсов и в то же время удовлетворяет допустимому времени завершения задач. Была рассмотрена проблема контроля доступа при распределении ресурсов JCC. Всесторонняя оценка производительности показывает, что предложенный метод может обеспечить высокую производительность для мобильных граничных вычислений по требованию.

SCA

Решается проблема минимизации общих затрат мобильной энергии на разгрузку при ограничениях по задержке в отношении параметров связи и вычислений, учитывая совместный характер приложений AR. Проблема распределения ресурсов решается с помощью метода последовательного выпуклого приближения (Successive Convex Approximation – SCA) [10], [11].

Рассматривается мобильная граничная вычислительная система, показанная на рисунке 7.

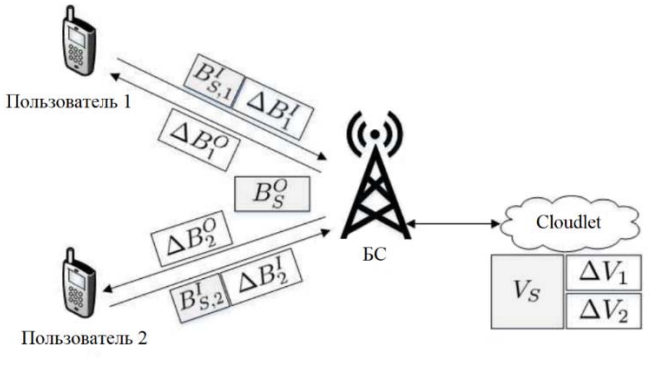


Рис. 7. Выгрузка AR-приложения в облако, подключенное к БС

Имеется набор пользователей, которые запускают AR-приложение на своих мобильных устройствах (МУ) с помощью облачного сервера (Cloudlet). Сервер подключен к базовой станции (BS), которая обслуживает всех пользователей в соте с использованием дуплексного канала с временным разделением (TDD) по каналу с гладким замиранием частоты. Выгруженное приложение имеет общие входные и выходные данные и вычислительные задачи, которые относятся к компонентам Tracker, Mapper и Object Recognizer. Мобильные приложения AR обладают свойствами совместной работы с точки зрения сбора данных по восходящей линии связи, вычислений на границе сети и доставки данных по нисходящей линии связи.

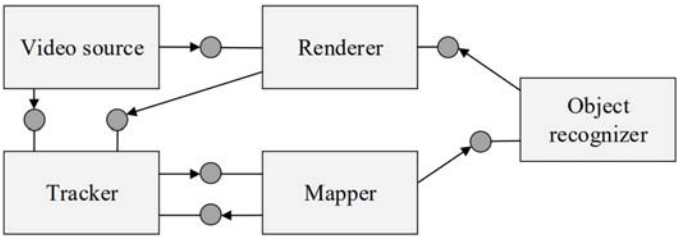


Рис. 8. Пример модели AR-приложения [12]

На рисунке 8 изображена модель AR-приложения. Она включает в себя компоненты Video Source и Renderer, которые необходимо запускать локально на МУ, а также три основных компонента – Mapper, Tracker и Object Recognizer, которые можно выгружать [13].

1. Video Source: получает необработанные видеокадры с мобильной камеры;
2. Tracker: отслеживает положение пользователя относительно окружающей среды;
3. pper: строит модель окружающей среды;
4. Object Recognizer: идентифицирует объекты в окружающей среде;
5. Renderer: подготавливает обработанные кадры к отображению.

Совместный характер компонентов Tracker, Mapper и Object Recognizer можно использовать для снижения энергопотребления мобильных устройств и задержки разгрузки. Несовместные компоненты также могут выполняться локально, если это снижает энергопотребление.

В таблице 3 указаны параметры, используемые для метода SCA.

1. Uplink transmission (Передача по восходящей линии связи):

Скорость для передачи входных бит пользователя по восходящей линии связи определяется выражением:

$$R_k^{ul}\left(P_k^{ul}\right) = \frac{W^{ul}}{K} log_2\left(1 + \frac{\gamma_k P_k^{ul}}{\frac{N_0 W^{ul}}{K}}\right)$$

Энергопотребление мобильной связи при передаче по восходящей линии связи:

$$E_k^{ul}\left(P_k^{ul}, B_{S,k}^I\right) = \left(\frac{P_k^{ul}}{R_k^{ul}(P_k^{ul})} + l_k^{ul}\right)\left(B_{S,k}^I + \Delta B_k^I\right)$$

Таблица 3

Параметры, используемые для метода SCA

Параметр	Описание
W^{ul}	Пропускная способность восходящей линии связи
K	Пользователи
γ_k	Коэффициент усиления
P_k^{ul}	Мощность передачи МУ пользователя
V_S	Общие циклы ЦП (центрального процессора)
ΔV_k	Циклы ЦП для пользователя
f_S	Доля вычислительной мощности F_C для выполнения V_S, подлежащая оптимизации
F_C	Мощность облачного сервера
N_0	Спектральная плотность мощности шума в приемнике
R_k^{ul}	Скорость передачи входных бит пользователя по восходящей линии связи
l_k^{ul}	Параметр, указывающий количество энергии, затраченное МУ на извлечение каждого бита выгруженных данных из Video Source
$B_{S,k}^I$	Раздельные биты
W^{dl}	Пропускная способность нисходящей линии связи
P_k^{dl}	Мощность передачи БС, выделенная для обслуживания пользователя

2. oudlet processing (Облачная обработка):

Время выполнения общих циклов центрального процессора (ЦП):

$$T_S^C = \frac{V_S}{f_S F_C}$$

Время, необходимое для удаленного выполнения циклов ЦП, представляющих интерес для пользователя:

$$T_k^U = \frac{\Delta V_k}{f_k F_C}$$

3. Downlink transmission (Передача по нисходящей линии связи):

Выходные биты, предназначенные для каждого пользователя, передаются одноадресно (unicast) по нисходящей линии связи с использованием полосы пропускания со следующей скоростью:

$$R_k^{dl}\left(P_k^{dl}\right) = \frac{W^{dl}}{K} log_2\left(1 + \frac{\gamma_k P_k^{dl}}{\frac{N_0 W^{dl}}{K}}\right)$$

Общее потребление энергии мобильной связи нисходящей линии связи для пользователя:

$$E_k^{dl}\left(P_k^{dl}, P_M^{dl}\right) = \left(\frac{\Delta B_k^O}{R_k^{dl}(P_k^{dl})} + \frac{B_S^O}{R_{M,k}^{dl}(P_M^{dl})}\right) l_k^{dl}$$

Рассмотрим результаты использования совместного характера приложений AR для мобильных граничных вычислений для минимизации суммарной энергии мобильных устройств, необходимой для разгрузки всех пользователей в условиях ограничений по задержке и мощности.

Сравнивается производительность предложенного метода, в котором передачи по восходящей и нисходящей линии связи, а также вычисления в Cloudlet разделяются на следующие решения по разгрузке:

1. Совместная облачная обработка, передача по нисходящей линии связи (Shared Processing & Downlink): циклы ЦП и выходные данные используются совместно, а входные биты передаются каждым пользователем индивидуально.

2. Совместная передача по восходящей линии связи (Shared Uplink): совместно используются только входные биты, при этом совместного использования вычислений и передачи по нисходящей линии связи не происходит.

3. Включен результат, полученный при решении невыпуклой задачи P.1.

Как показано на рисунке 9, схема Shared Processing & Downlink обеспечивает экономию энергии около 37% по сравнению с раздельной разгрузкой (Separate Offloading).

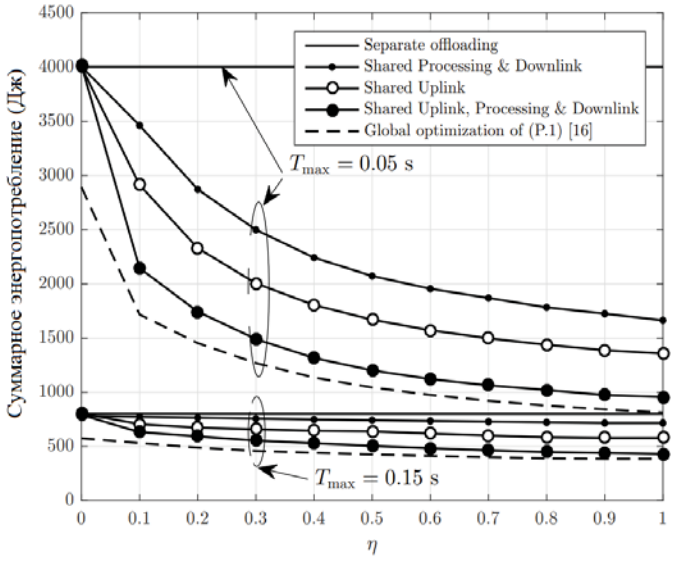


Рис. 9. Зависимость суммарного энергопотребления от доли η общих данных в восходящей и нисходящей линии связи, а также общих циклов ЦП, выполняемых в Cloudlet

Этот выигрыш можно объяснить увеличением времени, доступного для передачи по восходящей линии связи, благодаря более коротким периодам выполнения и передачи по нисходящей линии связи, что снижает соответствующую энергию разгрузки.

При тех же условиях экономия энергии около 50% по сравнению с Separate Offloading, обеспечиваемая Shared Uplink, обусловлена способностью системы регулировать доли общих данных, передаваемых каждым пользователем по восходящей линии связи, в зависимости от текущего состояния канала.

Эти два преимущества в совокупности обеспечивают экономию энергии предложенной схемы общей выгрузки данных по сравнению с общепринятой Separate Offloading, составляющую около 63%.

На рисунке 9 также показано, что в этом примере SCA дает решение, близкое к глобальному оптимуму невыпуклой задачи P.1.

Игра Штакельберга

Предлагается стратегия распределения вычислительных ресурсов для повышения производительности системы «Умный дом» [14] с MEC. Архитектура: пользовательское оборудование (UE), точки доступа (AP) и сервера MEC. AP может агрегировать данные, генерируемые UE, и улучшать граничную обработку, как показано на рисунке 10. На основе игры Штакельберга [15] разрабатывается эффективная стратегия распределения вычислительных ресурсов.

Имеется набор UE, сервер MEC имеет список предпочтений для всех точек доступа, AP имеет список предпочтений своих пользователей. Количество единиц вычислительных ресурсов, которые могут быть распределены сервером MEC определены как блок вычислительных ресурсов (CRB). Каждый CRB может предоставлять вычислительные услуги.

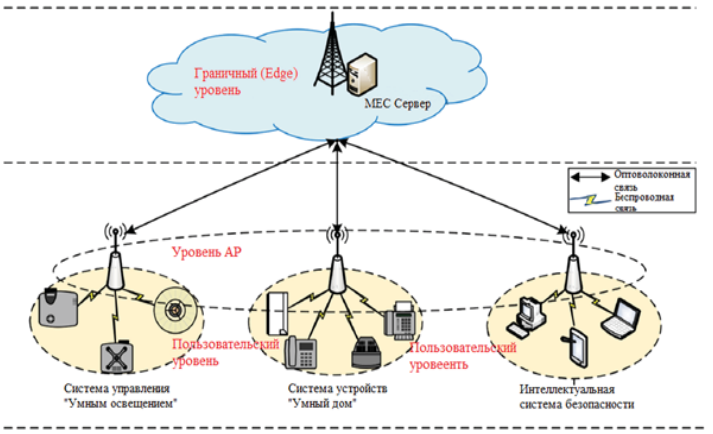


Рис. 10. Архитектура системы

Сервер может выделять ресурсы связанным беспроводным базовым станциям (BS). Каждая BS имеет AP, которая связана с обслуживаемым оборудованием UE. При выполнении задач в режиме реального времени UE может отправлять запросы на AP и выгружать задачи на сервер MEC через каналы связи после получения ответа от AP.

Таблица 4

Параметры, используемые для метода Игра Штакельберга

Параметр	Описание
M_k	Количество UE, подключенных к AP
λ_j	Скорость поступления рабочей нагрузки
μ	Скорость обслуживания каждого CRB
r_j^k	Установление цены AP над UE
p_k	Цена аренды, по которой AP арендует CRB с серверов
c_k	Цена передачи от сервера MEC к AP
Q_k	Количество CRB, которое требует AP
β_j^k, γ_j^k	Весовые коэффициенты функции полезности UE
t_{th}	Граница задержки
q_j	Количество CRB, которое требует UE
q_j^k	Оптимальное количество CRB для UE, чтобы получить максимальную полезность
$\tilde{Q}_k$	Количество CRB, выделенное AP

Сервер MEC напрямую связывается с точками доступа и распределяет между ними ресурсы. После того как AP приобретет вычислительные ресурсы у сервера, она может принять оптимальное решение о распределении подключенного UE для повышения полезности.

В таблице 4 указаны параметры, используемые для метода Игра Штакельберга.

В игре Шакельберга AP выступает в роли лидера (Leader) [16], прогнозируя цены на обслуживание и необходимое количество CRB для UE. UE действует как последователь (Follower), покупая CRB у AP, зная установленную цену. В течение всего процесса AP может наблюдать за действиями UE и окончательно определять цену услуги.

Полезность AP:

$$W_k^a = \sum_{i=1}^{M_k} \left(\frac{\lambda_j^k}{\mu} + \frac{\lambda_j^k}{\mu * \sqrt{\frac{\beta_j^k * r_j^k}{\gamma_j^k}}} * r_j^k - p_k * Q_k \right)$$

Оптимальная цена для получения максимальной полезности AP, то есть установка цены AP над UE:

$$r_j^k = \frac{\gamma_j^k}{\beta_j^k} * \left(\frac{\mu * t_{th} - \lambda_j^k}{\lambda_j^k} \right)^2$$

AP может получить оптимальное количество CRB, которое она планирует арендовать с сервера MEC. Количество CRB, выделенные для AP:

$$\tilde{Q}_k = \sum_{i=1}^{M_k} q_j^k$$

Зная количество CRB, которые AP хочет арендовать, сервер MEC принимает решение о том, как распределить CRB по AP. В зависимости от стоимости аренды и стоимости передачи сервер имеет разные предпочтения перед всеми AP. AP с более высоким предпочтением имеет приоритет в процессе выделения вычислительных ресурсов. Различные расстояния между сервером MEC и AP влияют на стоимость связи между ними. А полезность сервера MEC связана с ценой и стоимостью аренды.

Предпочтение сервера MEC для всех AP можно измерить ценой для связи и ценами на аренду:

$$L_i = p_i - c_i$$

После аренды CRB с сервера MEC, AP необходимо перераспределить их UE.

Результаты моделирования предложенного метода. Как показано на рисунке 11, схемы, основанные на игре Штакельберга, все время превосходят ситуацию с аукционной игрой. Равновесие Штакельберга гарантирует, что оптимальное количество CRB и цена услуги, выбранные UE, могут максимизировать полезность сервера MEC.

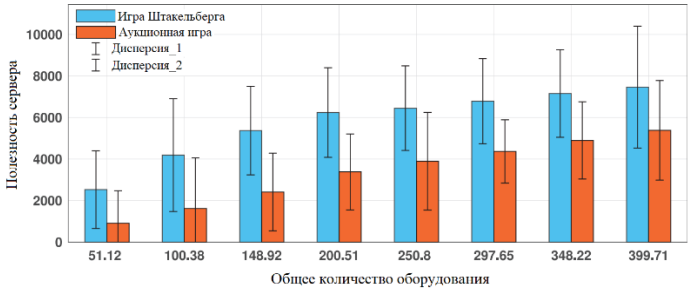


Рис. 11. Сравнение схем, основанных на игре Штакельберга и аукционной игре

На рисунке 12 показано, что полезность сервера MEC меняется в зависимости от скорости поступления рабочей нагрузки и скорости обслуживания CRB. Полезность возрастает с ростом рабочей нагрузки. Более высокая скорость рабочей нагрузки приводит к тому, что UE требуется больше CRB для выполнения своих задач, следовательно, сервер может распределить больше CRB и получать более высокую прибыль, пока все CRB не будут распределены. Кроме того, в зависимости от расстояния цена аренды, которую AP платит за сервер MEC, остается неизменной. Таким образом, доходы, которые получает сервер MEC, превышают расходы.

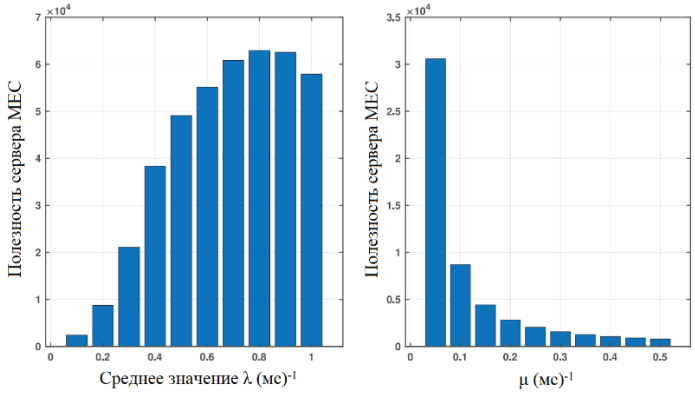


Рис. 12. Полезность сервера MEC

Также, исследовалась зависимость полезности сервера MEC от увеличивающихся количества доступных CRB, когда в сети развернуто различное количество AP. Когда CRB увеличиваются, сервер MEC может предоставлять вычислительные услуги для большего количества UE.

Игра Штакельберга предлагается для решения проблемы покупки ресурсов и проблемы ценообразования для AP и решения проблемы распределения вычислительных ресурсов для сервера MEC и AP.

При максимальной полезности, стратегия распределения может повысить эффективность распределения и гарантировать QoS в реальном времени для UE. Результаты демонстрируют влияние различных параметров, таких как количество CRB и скорость поступления рабочей нагрузки. По сравнению с механизмом распределения, основанным на аукционной игре, предлагаемая структура может обеспечить более высокую полезность для сервера. Сервер MEC может быть выгодным. Другими словами, вычислительные ресурсы могут быть распределены оптимально.

Актуальность и применение в перспективных работах

Данные методы были рассмотрены для различных сценариев и технологий, также для каждого метода были поставлены цели разного значения (таблица 5). Все методы оказались эффективными для достижения конкретных целей, а также являются актуальными для решения проблемы распределения ресурсов в настоящее время.

Таблица 5

Технологии/сценарии и цели методов распределения ресурсов

Методология	Технология/ Сценарий	Цели
DRL (Deep Reinforcement Learning)	SDN (Software Defined Network)	Сокращение времени обслуживания и балансировка распределения ресурсов
JCC (Joint Communication and Computation)	В общем	Распределение ресурсов по требованию
SCA (Successive Convex Approximation)	AR (Augmented Reality) приложения	Снижение энергопотребления при распределении ресурсов
Игра Штакельберга	Smart Home	Повышение эффективности распределения ресурсов, используя сервер MEC выгодно

Так, для исследований объединения серверов MEC для динамического распределения ресурсов, наиболее актуальным методом будет являться метод Игра Штакельберга

В ситуации нехватки ресурсов конкретного сервера MEC это решение может закупать виртуальные блоки вычислительных ресурсов (CRB) с разных серверов. Однако в этом случае задержка может стать большой, поскольку поставщики услуг предоставляет важные приложения, а также службы для граничных пользователей и устройств IoT. Таким образом, приобретение CRB с других ближайших серверов MEC может облегчить ситуацию. Опять же, приобретение любого количества CRB не является практическим подходом для преодоления пиковой нагрузки, так как дополнительные CRB останутся неиспользованными. Следовательно, сервер MEC должен определить оптимальное количество CRB, требуемое поставщикам граничных услуг, так как они приобретают CRB из MEC. Это должно максимизировать использование недоиспользованных граничных ресурсов и сократить время предоставления услуг одновременно.

Заключение

В данной статье были рассмотрены 4 метода: DRL, JCC, SCA, игра Штакельберга.

Каждый метод был применен для конкретных сценариев/технологий – SDN, в общем, AR, Smart Home. Также

была рассмотрена конкретная перспектива применения метода игра Штакельберга.

Перечисленные методы позволяют эффективно распределять ресурсы в сетях, с использованием серверов MEC, а также являются актуальными в настоящее время.

К наиболее актуальному методу на данный момент можно отнести «Игра Штакельберга», так как данный метод распределения ресурсов может повысить эффективность распределения и гарантировать QoS в реальном времени для пользовательских терминалов. По сравнению с остальными механизмами распределения предлагаемая структура может обеспечить более высокую полезность для сервера. Это может быть выгодно для того, чтобы избежать простоев для оборудования, также вместе с этим предлагается наиболее финансово выгодная перспектива, так как используются элементы сети по назначению, что позволяет прийти к минимуму бездействия.

Литература

1. *Khan M.A., Ahmadon M.A.* Trends and Challenges in Mobile Edge Computing for the Next Generation Massive Internet of Things // Evolution of Information, Communication and Computing System, 2023, vol. 4, pp. 28-42.

2. *Bebortta S., Singh A., Senapati D.* Performance analysis of multi-access edge computing networks for heterogeneous IoT systems // Materials Today: Proceedings, 2022, vol. 58, pp. 267-272.

3. *Liang B., Gregory M., Li S.* Multi-access Edge Computing fundamentals, services, enablers and challenges: A complete survey // Journal of Network and Computer Applications, 2022, vol. 199.

4. *Hu H., Wang P.* Computation Offloading Game for Multi-Channel Wireless Sensor Networks // Sensors, 2022, vol. 22, no. 22, p. 8718.

5. *Vhora F.* A comprehensive survey on mobile edge computing: Challenges, tools, applications // ICCMC, 2020, no. 6, pp. 1-6.

6. *Liu H., Long X., Li Z., Long S., Ran R.* Joint Optimization of Request Assignment and Computing Resource Allocation in Multi-Access Edge Computing // IEEE Transactions on Services Computing, 2023, vol. 16, no. 2, pp. 1254-1267.

7. *Li C., Zhang Y., Luo Y.* Deep reinforcement learning-based resource allocation and seamless handover in multi-access edge computing based on SDN // Knowledge and Information Systems, 2021, no. 63, pp. 2479-2511.

8. *Chen X.* Efficient resource allocation for on-demand mobile-edge cloud computing // IEEE Transactions on Vehicular Technology, 2018, no. 67, pp. 2-12.

9. *Inage S., Hebishima H.* Application of Monte Carlo stochastic optimization (MOST) to deep learning // Mathematics and Computers in Simulation, 2022, vol. 199, pp. 257-271.

10. *Al-Shuwaili A., Simeone O.* Energy-Efficient Resource Allocation for Mobile Edge Computing-Based Augmented Reality Applications // IEEE Wireless Communications Letters, June 2017, vol. 6, no. 3, pp. 398-401.

11. *Chen Y., Li C., Hu Y., Lu Z.* A parallel non-convex approximation framework for risk parity portfolio design // Parallel Computing, 2023, vol. 116.

12. *Yoo S., Jeong S., Kim J., Kang J.* Cache-assisted Mobile Edge Computing over Space-Air-Ground Integrated Networks for Extended Reality Applications // arXiv, 2023.

13. *Bohez S., Turck J. D., Verbelen T., Simoens P., Dhoedt B.* Mobile, Collaborative Augmented Reality Using Cloudlets // International Conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications, 2013, pp. 45-54.

14. *Badar A.Q.H., Anvari-Moghaddam A.* Smart Home Energy Management System – A Review //Advances in Building Energy Research, 2022, no. 16, pp. 118-143.

15. *Guo S.* Mobile edge computing resource allocation: A joint Stackelberg game and matching strategy // International Journal of Distributed Sensor Networks, 2019, vol. 15.

16. *Stein A., Salvioli M., Garjani H., Dubbeldam J., Viossat Y.* Stackelberg evolutionary game theory: how to manage evolving systems // Phil. Trans. R. Soc. B, 2023.

RESOURCE ALLOCATION METHODS OF MOBILE EDGE COMPUTING CONCEPT

ELENA V. CHIPSANOVA
St. Petersburg, Russia, enchip@mail.ru

VASILY S. ELAGIN
St. Petersburg, Russia, v.elagin@spbgut.ru

KEYWORDS: *MEC, resource allocation, DRL, JCC, SCA, Stackelberg game*

ABSTRACT

Introduction: Nowadays, there are a large number of applications that are computationally intensive and also need to have minimal latency. These applications can exceed the computing capabilities of mobile devices. A promising technology to support such applications is mobile edge computing (MEC). In this concept, edge servers are placed close to mobile devices to provide computational offloading. MEC can significantly increase the computing power of mobile devices. However, computation offloading requires data transfer from mobile devices to edge servers. This may result in additional data transmission delay and higher power consumption. In this regard, the issue of efficient resource allocation for the concept of mobile edge computing in telecommunication networks arises. **Purpose of the work:** the main objective is to familiarize with current methods that can improve resource allocation for the MEC concept. Also, among the methods considered, it is necessary to identify the most relevant for modern telecommunication networks. The analysis approach was chosen because it is the most suitable for the tasks at hand. **Result:** methods for resource allocation in mobile edge computing concept are considered, and the most relevant method among those presented is identified.

REFERENCES

1. Khan M. A., Ahmadon M. A. Trends and Challenges in Mobile Edge Computing for the Next Generation Massive Internet of Things. *Evolution of Information, Communication and Computing System*, 2023, vol. 4, pp. 28-42.

2. Bebortta S., Singh A., Senapati D. Performance analysis of multi-access edge computing networks for heterogeneous IoT systems. *Materials Today: Proceedings*, 2022, vol. 58, pp. 267-272.

3.Liang B., Gregory M., Li S. Multi-access Edge Computing fundamentals, services, enablers and challenges: A complete survey. *Journal of Network and Computer Applications*, 2022, vol. 199.

4. Hu H., Wang P. Computation Offloading Game for Multi-Channel Wireless Sensor Networks. *Sensors*, 2022, vol. 22, no. 22, p. 8718.

5. Vhora F. A comprehensive survey on mobile edge computing: Challenges, tools, applications. *ICCMC,* 2020, no. 6, pp. 1-6.

6. Liu H., Long X., Li Z., Long S., Ran R. Joint Optimization of Request Assignment and Computing Resource Allocation in Multi-Access Edge Computing. I*EEE Transactions on Services Computing*, 2023, vol. 16, no. 2, pp. 1254-1267.

7. Li C., Zhang Y., Luo Y. Deep reinforcement learning-based resource allocation and seamless handover in multi-access edge computing based on SDN. *Knowledge and Information Systems*, 2021, no. 63, pp. 2479-2511.

8. Chen X. Efficient resource allocation for on-demand mobile-edge cloud computing. *IEEE Transactions on Vehicular Technology,* 2018, no. 67, pp. 2-12.

9. Inage S., Hebishima H. Application of Monte Carlo stochastic optimization (MOST) to deep learning. *Mathematics and Computers in Simulation*, 2022, vol. 199, pp. 257-271.

10. Al-Shuwaili A., Simeone O. Energy-Efficient Resource Allocation for Mobile Edge Computing-Based Augmented Reality Applications. *IEEE Wireless Communications Letters*, June 2017, vol. 6, no. 3, pp. 398-401.

11. Chen Y., Li C., Hu Y., Lu Z. A parallel non-convex approximation framework for risk parity portfolio design. *Parallel Computing*, 2023, vol. 116.

12. Yoo S., Jeong S., Kim J., Kang J. Cache-assisted Mobile Edge Computing over Space-Air-Ground Integrated Networks for Extended Reality Applications. arXiv, 2023.

13. Bohez S., Turck J. D., Verbelen T., Simoens P., Dhoedt B. Mobile, Collaborative Augmented Reality Using Cloudlets. *International Conference on MOBILe Wireless MiddleWARE, Operating Systems, and Applications*, 2013, pp. 45-54.

14. Badar A. Q.H., Anvari-Moghaddam A. Smart Home Energy Management System – A Review. *Advances in Building Energy Research,* 2022, no. 16, pp. 118-143.

15. Guo S. Mobile edge computing resource allocation: A joint Stackelberg game and matching strategy. *International Journal of Distributed Sensor Networks*, 2019, vol. 15.

16. Stein A., Salvioli M., Garjani H., Dubbeldam J., Viossat Y. Stackelberg evolutionary game theory: how to manage evolving systems. *Phil. Trans. R. Soc. B*, 2023.

INFORMATION ABOUT AUTHOR:

Elena V. Chipsanova, Graduate student, The Bonch-Bruevich Saint-Petersburg State University of Telecommunications
Vasily S. Elagin, Head of the Department of Infocommunication Systems, The Bonch-Bruevich Saint-Petersburg State University of Telecommunications

For citation: Chipsanova E.V., Elagin V.S. Resource allocation methods of Mobile Edge Computing concept. *H&ES Reserch*. 2024. Vol. 16. No. 1. P. 4-13. doi: 10.36724/2409-5419-2024-16-1-4-13 (In Rus)

doi: 10.36724/2409-5419-2024-16-1-14-20

ПОСТРОЕНИЕ КОМБИНИРОВАННОГО КОДА НА ОСНОВЕ ТУРБОКОДА С ДОПОЛНИТЕЛЬНЫМ КОДИРОВАНИЕМ БИТ

СИДОРЕНКО
Александр Анатольевич[1]

АННОТАЦИЯ

Предмет: Методы повышения эффективности турбокодов. **Цель работы:** Оценка эффективности комбинированного кода, построенного на основе турбокода при выборочном дополнительном кодировании бит его кодового слова. **Методология проведения исследований:** Оценка эффективности осуществлялась на основании результатов декодирования при разном значении вероятности появления ошибочного бита в канале передачи данных p_B. В исследованиях использовался вариант построения турбокода на основе систематического сверточного кода. Предложена методика повышения эффективности выборочного кодирования путем умножения значений бит, подвергшихся и не подвергшихся дополнительному кодированию на различные коэффициенты. **Результаты исследований:** При дополнительном кодировании одного бита кодового слова турбокода зафиксировано снижение вероятности появления ошибочного бита в декодированном информационном сообщении p_D: при $p_B=10^{-1}$ в 1,34 раза, при $p_B=10^{-2}$ в 1,78 раза, при $p_B=10^{-3}$ в 2,18 раза. Изменение числа бит кодового слова турбокода, подвергшихся дополнительному кодированию, позволяет в широких пределах варьировать корректирующую способность кода. Предложенный метод повышения эффективности комбинированного кода показал свою действенность. При использовании данного метода для случая дополнительного кодирования трех информационных бит кодового слова снижение pD составило: 1,07 раза при $p_B=10^{-1}$, 2,76 раза при $p_B=10^{-2}$, 5,27 раза при $p_B=10^{-3}$. **Область применения результатов:** Полученные в ходе исследования результаты могут быть использованы при проектировании и построении разнообразных систем передачи и обработки данных. Разработанный программный тренажер, моделирующий работу комбинированного кода, может быть применен как в научных, так и в учебных целях. **Выводы:** В публикации рассмотрен метод повышения эффективности турбокода путем создания на его основе комбинированного кода. Повышение эффективности выражается как в росте корректирующей способности кода, так и в возможности гибкой адаптации параметров кода к изменениям состояния канала передачи данных.

Сведения об авторе:

[1] к.т.н., доцент, эксперт отделения погранологии, Международная академия информатизации, Москва, Россия, kotofeykin@mail.ru

КЛЮЧЕВЫЕ СЛОВА: *помехоустойчивость, телекоммуникационные системы, помехоустойчивое кодирование, сверточные коды, турбокоды*

Для цитирования: *Сидоренко А.А.* Построение комбинированного кода на основе турбокода с дополнительным кодированием бит // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16. № 1. С. 14-20. doi: 10.36724/2409-5419-2024-16-1-14-20

Введение

Информация играет все более возрастающую роль во всех видах человеческой деятельности. За последнее время резко повысились требования, предъявляемые к системам передачи информации. Необходимо вести передачу все больших объемов информации на все большие расстояния со все большей скоростью. При этом энергетические ресурсы передатчика, как правило, ограничены. Растут требования и к достоверности передачи данных. В реальной системе передачи информации достоверность определяется степенью искажения передаваемого сигнала. Эти искажения зависят от свойств и технического состояния системы, а также от интенсивности и характера помех. В правильно спроектированной и технически исправной системе передачи информации искажения сигналов в большей степени обусловлены воздействием помех. В этом случае достоверность передачи информации определяется помехоустойчивостью системы, то есть способностью системы противостоять вредному влиянию естественных и искусственных помех на передачу сообщений [1, 2]. Высокий уровень искусственных помех характерен для крупных городов, промышленных центров, а также для движущихся объектов (автомобилей, самолетов, электропоездов и т. п.).

Под воздействием неблагоприятных факторов в передаваемых данных возникают одиночные и групповые ошибки. Для повышения помехоустойчивости в телекоммуникационных системах все чаще используют помехоустойчивые коды [2 – 4]. Помехоустойчивым кодам не важна причина возникновения ошибок. При помехоустойчивом кодировании с целью обнаружения и исправления ошибок в информацию вводятся избыточные символы. Из всех помехоустойчивых кодов с исправлением ошибок турбокоды обладают лучшей корректирующей способностью в каналах передачи данных с высокой вероятностью появления ошибочного бита [1, 5].

Оценка эффективности комбинированного кода

В настоящее время турбокоды применяются в разнообразных телекоммуникационных системах, используются в стандартах связи и телевидения: CDMA-2000, INTELSAT, IEEE 802.16, WiMax, DVB-RCS, UMTS, 3GPP LTE, TIA-1008, DMR. Принципы построения турбокодов изложены в многочисленных источниках [1, 6, 7]. Основных принципов всего два:

- построение кодов с кодовыми словами, обладающими квазислучайными свойствами;
- построение декодеров, реализующих алгоритмы итеративного декодирования, в ходе которых осуществляется передача мягкого решения с выхода одного декодера на вход другого с повторением данной процедуры.

Как видно из структурной схемы N – размерного кодера турбокода (см. рис. 1) поток передаваемых данных состоит из последовательности информационных бит и N последовательностей проверочных бит [6, 8].

Одним из недостатков турбокода является его значительная избыточность, сильно снижающая реальную скорость передачи данных.

Турбокоды обладают достаточно высокой корректирующей способностью, однако их характеристики все равно остаются весьма далекими от теоретических пределов, определенных К. Шенноном. В публикации будет рассмотрен вариант повышения корректирующей способности турбокода путем создания на его основе комбинированного кода.

С целью повышения корректирующей способности помехоустойчивых кодов их объединяют в разнообразные кодовые конструкции. Очевидно, что дополнительное кодирование кодовых бит турбокода повышает их надежность, что должно благоприятно отразится на корректирующей способности кода в целом. Однако, следует учесть, что кодовое слово турбокода уже обладает существенной избыточностью. В связи с этим, предлагается производить дополнительное кодирование кодовых бит турбокода избирательно. Будем называть такой комбинированный код – турбокодом с дополнительным кодированием бит (далее – ТДК). Определим, какие биты кодового слова турбокода наиболее целесообразно подвергать дополнительному кодированию, после чего произведем анализ корректирующей способности ТДК.

В соответствии со структурой кодера турбокода (см. рис. 1) одни и те же информационные биты участвуют в формировании N кодовых слов, что определяет наибольшую ценность этих бит. Ошибка, возникшая в одном информационном бите, превращается в N ошибок при декодировании. Можно предположить, что наибольший негативный эффект на успешность декодирования кодового слова оказывают ошибки, возникшие в информационных битах.

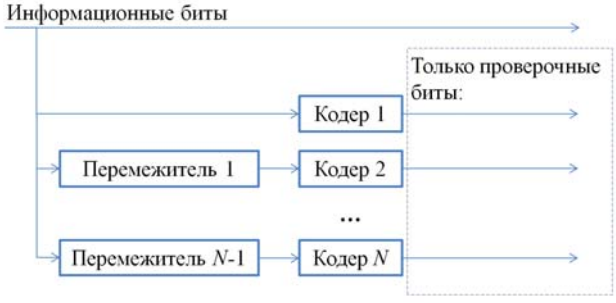


Рис. 1. Структурная схема кодера турбокода

Подтвердим высказанное предположение о более высокой чувствительности информационных бит, рассмотрев примеры декодирования кодовых слов турбокода, содержащих ошибки.

В качестве практического примера будем рассматривать турбокод на основе систематического сверточного кода (далее – ССК). Структурная схема кодера турбокода изображена на рисунке 2 и соответствует рисунку 1 при N=2. В таком случае кодовое слово турбокода состоит из блока информационных бит и двух блоков проверочных бит с выходов первого и второго кодеров ССК. Структурная схема кодового слова турбокода приведена на рисунке 3.

При декодировании будет применяться хорошо изученный и подробно описанный в разнообразных источниках [8, 9] алгоритм вычисления решения о значении принятого символа по максимуму апостериорной вероятности (далее – MAP от «maximum a posteriori»).

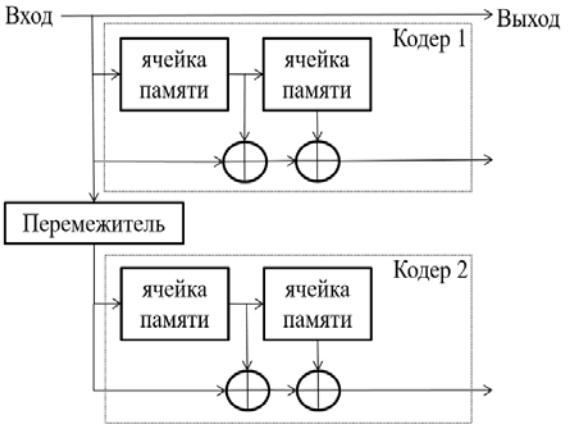


Рис. 2. Структурная схема кодера турбокода на основе ССК

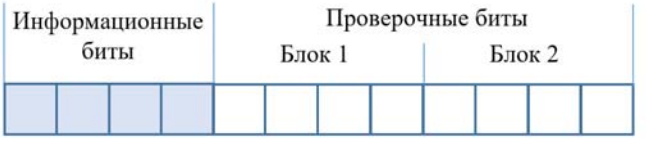


Рис. 3. Структура кодового слова турбокода

Известно, что корректирующая способность турбокодов значительно возрастает при декодировании мягких решений детектора, однако наличие на выходе детектора так называемых «сильных» и «слабых» бит [9] может помешать правильной оценке результатов исследования. В публикации будут рассмотрены результаты декодирования жесткого решения детектора. Декодирование кодового слова турбокода будет осуществляться параллельным декодером, структурная схема которого изображена на рисунке 4 [6]. Первый и второй декодеры полностью идентичны.

Кратко рассмотрим процесс формирования кодового слова турбокода для информационной последовательности 0100. В результате кодирования первым кодером ССК формируется последовательность проверочных бит 0111. Пусть при перемежении информационной последовательности первый по счету бит меняется с третьим, а второй с четвертым, тогда на вход второго кодера ССК поступает последовательность 0001. В результате кодирования на выходе второго кодера формируется последовательность проверочных бит 0001. В соответствии с рисунком 3, имеем следующую сформировавшуюся последовательность выхода кодера турбокода: 0100 0111 0001. Передача и декодирование бит будет осуществляться при их представлении биполярными сигналами, тогда кодовое слово турбокода имеет следующий вид: -11-1-1 -1111 -1-1-11.

Для оценки корректирующей способности описанного турбокода произведена разработка в MS Excel реализации декодера, позволяющего осуществлять декодирование одного кодового слова и функционирующего на основе алгоритма MAP.

Оценку корректирующей способности будем производить при наличии в кодовом слове турбокода трех ошибок и их расположении поочередно в каждом из трех блоков кодового слова. При расположении ошибок в блоке информационных бит кодовое слово имеет вид: 1-1-11 -1111 -1-1-11. Ошибочные символы здесь и далее подчеркнуты. Результаты по итогам трех итераций декодирования приведены в таблице 1. Ошибочно декодированы биты № 1 и № 4. При расположении ошибок в первом блоке проверочных бит кодовое слово имеет вид: -11-1-1 1-11-1 -1-1-11. Результаты декодирования приведены в таблице 2. Декодирование осуществлено без ошибок. При расположении ошибок во втором блоке проверочных бит кодовое слово имеет вид: -11-1-1 -1111 11-1-1. Результаты декодирования приведены в таблице 3. Декодирование осуществлено без ошибок. Анализ результатов декодирования кодового слова при расположении ошибок в его разных частях подтвердил предположение, что наиболее негативный эффект на результат декодирования оказывает появление ошибок в информационных битах кодового слова турбокода.

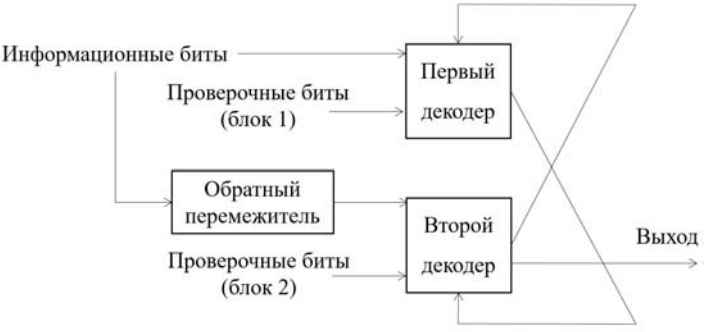


Рис. 4. Структурная схема параллельного декодера турбокода

Таблица 1

Значения мягкого выхода декодера турбокода при расположении ошибок в информационных битах кодового слова

№ бита	Значение переданного бита	Итерации декодирования		
		1	2	3
1	–1	5,5	10,2	21,4
2	1	-4,9	-9,1	-21,2
3	–1	-8,1	-19,3	-43,6
4	–1	4,8	8,2	15,4

Таблица 2

Значения мягкого выхода декодера турбокода при расположении ошибок в первом блоке проверочных бит кодового слова

№ бита	Значение переданного бита	Итерации декодирования		
		1	2	3
1	–1	-4,8	-9,9	-20,6
2	1	6,5	9,6	18,3
3	–1	-7,5	-17,5	-36,7
4	–1	-8,2	-13,1	-29,4

Таблица 3

Значения мягкого выхода декодера турбокода при расположении ошибок во втором блоке проверочных бит кодового слова

№ бита	Значение переданного бита	Итерации декодирования		
		1	2	3
1	–1	-7,5	-17,5	-36,7
2	1	8,2	13,1	29,4
3	–1	-4,8	-9,9	-20,6
4	–1	-6,5	-9,6	-18,3

С целью количественной оценки эффективности ТДК, на языке программирования *Python* разработан прогаммный тренажер [10], содержащий в своем составе следующие основные компоненты:

– генератор информационной последовательности;
– одер турбокода;
– блок моделирования выхода детектора приемника с заданной вероятностью появления ошибочного бита p_B;
– блок моделирования выхода декодера дополнительного кода с вероятностью появления ошибочного бита p_V;
– де одер турбокода.

В программном тренажере генерация массивов информационных бит и массивов ошибок осуществляется с использованием библиотеки «*scipy*» и ее субпакета «*stats*», содержащего статистические функции [11]. В качестве дополнительного кода будем использовать хорошо известный код Голея. В таблице 4 приведена справочная информация о значении вероятности битовой ошибки в декодированном информационном сообщении для кода Голея при разных значениях вероятности появления ошибочного бита в канале передачи данных p_B [8].

Таблица 4

Вероятность появления ошибочного бита в декодированном информационном сообщении для кода Голея

$p_B=10^{-1}$	$p_B=10^{-2}$	$p_B=10^{-3}$
$3{,}1\cdot10^{-2}$	$1{,}9\cdot10^{-5}$	$2{,}3\cdot10^{-9}$

Рассмотрим итоги работы программного тренажера. В таблице 5 приведены результаты рассчета вероятности появления ошибочного бита в декодированном информационном сообщении p_D при кодировании дополнительным кодом двух информационных бит и двух проверочных бит. Произведя анализ полученных данных, можно сделать вывод о значительном выигрыше в корректирующей способности ТДК при дополнительном кодировании информационных бит кодового слова.

При осуществлении дополнительного кодирования бит кодового слова турбокода, наиболее целесообразно кодировать информационные биты. Структурная схема возможного варианта дополнительного кодирования кодом Хемминга информационных бит четырех кодовых слов турбокода представлена на рисунке 5. Направление кодирования кодом Хемминга показано пунктирными линиями.

Таблица 5

Корректирующая способность ТДК при введении дополнительного кодирования в разные части кодового слова турбокода

Особенности введения дополнительного кодирования в кодовое слово турбокода	$p_B=1\cdot10^{-1}$	$p_B=1\cdot10^{-2}$	$p_B=1\cdot10^{-3}$
Дополнительное кодирование двух информационных бит	$1{,}55\cdot10^{-2}$	$1{,}34\cdot10^{-5}$	$1{,}10\cdot10^{-8}$
Дополнительное кодирование двух проверочных бит	$2{,}54\cdot10^{-2}$	$1{,}15\cdot10^{-4}$	$1{,}02\cdot10^{-7}$

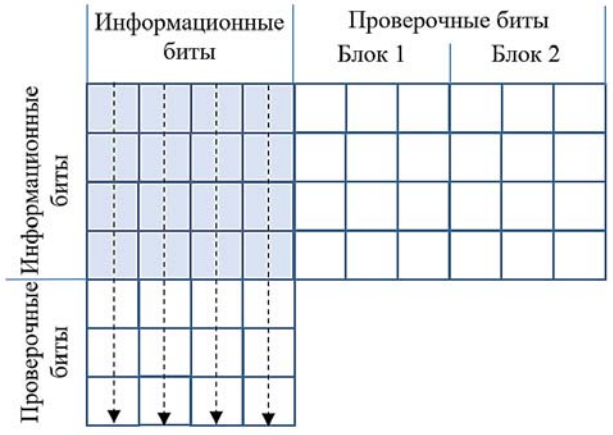


Рис. 5. Структурная схема реализации дополнительного кодирования информационных бит кодового слова турбокода

Возникающую в ТДК избыточность можно снизить, дополнительно кодируя только часть информационных бит. Часто при построении кодового слова турбокода с целью повышения надежности последних информационных бит, к ним добавляют терминирующие биты, значения которых заранее известны. При наличии в информационной последовательности терминирующих бит, подвергать их повторному кодированию не целесообразно. Схематично такой вариант кодирования изображен на рисунке 6. Можно применять дополнительное кодирование некоторых информационных бит, например через один, как показано на рисунке 7.

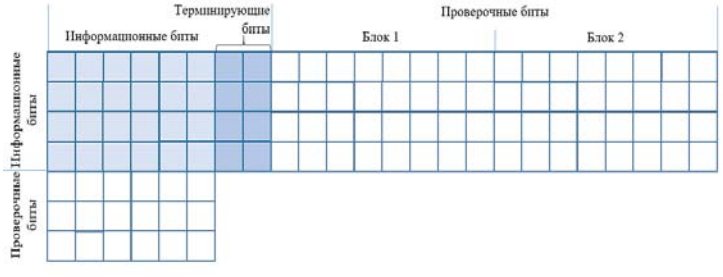


Рис. 6. Структурная схема реализации дополнительного кодирования информационных бит кодового слова турбокода за исключением терминирующих бит

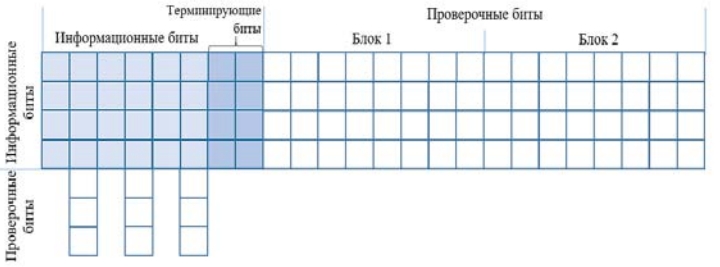


Рис. 7. Структурная схема реализации дополнительного кодирования части информационных бит

Применение терминирующей последовательности для защиты последнего информационного бита можно заменить его дополнительным кодированием. Аналогично можно защитить и тот информационный бит, который становится последним после перемежения. При этом можно наблюдать значительное снижение избыточности. Так, например, при длине информационной последовательности равной 6 бит и дополнительном кодировании кодом Хемминга можно перейти от структуры размером 96 бит к структуре из 78 бит, как показано на рисунке 8.

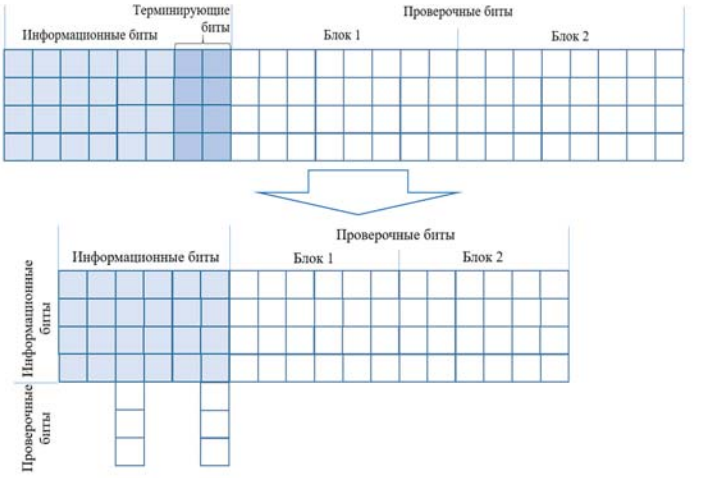


Рис. 8. Структурная схема реализации повышения защищенности последних информационных бит кодового слова турбокода

Код Хемминга обладает минимальной избыточностью и способен исправить только один ошибочный бит, для случая декодирования жесткого решения. Но даже применение простейшего кода Хемминга для дополнительного кодирования информационных бит турбокода может значительно повысить корректирующую способность. Продемонстрируем возможности по исправлению множественных ошибок в блоке информационных бит при их дополнительном кодировании кодом Хемминга.

Пусть в блоке информационных бит, закодированных по столбцам кодом Хемминга, одно информационное слово полностью искажено пакетом ошибок, а также имеются две независимые ошибки. Описанный блок схематично изображен на рисунке 9 слева, а результаты его декодирования декодером кода Хемминга схематично изображены на рисунке 9 справа. Ошибочные биты на рисунке отмечены х-образным перечеркиванием. В процессе декодирования декодер кода Хемминга без ошибок декодировал четыре кодовых слова, содержащих по одному ошибочному биту. Два кодовых слова содержали по два ошибочных бита и были декодированы с ошибками (по итогам ошибочного декодирования в кодовом слове появился еще один ошибочный бит). Рассмотрим четыре информационных слова турбокода, декодированные кодом Хемминга. Одно информационное слово не имеет ошибок, одно – содержит две ошибки, два – содержат одну ошибку. С учетом того, что декодер турбокода способен успешно исправлять все варианты расположения двух ошибок, информационное сообщение будет декодировано без ошибок. Можно сделать вывод об эффективности использования дополнительного кода Хемминга для противостояния множественным ошибок.

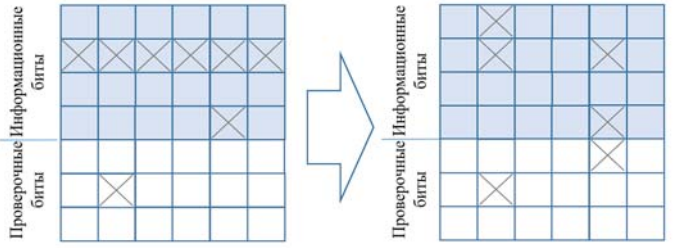


Рис. 9. Схематичная демонстрация результатов декодирования блока информации

Значения символов, подвергшихся дополнительному кодированию, являются более достоверными, чем значения остальных символов. При сверточном кодировании ближайшие кодовые биты связаны между собой, поэтому повышение достоверности отдельных бит может быть распространено и на соседние биты. Учесть большую степень достоверности дополнительно закодированных символов можно путем умножения их значений перед декодированием на повышающий коэффициент K_p>1, или путем умножения значений менее достоверных символов на понижающий коэффициент K_p<1. Предполагается, что при этом эффективность ТДК повысится. Подтвердим данное предположение, рассмотрев пример декодирования кодового слова турбокода при дополнительном кодировании его информационных бит и наличии ошибок.

В качестве кодового слова турбокода будем использовать ранее сформированное слово: -11-1-1 -1111 -1-1-11. При защите информационных бит дополнительным кодированием возникновение ошибок будет более вероятным в проверочных битах кодового слова. Пусть имеется три ошибки в блоках проверочных бит, тогда кодовое слово принимает вид: -11-1-1 -11-1-1 -1-1-1-1. Чтобы учесть большую достоверность информационных бит будем умножать проверочные биты на понижающий коэффициент K_p<1. В таблице 6 приведены значения мягкого выхода декодера турбокода, полученные в результате трех итераций декодирования кодового слова. Проведя анализ результатов декодирования, можно сделать вывод о повышении корректирующей способности кода при снижении значения K_p для случая наличия ошибок в проверочных битах кодового слова. Определение наилучшего значения K_p представляет интерес для дальнейшего исследования.

Таблица 6

Мягкий выход декодера турбокода при разных значениях K_p

№ бита	Значение переданного бита	Значения мягкого выхода декодера турбокода после трех итераций декодирования			
		K_p=1	K_p=0,8	K_p=0,7	K_p=0,6
1	–1	-41,3	-34,4	-30,9	-27,9
2	1	-10,4	-4,8	-1,9	0,9
3	–1	-67,0	-65,8	-65,2	-64,6
4	–1	-66,8	-65,8	-63,6	-62,7

С целью количественной оценки корректирующей способности ТДК с использованием программного тренажера произведем нахождение вероятности появления ошибки в декодированном информационном сообщении p_D при разном числе дополнительно закодированных информационных бит и различных значениях K_p. В отличии от рассмотренных примеров декодирования, в программном тренажере вместо умножения на понижающий коэффициент значений символов, не подвергшихся кодированию дополнительным кодом, было реализовано умножение на повышающий коэффициент K_p>1 значений символов, закодированных дополнительным кодом. В качестве дополнительного кода использован код Голея. Обозначим число информационных бит кодового слова турбокода, закодированных кодом Голея, как T_D.

Результаты исследования ТДК приведены в таблице 7. В таблице в скобках приведены значения K_p, для которых получены наилучшие значения корректирующей способности.

Таблица 7

Корректирующая способность ТДК

T_D	p_B=10^{-1}, p_V=3,1·10^{-2}		p_B=10^{-2}, p_V=1,9·10^{-5}		p_B=10^{-3}, p_V=2,3·10^{-9}	
	K_p =1	K_p >1	K_p =1	K_p >1	K_p =1	K_p >1
0	3,0·10^{-2}	–	1,3·10^{-4}	–	8,5·10^{-7}	–
1	2,32·10^{-2}	2,23·10^{-2} (K_p=1,2)	7,31·10^{-5}	4,18·10^{-5} (K_p=3)	3,9·10^{-7}	2,0·10^{-7} (K_p=3)
2	1,55·10^{-2}	1,49·10^{-2} (K_p=1,1)	1,34·10^{-5}	7,80·10^{-6} (K_p=2)	1,1·10^{-8}	2,35·10^{-9} (K_p=3)
3	1,20·10^{-2}	1,12·10^{-2} (K_p=1,1)	8,01·10^{-6}	2,9·10^{-6} (K_p=2)	4,0·10^{-9}	7,58·10^{-10} (K_p=3)

Проанализировав результаты, можно сделать вывод о повышении корректирующей способности турбокода при дополнительном кодировании. Для исследуемого турбокода при дополнительном кодировании лишь одного информационного бита зафиксировано снижение вероятности появления ошибочного бита в декодированном информационном сообщении: в 1,34 раза при p_B=10^{-1}, в 1,78 раза при p_B=10^{-3}, в 2,18 раза при p_B=10^{-3}. При дополнительном кодировании возможно изменение в широких пределах корректирующей способности, путем регулировки числа дополнительно закодированных бит.

Применение повышающего коэффициента к значениям символов кодового слова турбокода, закодированных дополнительным кодом, приводит к повышению корректирующей способности. Эффект от умножения на коэффициент увеличивается при снижении вероятности ошибочного бита в декодируемой последовательности. При T_D=3 за счет применения K_p снижение p_D составило: 1,07 раза при p_B=10^{-1}, 2,76 раза при p_B=10^{-3}, 5,27 раза при p_B=10^{-3}.

Выводы

Повысить эффективность турбокода можно путем выборочного дополнительного кодирования бит кодового слова. Таким образом, на основе турбокода образуется комбинированный код. С целью повышения кодовой скорости представляется целесообразным осуществлять дополнительное кодирование избирательно. В ходе исследований доказано, что наиболее целесообразно осуществлять кодирование дополнительным кодом информационных бит кодового слова. Изменяя число дополнительно закодированных бит, можно в широких пределах варьировать корректирующую способность кода. Предложен метод повышения эффективности такого комбинированного кода путем умножения значений дополнительно кодированных бит на повышающий коэффициент K_p>1 (или, что аналогично, умножению значений остальных бит на понижающий коэффициент K_p<1).

Определены значения K_p, обеспечивающие минимальную величину битовой ошибки в декодированном информационном сообщении. Выявлено, что рост корректирующей способности за счет применения коэффициентов увеличивается при снижении вероятности битовой ошибки в канале передачи данных. Результаты исследований могут быть применены при построении разнообразных телекоммуникационных систем, особенно использующих канал связи с высокой степенью нестационарности.

Литература

1. *Аджемов А.С., Санников В.Г.* Общая теория связи: учебник для вузов. М.: Горячая линия – Телеком, 2023. 624 с.

2. *Васильев К.К., Глушков В.А., Нестеренко А.Г.* Теория электрической связи. Вологда: Инфра-Инженерия, 2021. 468 с.

3. *Кудряшев Б.Д.* Основы теории кодирования: учебное пособие. СПб.: БХВ-Петербург, 2016. 400 с.

4. *Золоторев В.В.* Кодирование для цифровой связи и систем памяти: монография. М.: Горячая линия – Телеком, 2023. 192 с.

5. *Клюев Л.Л.* Теория электрической связи: учебник. М.: ИНФРА-М, 2019. 448 с.

6. *Варгузин В.А., Цикин И.А.* Методы повышения энергетической и спектральной эффективности цифровой радиосвязи. СПб.: БХВ-Петербург, 2013. 352 с.

7. *Голиков А.М.* Модуляция, кодирование и моделирование в телекоммуникационных системах. Теория и практика: учебное пособие для вузов. СПб.: Лань, 2022. 452 с.

8. *Скляр Б.* Цифровая связь. Теоретические основы и практическое применение. Пер с англ. М.: Вильямс, 2016. 1104 с.

9. *Морелос-Сарагоса Р.* Искусство помехоустойчивого кодирования. Методы, алгоритмы, применение. Пер с англ. М.: Техносфера, 2006. 319 с.

10. *Сидоренко А.А.* Программный тренажер исследования эффективности турбокода с изменяемыми коэффициентами значимости декодируемых символов // Свидетельство о государственной регистрации программы для ЭВМ. № 2023619022. 2023.

11. *Нуньес-Иглесиас Х., Уолт Ш., Дэшноу Х.* Элегантный SciPy. Перевод с английского – ДМК Пресс, 2018. 266 с.

BUILDING A COMBINED CODE BASED ON TURBO CODE WITH ADDITIONAL BIT ENCODING

ALEKSANDR A. SIDORENKO
Moscow, Russia, kotofeykin@mail.ru

KEYWORDS: *noise immunity, telecommunication systems, noise-resistant coding, convolutional codes, turbo codes.*

ABSTRACT

Subject: Methods for improving the efficiency of turbo codes. The purpose of the work is to evaluate the effectiveness of a combined code based on a turbo code with selective additional encoding of bits of its codeword. **Research methodology:** The effectiveness was evaluated based on the decoding results with different values of the probability of an erroneous bit appearing in the pB data transmission channel. The research used a variant of building a turbo code based on a systematic convolutional code. A technique is proposed to increase the efficiency of selective coding by multiplying the values of bits that have been and have not been subjected to additional coding by various coefficients. **Result:** With additional encoding of one bit of the turbo code codeword, a decrease in the probability of an erroneous bit appearing in the decoded information message was recorded p_D: at $p_B=10^{-1}$ by 1.34 times, at $p_B=10^{-2}$ by 1.78 times, at $p_B=10^{-3}$ by 2.18 times. Changing the number of bits of the turbo code codeword that have undergone additional coding allows you to vary the correcting ability of the code within a wide range. The proposed method of increasing the efficiency of the combined code has shown its effectiveness. When using this method for the case of additional encoding of three information bits of the codeword, the decrease in pD was: 1.07 times at $p_B=10^{-1}$, 2.76 times at $p_B=10^{-2}$, 5.27 times at $p_B=10^{-3}$. **Scope of the results:** The results obtained in the course of the study can be used in the design and construction of various data transmission and processing systems. The developed software simulator simulating the operation of the combined code can be used both for scientific and educational purposes. **Conclusions:** The publication discusses a method for improving the efficiency of turbo code by creating a combined code based on it. The increase in efficiency is expressed both in the growth of the correcting ability of the code, and in the possibility of flexible adaptation of the code parameters to changes in the state of the data transmission channel.

REFERENCES

1. Ajemov A.S. and Sannikov V.G. General theory of communication. Moscow: Hotline – Telecom Publ., 2023.

2. Vasiliev K.K., Glushkov V.A. and Nesterenko A.G. Theory of electrical communication. Vologda: Infra-Engineering Publ., 2021.

3. Kudryashev B.D. Fundamentals of coding theory. Saint-Petersburg: BHV-Petersburg Publ., 2016.

4. Zolotarev V.V. Coding for digital communication and memory systems, Moscow: Hotline – Telecom Publ., 2023.

5. Kliuev L.L. Theory of electrical communication, Moscow: Infra-M Publ., 2019.

6. Varguzin V.A. and Tsikin I.A. Methods for improving the energy and spectral efficiency of digital radio communications. Saint-Petersburg: BHV-Petersburg Publ., 2013.

7. Golikov A.M. Modulation, coding and modeling in telecommunication systems. Theory and practice. Saint-Petersburg: Lan Publ., 2022.

8. Sklyar B. Digital communication. Theoretical foundations and practical application, Prentice Hall, New Jersey, 2016.

9. Morelos-Zaragoza R. The art of noise-resistant coding. Methods, algorithms, application, Jon Wiley and Sons, Chichester, 2006.

10. Sidorenko A.A. Program simulator for studying the efficiency of a turbo code with variable coefficients of significance of decoded characters. The Certificate on Official Registration of the Computer Program in Russia, Moscow, Russia, No. 2023619022, 2023.

11. Nunez-Iglesias H., Walt S. and Dashnow H. Elegant SciPy, O'Reilly Media, Sebastopol, 2018.

INFORMATION ABOUT AUTHOR:
Aleksandr A. Sidorenko, Expert of the Department of Foreign Studies, International Academy of Informatization, Moscow, Russia

For citation: Sidorenko A.A. Building a combined code based on turbo code with additional bit encoding. *H&ES Reserch*. 2024. Vol. 16. No. 1. P. 14-20. doi: 10.36724/2409-5419-2024-16-1-14-20 (In Rus)

doi: 10.36724/2409-5419-2024-16-1-21-30

ОТ FOXPRO К POSTGRESQL: ОПТИМИЗАЦИЯ, ЭФФЕКТИВНОЕ УПРАВЛЕНИЕ ДАННЫМИ И ГЕНЕРАЦИЯ ОТЧЕТОВ

МАРКЕВИЧ
Даниил Владимирович [1]

ХОМОНЕНКО
Анатолий Дмитриевич [2]

ЕРМАКОВ
Сергей Геннадьевич [3]

АННОТАЦИЯ

Введение: в условиях быстро развивающихся технологических трендов и повышающейся потребности в эффективном управлении базами данных, статья представляет практическую реализацию перехода от устаревшей системы управления базами данных Foxpro к мощной и гибкой СУБД PostgreSQL. Основное внимание уделяется разработке функционала, который автоматически генерирует отчеты в формате PDF с использованием инструмента LibreOffice Base. **Цель исследования:** изучение и демонстрация эффективной переходной технологии от устаревшей СУБД Foxpro к современной СУБД PostgreSQL, а также разработка механизма генерации автоматических отчетов в формате PDF с использованием LibreOffice Base. **Методы и средства:** исследование и реализация переходной технологии выполняются с использованием СУБД PostgreSQL, а генерация отчетов в формате PDF осуществляется через инструмент LibreOffice Base. **Результаты:** результаты исследования включают в себя успешный переход от Foxpro к PostgreSQL, а также разработку функционала для автоматической генерации отчетов в формате PDF с помощью LibreOffice Base для повышения эффективности управления базами данных и снижения затрат на лицензирование программного обеспечения, обеспечивая надежность и безопасность данных. **Практическая значимость:** главной практической значимостью реализованной технологии является полная совместимость двух разных систем: базы данных, предоставляющей обширные средства, и платформы для аналитики. Представлены причины и преимущества интеграции этих систем, а также их применение при решении конкретных задач. Рассматриваются возможности оптимизации управления данными и увеличения качества принимаемых решений. **Обсуждение:** при решении задачи обеспечения перехода рассмотрен весь процесс на конкретном проекте с использованием СУБД PostgreSQL и инструмента LibreOffice Base для создания баз данных и генерации отчетов. Дальнейшие исследования могут быть направлены на улучшение аналитики больших объемов данных из нескольких баз данных и повышение качества управленческих решений, основанных на интеграции этих систем.

КЛЮЧЕВЫЕ СЛОВА: *система управления базами данных, Foxpro, PostgreSQL, генерация отчетов, LibreOffice Base, снижение затрат, безопасность данных.*

Сведения об авторах:

[1] магистрант кафедры "Информационные и вычислительные системы" Петербургского государственного университета путей сообщения Императора Александра I, г. Санкт-Петербург, Россия, dmarkevich811@mail.ru

[2] д.т.н., профессор, профессор кафедры "Информационные и вычислительные системы" Петербургского государственного университета путей сообщения Императора Александра I, г. Санкт-Петербург, Россия, khomon@mail.ru

[3] д.т.н., профессор, заведующий кафедрой "Информационные и вычислительные системы" Петербургского государственного университета путей сообщения Императора Александра I, г. Санкт-Петербург, Россия, ermakov@pgups.ru

Для цитирования: Маркевич Д.В., Хомоненко А.Д., Ермаков С.Г. От Foxpro к PostgreSQL: оптимизация, эффективное управление данными и генерация отчетов // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16. № 1. С. 21-30. doi: 10.36724/2409-5419-2024-16-1-21-30

Введение

Современные организации все больше осознают важность эффективного управления базами данных для обеспечения бесперебойной работы и принятия информированных решений. С этой целью, многие из них сталкиваются с необходимостью перехода от устаревших и ограниченных систем управления базами данных (СУБД) к более современным и мощным решениям. В рамках статьи рассматривается проект, ориентированный на переход от СУБД Foxpro к PostgreSQL и разработку функционала для автоматической генерации отчетов в формате PDF при помощи LibreOffice Base [1].

СУБД Foxpro, популярная в прошлом, стала неэффективной и устаревшей для современных бизнес-задач. Переход на PostgreSQL представляет собой логичное решение, учитывая его гибкость, масштабируемость и активное сообщество разработчиков. Однако, одним только переходом на новую СУБД дело не ограничивается.

Ключевой компонент проекта – разработка функционала генерации отчетов в формате PDF. Этот функционал значительно упрощает процесс анализа и предоставления данных, делая его более эффективным и прозрачным. Более того, использование LibreOffice Base позволяет существенно снизить затраты на лицензирование программного обеспечения, поскольку это бесплатное программное обеспечение с открытым исходным кодом. Это означает, что не нужно покупать лицензии для его использования, в отличие от некоторых других программных продуктов, таких как Microsoft Access.

В статье подробно рассмотрен весь процесс перехода на PostgreSQL и разработки функционала генерации отчетов в формате PDF, а также проанализированы различные аспекты этого решения.

СУБД Foxpro: проблемы и причины перехода на PostgreSQL

Системы управления базами данных играют важную роль в современном мире бизнеса и информационных технологий. Они служат основой для хранения, управления и обработки данных, необходимых для принятия бизнес-решений. Однако, с течением времени и с развитием технологий, многие старые СУБД становятся устаревшими и неспособными удовлетворить современные требования к управлению данными. В этом исследовании рассмотрена система управления базами данных Foxpro и проблемы, связанные с ее использованием, а также необходимость перехода на современную и мощную СУБД PostgreSQL.

СУБД Foxpro разработана компанией Fox Software в конце 1980-х годов и быстро приобрела популярность благодаря своей интеграции с системами Windows и легкости в использовании. Foxpro предоставляла возможности для создания десктопных приложений с использованием баз данных в формате .DBF.

Со временем, Foxpro начала проявлять свои ограничения:

1) Foxpro не предоставляет масштабируемости для обработки больших объемов данных. Это делает ее несостоятельной для предприятий, которые должны управлять огромными объемами информации.

2) Ограниченные возможности Foxpro в обработке сложных запросов и анализе данных делают ее неэффективной для современных бизнес-задач.

3) С развитием операционных систем, поддержка Foxpro существенно сократилась. Это означает, что пользователи сталкиваются с риском потери данных и недоступности технической помощи в случае проблем.

Современный бизнес невозможно представить без эффективной работы с данными. Базы данных являются фундаментом для хранения информации о клиентах, продуктах, операциях и многом другом. Бизнес-решения, основанные на данных, помогают предсказывать тренды, оптимизировать процессы и принимать информированные решения [2].

Актуальность перехода на современную СУБД, такую как PostgreSQL, заключается в следующем [3]:

1) Современные бизнес-задачи требуют обработки больших объемов данных и выполнения сложных аналитических запросов, а Foxpro не способна эффективно справляться с такими задачами.

2) PostgreSQL предлагает более высокий уровень надежности и безопасности по сравнению с Foxpro, что особенно важно при работе с критическими данными.

3) PostgreSQL предоставляет гибкость и масштабируемость для роста вашей организации и обработки больших объемов данных.

4) PostgreSQL является СУБД с открытым исходным кодом, она распространяется бесплатно и доступна для использования без необходимости покупки дорогостоящих лицензий. Это приводит к непосредственному снижению затрат на программное обеспечение, которые обычно связаны с коммерческими СУБД, такими как Microsoft SQL Server или Oracle Database [4].

5) PostgreSQL поддерживает стандарты SQL и обладает широкой совместимостью с другими инструментами и технологиями. Это означает, что не нужно приобретать дополнительные инструменты или платить за лицензии для работы с данными, так как PostgreSQL интегрируется легко с различными средствами анализа, отчетности и приложениями.

По сравнению с Foxpro СУБД PostgreSQL обладает выдающейся гибкостью и масштабируемостью. Она способна обслуживать как небольшие организации, так и крупные корпорации, обрабатывая огромные объемы данных без потери производительности. PostgreSQL предлагает богатый набор функций для работы с данными. Одним из ключевых преимуществ PostgreSQL является открытый исходный код. Пользователи могут использовать его бесплатно, что существенно снижает затраты на лицензирование программного обеспечения. Кроме того, PostgreSQL успешно используется во многих организациях по всему миру. Надежность и производительность доказаны многолетним опытом использования в критически важных приложениях [5].

Таким образом, в связи с выявленными проблемами, с которыми сталкиваются пользователи СУБД Foxpro, обоснована необходимость перехода на PostgreSQL. PostgreSQL предоставляет более гибкие, масштабируемые и безопасные решения для управления данными, что делает его оптимальным выбором для современных бизнес-задач. Регулярные обновления

и активное сообщество разработчиков делают PostgreSQL надежным и актуальным решением для организации.

Что касается LibreOffice Base, то это бесплатная и мощная программа для управления базами данных, которая входит в состав офисного пакета LibreOffice. Она предоставляет удобный пользовательский интерфейс и широкий набор функций, позволяющих создавать, управлять и анализировать базы данных. Это свободно распространяемая СУБД, которая входит в состав офисного пакета LibreOffice. Она также доступна бесплатно и не требует покупки лицензии. Используя LibreOffice Base для генерации отчетов в формате PDF, можно избежать затрат на коммерческие решения для создания отчетов или лицензирование других СУБД, которые могли бы потребоваться для этой цели. Поскольку PostgreSQL и LibreOffice Base являются свободно доступными и широко используемыми инструментами, обучение персонала и обслуживание базы данных становятся более доступными и дешевыми. Не нужно тратить средства на обучение персонала работе с дорогостоящими коммерческими продуктами [5].

LibreOffice Base позволяет создавать новые базы данных с нуля или импортировать существующие данные из различных форматов, таких как CSV, Microsoft Access, dBase и другие. Кроме того, интерфейс LibreOffice Base интуитивно понятен и удобен в использовании. Он содержит разделы для таблиц, запросов, отчетов и форм, что облегчает работу с данными [6].

LibreOffice Base поддерживает различные базы данных, включая PostgreSQL, MySQL, SQLite, Microsoft Access и другие. Это дает возможность работать с разными СУБД и переносить данные между ними. LibreOffice Base поддерживает создание макросов для автоматизации повторяющихся задач. Макросы могут выполняться автоматически при определенных событиях или по запросу пользователя.

В результате, с помощью программы возможно печатать данные и отчеты из LibreOffice Base или экспортировать их в различные форматы, такие как PDF, CSV, Microsoft Excel и другие [7].

Для реализации перехода на новую технологию сперва необходимо изучить принцип работы на примере, в котором в качестве исходных данных будут рассматриваться три таблицы в формате DBF, которые необходимо импортировать в базу данных PostgreSQL, а затем создать файл ODB в инструменте разработки и администрирования баз данных с открытым исходным кодом и настроить автоматическую генерацию отчетов в формате PDF по динамическим данным таблиц.

Первый шаг алгоритма – импорт данных из исходных файлов DBF в базу данных PostgreSQL. Для этого можно использовать различные методы, такие как использование командной строки, языков программирования или специализированных инструментов импорта данных. Для прохождения этого шага выбран специализированный инструмент импорта данных: команда COPY. Она предоставила возможность реализации импорта данных со всеми данными в таблице.

Второй шаг – создание файла базы данных с помощью инструмента разработки и администрирования баз данных с открытым исходным кодом, в качестве которого была выбрана программа LibreOffice Base. Это офисное приложение предоставляет графический интерфейс для создания таблиц, запросов и отчетов. При создании нового файла базы данных (с форматом ODB) его можно связать с существующей базой данных в PostgreSQL для открытия всех файлов внутри неё. Затем производится настройка запросов и отчетов для динамической генерации данных. Файл ODB представляет собой базу данных LibreOffice, в которой можно создавать таблицы, формы, запросы и отчеты. Это предоставляет удобный способ управления данными и настройки отчетов.

Третий шаг – настройка автоматической генерации отчетов в формате PDF по динамическим данным таблиц. Это позволит создавать отчеты на основе актуальных данных в таблицах. Шаблон, на который необходимо ориентироваться, приведен на рисунке 1.

Лист 1
08/06/23

УЧЕТНЫЕ КАРТОЧКИ АВТОМОБИЛЕЙ

Приложение к акту списания топлива за май 2023 г.
по автомобилю ФОЛЬКСВАГЕН МУЛЬТИВЕНГос.No X215 Гар.No 19

Пачка №	Дата выдачи	№ путевого листа	Таб.№ водителя (1)	Фамилия водителя	Таб.№ водителя (2)	Код топлива	Остаток при выезде	Выдано топлива (1)	Код топлива	Выдано топлива (2)	Куплено топлива	Остаток при возврате	Расход по норме	Расход фактический	В т.ч для спец. обор.	Экономия / пережег (+/-)	Спидометр при выезде	Спидометр при возвр.	Пробег
57	15/05/23	4757	389	ШУРИНОВ А.В.		95	65					49	17.72	16.00		1.72	115300	115386	86
57	16/05/23	4795	389	ШУРИНОВ А.В.		95	49					40	9.48	9.00		0.48	115386	115432	46
57	17/05/23	4839	389	ШУРИНОВ А.В.		95	40	40.42				72	9.48	8.42		1.06	115432	115478	46
57	18/05/23	4871	389	ШУРИНОВ А.В.		95	72					63	9.89	9.00		0.89	115478	115526	48
57	19/05/23	4911	389	ШУРИНОВ А.В.		95	63	70.49				80	64.61	53.49		11.12	115526	115884	358
Итого:								110.91		0.00	0.00		111.18	95.91		15.27			584

Главный специалист _____ /Нескоромная И.В./

Рис. 1. Шаблон необходимого вывода данных

LibreOffice Base, благодаря своим возможностям, является отличным выбором для многих пользователей, нуждающихся в эффективном инструменте для работы с данными, поскольку здесь можно создать запросы, которые извлекают необходимые данные из таблиц. Затем можно настроить отчеты, указав макет, структуру и форматирование. После этого осуществляется настройка автоматической генерации отчетов на основе актуальных данных через встроенные функции LibreOffice Base. Оценка эффективности генерации отчетов зависит от сложности запросов, объема данных и требований к времени выполнения [8].

Создание базы данных в PostgreSQL, подключение к LibreOffice Base и импорт данных

Далее необходимо более развернуто рассмотреть процесс создания базы данных PostgreSQL и соединения с ней через LibreOffice Base – приложения для работы с базами данных в открытом доступе. После установки PostgreSQL нужно создать базу данных, в которой будут храниться данные. Создание производится с помощью следующей команды на языке запросов SQL:

CREATE DATABASE PortDatabase;

Созданная база данных в программном продукте для администрирования и разработки баз данных pgAdmin4 выглядит следующим образом (рис. 2):

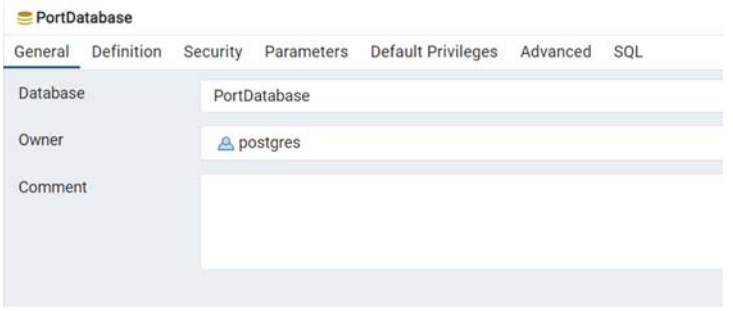


Рис. 2. Окно созданной базы данных

Теперь, когда база данных PostgreSQL создана, необходимо создание соединения LibreOffice Base с базой данных. Для этого при открытии программы требуется открыть «Подключение к существующей базе данных», выбрать PostgreSQL и заполнить следующие поля:

- сервер;
- пользователь;
- пароль;
- база данных.

После нажатия «Далее» и подтверждения параметров подключения LibreOffice Base установит соединение с базой данных PostgreSQL. Само соединение представлено на рисунке 3.

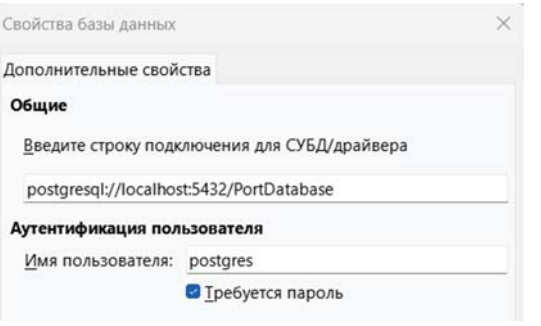


Рис. 3. Свойства базы данных

Теперь, когда установлено соединение с базой данных в LibreOffice Base, можно создать формы и отчеты для работы с данными. LibreOffice Base предоставляет мощные инструменты для создания пользовательского интерфейса и отображения результатов запросов.

После создания базы данных PostgreSQL и соединения с ней через LibreOffice Base необходимо создавать и импортировать таблицы в базу данных. Исходные материалы предоставлены в формате DBF и для их импорта проведена их конвертация в CSV-файл, а затем и импорт в саму базу данных. Исходные таблицы выглядят следующим образом (рис. 4-6):

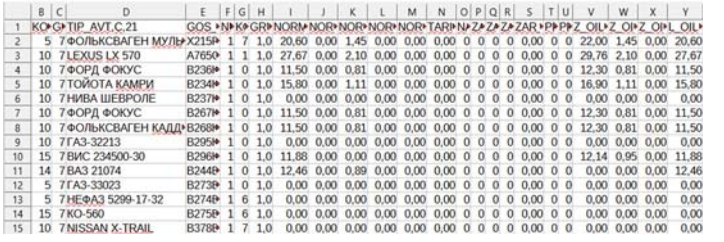


Рис. 4. Исходная таблица №1

Рис. 5. Исходная таблица №2

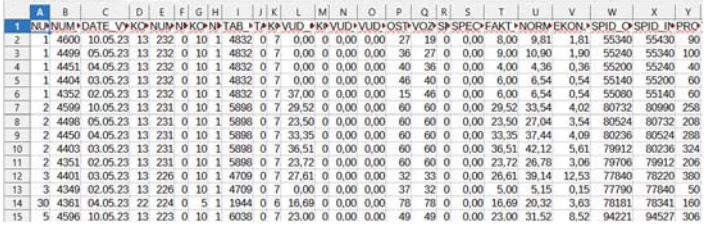


Рис. 6. Исходная таблица №3

Перед тем как начать импорт данных, создается новая таблица, в которую будут импортироваться данные. Это можно сделать также с помощью программного продукта pgAdmin4. Результат создания трех таблиц отражен на рисунках 7, 8 и 9.

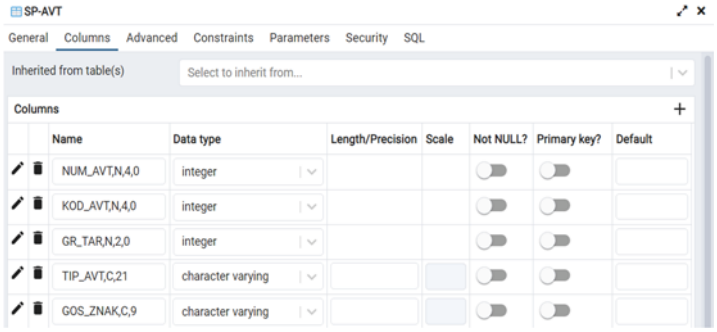


Рис. 7. Окно созданной в БД таблицы №1

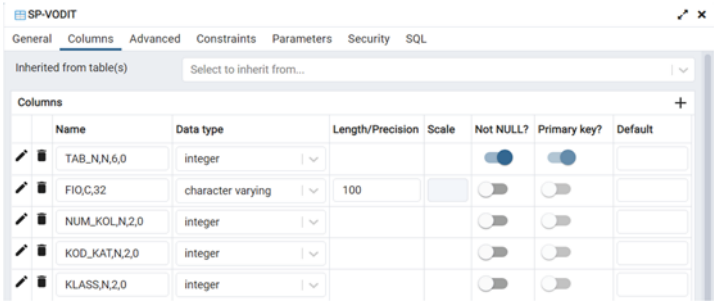


Рис. 8. Окно созданной в БД таблицы №2

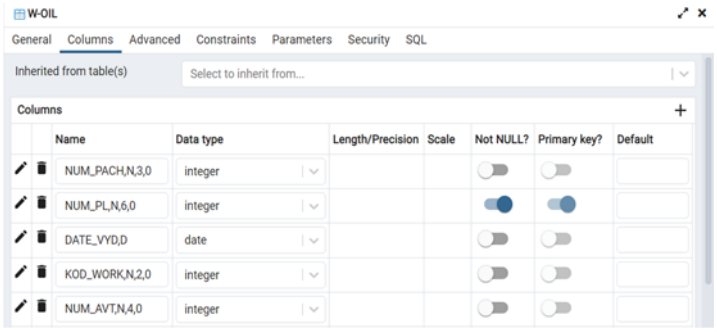


Рис. 9. Окно созданной в БД таблицы №3

Следующим шагом выполняется импорт данных из исходных таблиц. Одним из распространенных способов импорта данных в PostgreSQL является использование CSV-файлов. CSV – это текстовый формат, в котором данные разделены запятыми или другими символами-разделителями. В PostgreSQL существует удобный инструмент COPY, который позволяет импортировать данные из CSV-файла в таблицу [9].

Для импорта данных из файла в таблицу задается команда в следующим виде:

copy "SP-AVT" from 'C:\SP_AVTss.csv' DELIMITER ';' csv header;

В ней указывается название таблицы, в которую импортируются данные. Затем указывается путь к CSV-файлу, символ-разделитель (';') и опция CSV HEADER, которая указывает, что первая строка в файле содержит заголовки столбцов. Остальные таблицы импортируются таким же образом.

В результате проделанных действий данные таблиц были перенесены в базу данных, что отражено на рисунках 10, 11 и 12.

	NUM_AVT,N,4 integer	KOD_AVT,N,4 integer	GR_TAR,N,2,0 integer	TIP_AVT,C,21 character var	GOS_ZNAK,C character var	NUM_KOL,N,2 integer	KOD_OIL,N,2,0 integer	GRUZOPOD,N double precis	NORMA_OIL,N double precis	NOR_OIL_SD double precis	NOR_OIL_SQ double precis
1	19	5	7	ФОЛЬКС...	X215PP1...	1	7	1	20.6	0	1.45
2	132	10	7	LEXUS LX...	A765CY1...	1	1	1	27.67	0	2.1
3	87	10	7	ФОРД Ф...	B236KE1...	1	0	1	11.5	0	0.81
4	85	10	7	ТОЙОТА ...	B234KE1...	1	0	1	15.8	0	1.11
5	88	10	7	НИВА ШЕ...	B237KE1...	1	0	1	0	0	0
6	90	10	7	ФОРД Ф...	B267KE1...	1	0	1	11.5	0	0.81
7	91	10	7	ФОЛЬКС...	B268KE1...	1	0	1	11.5	0	0.81
8	99	10	7	ГАЗ-32213	B295KE1...	1	0	1	0	0	0
9	100	15	7	ВИС 2345...	B296KE1...	1	0	1	11.88	0	0
10	109	14	7	ВАЗ 21074	B244EO1...	1	0	1	12.46	0	0.89

Рис. 10. Импортированные в БД данные из таблицы №1

	TAB_N,N,6,0 [PK] integer	FIO,C,32 character vary	NUM_KOL,N,2 integer	KOD_KAT,N,2, integer	KLASS,N,2,0 integer	KF_TAR,N,3,0 integer	N_UD,C,10 character var	MED_S,D date
1	1378	СОЛОВЬЕ...	10	10	25	0	78ME161...	2016-10-09
2	1673	ЛУКЬЯНО...	9	10	0	0	78CT104...	[null]
3	1908	МАХОВ В...	10	10	25	0	78BT001...	2014-03-23
4	1909	КИРИЛЛ...	10	10	25	0	78TB040...	2017-09-07
5	1910	ДУДЧЕНК...	9	10	10	0	78PM011...	[null]
6	1911	КУТЕПОВ ...	1	1	25	0	78XE018...	2016-10-27
7	1916	МАКАРИ...	10	10	25	0	78CT000...	2017-04-02
8	1919	ШИНКАР...	1	1	25	0	78TB013...	2018-03-28

Рис. 11. Импортированные в БД данные из таблицы №2

	NUM_PACH,N integer	NUM_PL,N,6,0 [PK] integer	DATE_VYD,D date	KOD_WORK,N integer	NUM_AVT,N,4 integer	NUM_PRIC,N integer	KOD_AVT,N,4 integer	NUM_KOL,N,2 integer	TAB_NUM1,N integer	TAB_NUM2,N integer	KOD_OIL,N,2,0 integer	VUD_OIL,N,7, double precis
1	27	4300	2023-05-01	22	234	0	5	1	1960	0	6	32.43
2	42	4301	2023-05-01	22	177	0	5	1	4490	0	6	33.66
3	46	4302	2023-05-01	22	152	0	5	1	6387	0	6	15
4	46	4303	2023-05-01	22	152	0	5	1	5682	0	6	11.6
5	40	4304	2023-05-01	22	179	0	5	1	4062	0	6	10.03
6	40	4305	2023-05-01	22	179	0	5	1	5700	0	6	12.28
7	35	4306	2023-05-01	22	209	0	5	1	5122	0	3	8.11
8	29	4307	2023-05-01	22	225	0	5	1	6258	0	3	23.36

Рис. 12. Импортированные в БД данные из таблицы №3

Таким образом, реализован процесс создания таблиц в PostgreSQL и импорта данных в них из CSV-файлов.

Работа с данными и создание отчетов в LibreOffice Base

После импорта данных в таблицы начинается работа в LibreOffice Base. Перед созданием отчета, необходимо создать запрос для объединения, группировки и сортировки данных из таблиц и получения необходимых результатов [10].

LibreOffice Base предоставляет удобный инструмент для создания запросов к базе данных. Этот инструмент позволяет выбирать таблицы, указывать условия и определять сортировку данных. Для начала требуется открыть LibreOffice Base и выбрать раздел «Запросы», а после перейти к созданию запросов в режиме SQL (рис. 13) [11].

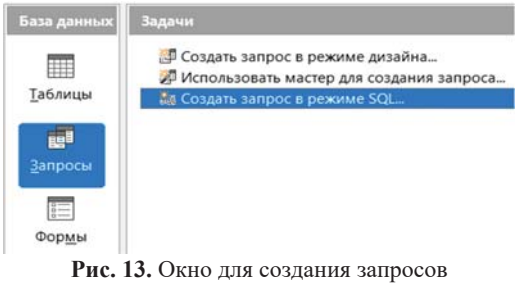


Рис. 13. Окно для создания запросов

После этого вводится запрос на языке SQL, в котором будут отражаться данные в соответствии с нужной группировкой и сортировкой:

```
SELECT
"NUM_AVT,N,4,0", "NUM_PACH,N,3,0", "DATE_VYD,D",
"NUM_PL,N,6,0", "TAB_NUM1,N,6,0", "FIO,C,32",
"TAB_NUM2,N,6,0", "KOD_OIL,N,2,0", "OST_OIL,N,4,0",
"VUD_OIL,N,7,2", "KOD_O_1,N,2,0", "VUD_O_1,N,6,2",
"VUD_OILR,N,6,2", "VOZ_OIL,N,4,0", "NORMA_OIL,N,7,2",
"FAKT_OIL,N,7,2", "SPEC_OIL,N,6,2", "EKON,N,6,2",
"SPID_OUT,N,6,0", "SPID_IN,N,6,0", "PROBEG,N,5,0",
"TIP_AVT,C,21", "GOS_ZNAK,C,9", NULL AS
"SUM_VUD_OIL,N,7,2", NULL AS "SUM_VUD_O_1,N,6,2", NULL AS
"SUM_VUD_OILR,N,6,2", NULL AS "SUM_NORMA_OIL,N,7,2",
NULL AS "SUM_FAKT_OIL,N,7,2", NULL AS "SUM_EKON,N,6,2",
NULL AS "SUM_PROBEG,N,5,0" FROM "result"
WHERE "NUM_AVT,N,4,0" IS NOT NULL
UNION ALL
SELECT "NUM_AVT,N,4,0", NULL, NULL, NULL, NULL, NULL,
NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL,
NULL, NULL, NULL, NULL, NULL, NULL, NULL, NULL,
SUM("VUD_OIL,N,7,2"), SUM("VUD_O_1,N,6,2"),
SUM("VUD_OILR,N,6,2"), SUM("NORMA_OIL,N,7,2"),
SUM("FAKT_OIL,N,7,2"), SUM("EKON,N,6,2"),
SUM("PROBEG,N,5,0") FROM "result"
WHERE "NUM_AVT,N,4,0" IS NOT NULL
GROUP BY "NUM_AVT,N,4,0"
ORDER BY "NUM_AVT,N,4,0", "DATE_VYD,D"
```

Как видно по коду, принято решение предварительно создать отдельную таблицу «Result» перед работой с запросами для того, чтобы в ней отражались необходимые колонки трех таблиц. Кроме того, в результат запроса добавляется отдельная строка, которая высчитывает суммарное значение необходимого столбца.

После создания запросов можно выполнить их, нажав на кнопку «Выполнить запрос» в LibreOffice Base, после чего результаты запроса будут отображены в новом окне (рис. 14).

Безопасность и целостность данных – это критически важные аспекты при работе с базой данных. В PostgreSQL права доступа управляются на уровне базы данных и таблиц. Для обеспечения безопасности данных следует предоставлять минимально необходимые права пользователям, чтобы они могли выполнять только необходимые операции [12].

В рамках рассматриваемого прмера безопасность данных, хранящихся в БД PostgreSQL, осуществляется благодаря необходимости в авторизации при подключении. Выглядит это следующим образом (рис. 15).

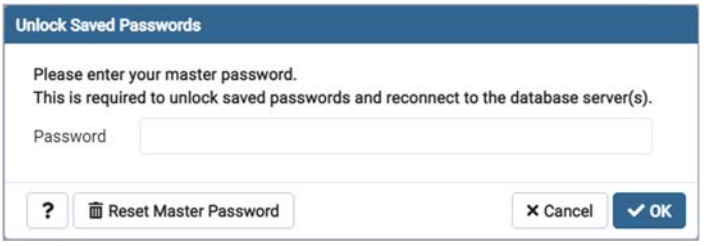


Рис. 15. Окно авторизации при подключении к базе данных в PostgreSQL

В LibreOffice Base также нужно указать учетные данные пользователя при создании соединения с базой данных. Перед авторизацией необходимо убедиться, что у пользователя есть только те права, которые ему требуются для выполнения своих задач [13]. Авторизация пользователей отражена на рисунке 16.

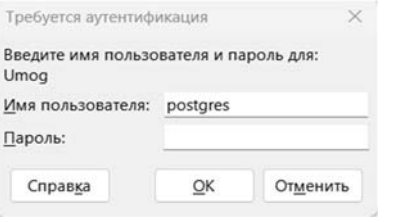


Рис. 16. Окно авторизации при подключении к базе данных в PostgreSQL

NUM_...	NUM_...	DATE_V...	NUM...	TAB_...	FIO,C,32	TA...	KO...	O...	VUD...				V...	NOR...	FAKT...	SPE...	EKO...	SPID_...	SPID_I...	PR...	TIP_AVT,C,21	GOS_...	SUM_...	SUM_...	SUM...	SUM_...	SUM...	SUM...	SU...
19	57	15.05.23	4757	3897	ШУРИНОВ АН/	0	7	65	0	0	0	0	49	17,72	16	0	1,72	115300	115386	86	ФОЛЬКСВАГЕН Т	X215PP							
19	57	16.05.23	4795	3897	ШУРИНОВ АН/	0	7	49	0	0	0	0	40	9,48	9	0	0,48	115386	115432	46	ФОЛЬКСВАГЕН Т	X215PP							
19	57	17.05.23	4839	3897	ШУРИНОВ АН/	0	7	40	40,42	0	0	0	72	9,48	8,42	0	1,06	115432	115478	46	ФОЛЬКСВАГЕН Т	X215PP							
19	57	18.05.23	4871	3897	ШУРИНОВ АН/	0	7	72	0	0	0	0	63	9,89	9	0	0,89	115478	115526	48	ФОЛЬКСВАГЕН Т	X215PP							
19	57	19.05.23	4911	3897	ШУРИНОВ АН/	0	7	63	70,49	0	0	0	80	64,61	53,49	0	11,12	115526	115884	358	ФОЛЬКСВАГЕН Т	X215PP							
19																							110,91	0	0	111,18	95,91	15,27	584
129	61	22.05.23	4973	4132	КОЧАРЛИНСКІ	0	7	63	0	0	0	0	62	1,26	1	0	0,26	247677	247685	8	ТОЙОТА КАМРИ	K553OT							
129	61	24.05.23	5064	3319	ЕФРЕМОВ ЕВГЕ	0	7	62	0	0	0	0	60	7,11	2	1,11	5,11	247685	247723	38	ТОЙОТА КАМРИ	K553OT							
129	61	25.05.23	5112	3319	ЕФРЕМОВ ЕВГЕ	0	7	60	0	0	0	0	58	6,8	2	1,11	4,8	247723	247759	36	ТОЙОТА КАМРИ	K553OT							
129																							0	0	0	15,17	5	10,17	82
130	25	02.05.23	4321	4132	КОЧАРЛИНСКІ	0	7	58	0	0	0	0	54	5,24	4	0	1,24	218713	218753	40	ТОЙОТА КАМРИ	B675HH							
130	25	03.05.23	4375	4132	КОЧАРЛИНСКІ	0	7	54	21,5	0	0	0	70	10,48	5,5	0	4,98	218753	218833	80	ТОЙОТА КАМРИ	B675HH							
130	25	04.05.23	4423	4132	КОЧАРЛИНСКІ	0	7	70	0	0	0	0	61	9,04	9	0	0,04	218833	218902	69	ТОЙОТА КАМРИ	B675HH							
130	25	05.05.23	4472	4132	КОЧАРЛИНСКІ	0	7	61	0	0	0	0	57	5,76	4	0	1,76	218902	218946	44	ТОЙОТА КАМРИ	B675HH							
130	25	10.05.23	4575	4132	КОЧАРЛИНСКІ	0	7	57	0	0	0	0	53	4,32	4	0	0,32	218946	218979	33	ТОЙОТА КАМРИ	B675HH							
130	25	11.05.23	4612	4132	КОЧАРЛИНСКІ	0	7	53	26,67	0	0	0	70	16,38	9,67	0	6,71	218979	219104	125	ТОЙОТА КАМРИ	B675HH							
130	25	12.05.23	4655	4132	КОЧАРЛИНСКІ	0	7	70	0	0	0	0	66	5,37	4	0	1,37	219104	219145	41	ТОЙОТА КАМРИ	B675HH							
130	25	16.05.23	4776	4132	КОЧАРЛИНСКІ	0	7	66	0	0	0	0	63	3,93	3	0	0,93	219145	219175	30	ТОЙОТА КАМРИ	B675HH							
130	25	17.05.23	4822	4132	КОЧАРЛИНСКІ	0	7	63	0	0	0	0	57	5,64	6	1,84	-0,36	219175	219204	29	ТОЙОТА КАМРИ	B675HH							
130	25	18.05.23	4854	4132	КОЧАРЛИНСКІ	0	7	57	0	0	0	0	52	5,5	5	0	0,5	219204	219246	42	ТОЙОТА КАМРИ	B675HH							
130	25	19.05.23	4894	4132	КОЧАРЛИНСКІ	0	7	52	0	0	0	0	50	1,97	2	0	-0,03	219246	219261	15	ТОЙОТА КАМРИ	B675HH							
130	25	23.05.23	5017	4132	КОЧАРЛИНСКІ	0	7	50	0	0	0	0	45	5,24	5	0	0,24	219261	219301	40	ТОЙОТА КАМРИ	B675HH							
130	25	24.05.23	5060	4132	КОЧАРЛИНСКІ	0	7	45	0	0	0	0	41	5,9	4	0,92	1,9	219301	219339	38	ТОЙОТА КАМРИ	B675HH							

Рис. 14. Результат выполнения запроса

Кроме аутентификации осуществляется шифрование данных, которое помогает обеспечить безопасность данных при хранении и передаче. PostgreSQL поддерживает шифрование на уровне хранилища данных с помощью различных методов шифрования. Можно выбрать подходящий метод шифрования, чтобы защитить данные от несанкционированного доступа.

Приложения, такие как LibreOffice Base, могут выполнять проверку данных на стороне клиента, но важно также проводить проверку данных на стороне сервера. PostgreSQL предоставляет механизмы для создания правил проверки данных, триггеров и хранимых процедур, которые обеспечивают целостность данных на уровне базы данных [14].

Регулярное создание резервных копий данных является критически важной процедурой для обеспечения безопасности. В случае сбоя или несанкционированного доступа к данным, резервные копии позволяют восстановить данные до последнего рабочего состояния [15].

PostgreSQL предоставляет инструменты для создания резервных копий баз данных и восстановления данных из них. Можно использовать pg_dump для создания текстовых резервных копий или pg_dumpall для создания резервных копий всех баз данных в PostgreSQL-кластере.

Обеспечение безопасности и целостности данных – это важная задача при работе с базой данных в LibreOffice Base и PostgreSQL. Правильное управление правами доступа, использование шифрования данных, проверка данных на стороне сервера, резервное копирование и регулярное обновление программного обеспечения помогают защитить данные от угроз и обеспечить их сохранность. Следование этим рекомендациям поможет обеспечить безопасность и надежность вашей базы данных и обеспечить спокойствие при работе с важными данными [16].

Форматирование и генерация PDF – также важные задачи при работе с данными и документами в LibreOffice Base и PostgreSQL. В этом разделе рассмотрена возможности форматирования данных и создания PDF-файлов с помощью LibreOffice Base и PostgreSQL [17].

LibreOffice Base предоставляет широкие возможности для форматирования данных в отчетах и формах. Так, можно применять стили, устанавливать форматы дат, чисел и текстов, а также добавлять изображения и графику для создания профессионально выглядящих отчетов и документов.

В формах и отчетах можно настроить отображение чисел с заданным количеством знаков после запятой, установить формат даты и времени, а также использовать различные форматы чисел (например, валюты). Также LibreOffice Base предоставляет возможность создания и применения стилей, которые позволяют единообразно форматировать различные элементы документов, такие как заголовки, текст, таблицы и т.д. Можно добавлять изображения и графику в отчеты и формы, чтобы улучшить их визуальное представление и сделать их более информативными [18].

LibreOffice Base позволяет генерировать PDF-файлы из отчетов и форм. Это особенно удобно, если необходимо создавать документы для распространения или печати. Создавать отчеты можно в LibreOffice Base и затем сохранять их в формате PDF. Для этого нужно открыть отчет, выбрать «Файл» > «Экспорт в PDF» и сохранить документ в нужной директории. Окно при экспорте выглядит следующим образом (рис. 17):

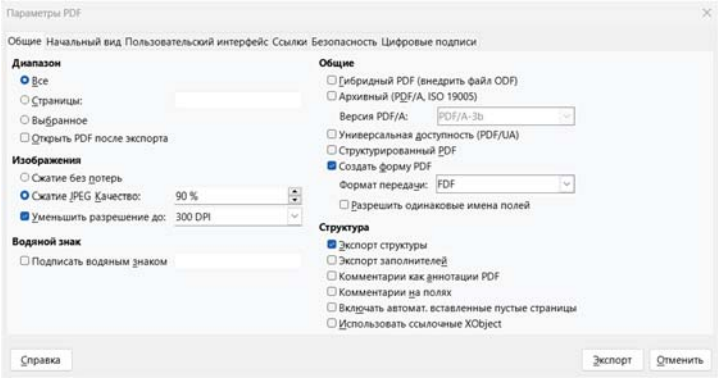


Рис. 17. Окно при экспорте результатов в формате PDF

Форматирование данных и генерация PDF-файлов – это мощные инструменты для работы с данными и документами в LibreOffice Base и PostgreSQL. Можно настроить отображение данных, создавать профессионально выглядящие отчеты и документы, а также генерировать PDF-файлы для распространения и печати. Эти возможности значительно облегчают работу с данными и улучшают визуальное представление результатов работы с базой данных [19].

Создание отчетов является важной задачей при работе с базами данных в LibreOffice Base. Отчеты представляют структурированную информацию, которая помогает анализировать данные, принимать решения и предоставлять информацию пользователям. Так что необходимо рассмотреть процесс создания отчета в режиме дизайна с использованием динамических полей, которые позволяют генерировать отчеты с автоматическим обновлением на основе данных из таблицы [20]. После определения способа вывода данных далее осуществляется формирование шаблона отчета, который должен в нужном формате выводить табличные данные.

Для создания отчета с динамическими полями сперва требуется перейти к разделу «Отчеты», а после создать новый отчет через опцию «Создать ответ в режиме дизайна» из меню или на панели инструментов.

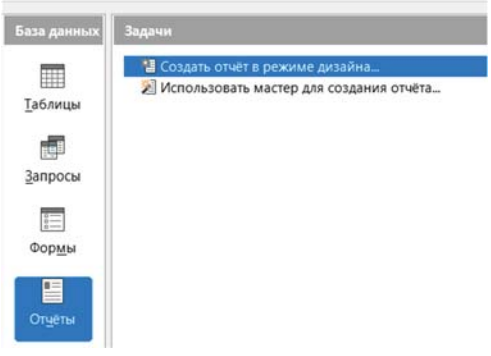


Рис. 18. Окно для создания отчетов

В режиме дизайна добавляются необходимые поля из таблицы на отчет. Поля можно перетаскивать на отчет из панели

браузера базы данных. Для создания динамических полей, требуется выделить поле на отчете и открыть его свойства.

Рис. 19. Раздел для указания источника данных

В разделе «Данные» указать в качестве типа содержимого указать необходимый SQL-запрос.

Далее следует дополнение отчета другими элементами, такими как заголовки, подвалы, разделители и графики, чтобы улучшить его визуальное представление и информативность. После завершения дизайна, будет готовый отчет с динамическими полями, который автоматически обновляется на основе данных из таблицы.

Таким образом, шаблон выглядит следующим образом (рис. 20).

В результате сгенерированный отчет представлен на рисунке 21.

Таким образом, создание отчета в режиме дизайна с динамическими полями позволяет создавать гибкие и динамические отчеты, которые обновляются на основе данных из таблицы.

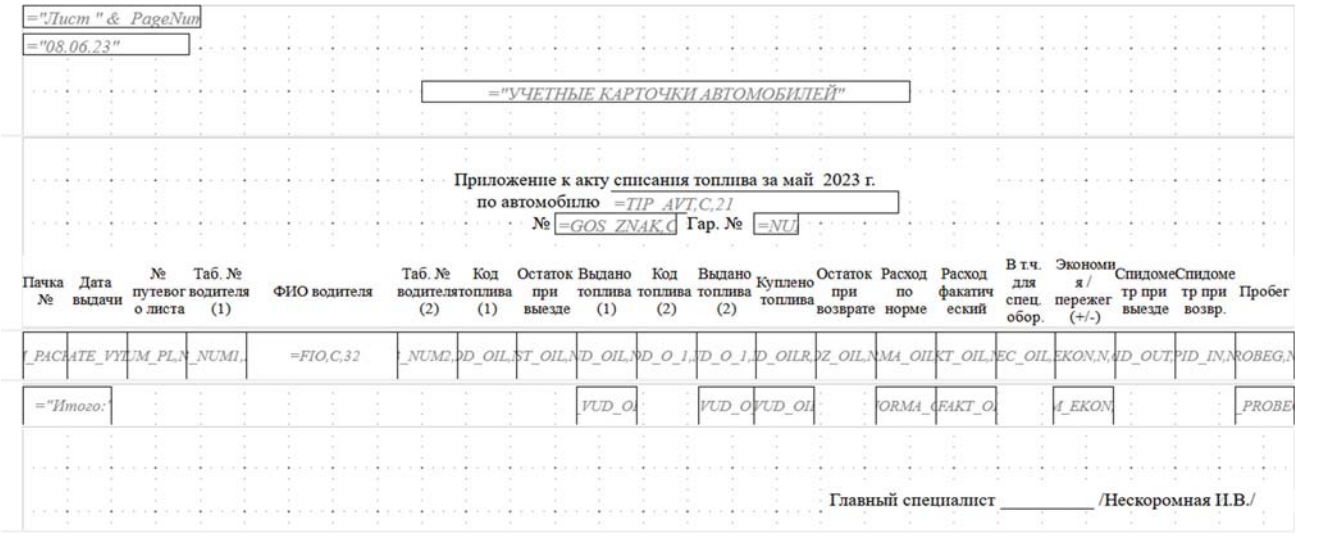


Рис. 20. Сформированный шаблон отчета для необходимого формата вывода данных

Лист 1
08.06.23

УЧЕТНЫЕ КАРТОЧКИ АВТОМОБИЛЕЙ

Приложение к акту списания топлива за май 2023 г.
по автомобилю ФОЛЬКСВАГЕН МУЛЬТИВЕН
№ Х215Р Гар. № 19

Пачка №	Дата выдачи	№ путевого листа	Таб. № водителя (1)	ФИО водителя	Таб. № водителя (2)	Код топлива (1)	Остаток при выезде	Выдано топлива (1)	Код топлива (2)	Выдано топлива (2)	Куплено топлива	Остаток при возврате	Расход по норме	Расход фактический	В т.ч. для спец. обор.	Экономия / пережег (+/-)	Спидометр при выезде	Спидометр при возвр.	Пробег
57	15.05.23	4757	389	ШУРИНОВ АНДРЕЙ ВАЛЕРЬЕВИЧ		7	65					49	17,72	16		1,72	115300	115386	86
57	16.05.23	4795	389	ШУРИНОВ АНДРЕЙ ВАЛЕРЬЕВИЧ		7	49					40	9,48	9		0,48	115386	115432	46
57	17.05.23	4839	389	ШУРИНОВ АНДРЕЙ ВАЛЕРЬЕВИЧ		7	40	40,42				72	9,48	8,42		1,06	115432	115478	46
57	18.05.23	4871	389	ШУРИНОВ АНДРЕЙ ВАЛЕРЬЕВИЧ		7	72					63	9,89	9		0,89	115478	115526	48
57	19.05.23	4911	389	ШУРИНОВ АНДРЕЙ ВАЛЕРЬЕВИЧ		7	63	70,49				80	64,61	53,49		11,12	115526	115884	358
Итого:								110,91					111,18	95,91		15,27			584

Главный специалист _____ /Нескоромная И.В./

Рис. 21. Результат вывода данных в необходимом формате

Это значительно упрощает процесс форматирования и генерации отчетов в LibreOffice Base, что важно для анализа данных и предоставления информации пользователям. Использование динамических полей помогает создать информативные и актуальные отчеты, способствуя принятию решений и оптимизации работы с базами данных.

Заключение

В ходе проделанной работы проведено исследование возможностей взаимодействия с базой данных PostgreSQL и разработан соответствующий функционал для повышения эффективности управления СУБД. Кроме того, создан функционал, позволяющий принимать динамические данные из базы данных и выводить результаты в формате PDF.

После изучения различных технологий взаимодействия с базой данных PostgreSQL выбраны наиболее эффективные методы выполнения запросов. Разработанное решение предоставляет пользователям возможность легко получать актуальные данные из базы данных.

Для повышения качества взаимодействия с базами данных применена оптимизация используемых запросов и организован механизм безопасности данных. Тестирование показало, что функционал работает корректно и соответствует заявленным требованиям.

Особо отметим возможность форматирования и сохранения результатов запросов в формате PDF. Благодаря этому инструменту, пользователи получают удобные и понятные отчеты по динамическим данным из базы PostgreSQL. Этот функционал значительно облегчает процесс управления СУБД и предоставляет ценные инсайты для принятия бизнес-решений. Кроме того, переход от Foxpro к PostgreSQL и использование LibreOffice Base позволяет организациям существенно сократить расходы на лицензирование программного обеспечения, что делает это решение более экономически выгодным и доступным.

Дальнейшие исследования в интересах импортозамещения и повышения эффективности бизнес-решений, на наш взгляд, целесообразно продолжить в направлениях расширения и совершенствования способов интеграции современных инструментальных программных систем, примеру СУБД PostgreSQL и системы бизнес-аналитики [21], а также практического использования инструментальных средств робототехники для повышения степени автоматизации решения задач при переходе на новые болет перспективные технологии [22].

Литература

1. *Алпатов А.В.* Перспективы использования систем управления базами данных в разработке документов по планированию обучения // Актуальные проблемы защиты и безопасности. 2021. С. 381-384.

2. *Гуринов А.В., Воронов А.А.* Особенности использования программного обеспечения в системе документооборота // Молодежь и XXI век. 2022. С. 46-49.

3. *Шадрин П.В.* Российские и свободные альтернативы программного обеспечения, используемого на предприятии // Вектор развития. 2022. № 11. С. 103-107.

4. *Карпов В.В., Рындин А.А.* Импортозамещение в сфере информационных технологий на примере миграции с СУБД «Oracle» на СУБД PostgreSQL // Управление социально-экономическим развитием регионов: проблемы и пути их решения. 2023. С. 425-428.

5. *Ванина М.Ф., Ерохин А.Г., Тутова Н.В., Фролова Е.А.* Возможности использования отечественных операционных систем в современном высшем экономическом образовании // Методические вопросы преподавания инфокоммуникаций в высшей школе. 2023. № 1. С. 15-20.

6. *Устенко Г.А.* Использование кроссплатформенного пакета LibreOffice для оптимизации процесса работы с данными на торговых предприятиях // Advances in science and technology. 2018. С. 74-75.

7. *Шамшурин Д.А.* Пакет программ LibreOffice в сравнении с Microsoft Office // Инновационные процессы в научной среде. 2021. С. 92-97.

8. *Маркелов М.А., Коростелев Д.А.* Сравнительный анализ производительности СУБД MS SQL Server и PostgreSQL при одинаковой нагрузке // Новые горизонты. 2022. С. 163-166.

9. *Виноградова М.В., Барашкова Е.С., Березин И.С., Ореликов М.Г., Лузин Д.С.* Обзор системы полнотекстового поиска в постреляционной базе данных PostgreSQL // E-Scio. 2020.

10. *Заика И.В., Кумов Д.О.* Использование SQL-запросов как инструмента для взаимодействия с базами данных // Вестник Таганрогского института имени А.П. Чехова. 2023. № 2. С. 29-34.

11. *Махмутова М.В., Новик О.А.* Применение средств моделирования на этапах проектирования базы данных // Отходы и ресурсы. 2019. № 4. С. 18.

12. *Чибышева В.Р.* Организация системы защиты данных с использованием СУБД PostgreSQL // Обеспечение безопасности личности, общества и государства в условиях глобализации: правовые и организационные проблемы и перспективы. 2022. С. 549-551.

13. *Широков И.А.* Исследование возможностей хранения и обработки многоуровневых данных в PostgreSQL // Материалы X Международной научно-практической конференции «Культура, наука, образование: проблемы и перспективы». Санкт-Петербург, 2022. С. 451-457.

14. *Мантусов А.Б., Доржинова З.Б.* Формирование вычислительных компетенций у будущих экономистов в процессе использования программы Libreoffice Calc // Вектор науки Тольяттинского государственного университета. 2018. № 4. С. 26-33.

15. *Пантилимонов М.В., Бучацкий Р.А., Жуйков Р.А.* Кэширование машинного кода в динамическом компиляторе SQL-запросов для СУБД PostgreSQL // Труды Института системного программирования РАН. 2020. № 32.

16. *Никулина Е.Ю., Мещеряков М.О.* Сравнительный анализ программных продуктов Microsoft Office и LibreOffice // Охрана, безопасность, связь. 2020. № 5. С. 253-256.

17. *Смирнов А.А.* Оптимизация запросов в PostgreSQL: стратегии и приемы // Научный Лидер. 2023. № 29. С. 34-36.

18. *Давыдовский М.А.* Поиск кратчайших путей в базе данных PostgreSQL и их отображение в геоинформационной системе QGIS // Цифровая трансформация транспорта: проблемы и перспективы. 2021. С. 85-88.

19. *Моргунов Е.П., Моргунова О.Н.* Технология использования метода «анализ среды функционирования» в виде расширения СУБД PostgreSQL // Системный анализ в проектировании и управлении. 2021. С. 134-139.

20. *Романов В., Солнцева О., Трескова А.* Опыт перехода на альтернативное программное обеспечение в учебном процессе УЛГАУ // Свободное программное обеспечение в высшей школе. 2020. С. 177–180.

21. *Маркевич Д.В., Харланова В.В., Хомоненко А.Д.* Интеграция систем бизнес-аналитики с системами управления базами данных на транспорте // Наукоемкие технологии в космических исследованиях Земли. 2023. Т. 15. № 2. С. 41-48.

22. *Ермаков С.Г., Баталов Д.И., Мельников И.С.* Использование платформы Robin RPA в процессе цифровой трансформации транспортных компаний // Интеллектуальные технологии на транспорте. 2023, 1(33). С. 5-14.

FROM FOXPRO TO POSTGRESQL: OPTIMIZATION, EFFECTIVE DATA MANAGEMENT, AND REPORT GENERATION

DANIIL V. MARKEVICH,
St. Petersburg, Russia, dmarkevich811@mail.ru

ANATOLY D. KHOMONENKO,
St. Petersburg, Russia, khomon@mail.ru

SERGEY G. ERMAKOV,
St. Petersburg, Russia, ermakov@pgups.ru

KEYWORDS: *database management system, Foxpro, PostgreSQL, report generation, LibreOffice Base, cost reduction, data security.*

ABSTRACT

Introduction: In the rapidly evolving technological landscape and increasing demand for efficient database management, this article presents a practical implementation of transitioning from the outdated Foxpro database management system (DBMS) to the powerful and flexible PostgreSQL DBMS. The primary focus is on the development of functionality that automatically generates PDF reports using the LibreOffice Base tool. **Purpose:** The research aims to study and demonstrate an effective transition technology from the outdated Foxpro DBMS to the modern PostgreSQL DBMS, as well as to develop a mechanism for generating automated PDF reports using LibreOffice Base. **Methods:** The transition technology is implemented using the PostgreSQL DBMS, and PDF report generation is carried out through the LibreOffice Base tool. **Results:** The research results include a successful transition from Foxpro to PostgreSQL and the development of functionality for automated PDF report generation using LibreOffice Base to enhance database management efficiency and reduce software licensing costs, ensuring data reliability and security. **Practical relevance:** The main practical significance of the implemented technology lies in the complete compatibility of two different systems: a database providing extensive features and an analytics platform. The reasons and benefits of integrating these systems are presented, along with their application in solving specific tasks. The possibilities of optimizing data management and improving the quality of managerial decisions based on the integration of these systems are also discussed. **Discussion:** The entire transition process is discussed in the context of a specific example using the PostgreSQL DBMS and the LibreOffice Base tool to create databases and generate reports. Further research may focus on enhancing analytics for large volumes of data from multiple databases and improving the quality of managerial decisions based on the integration of these systems.

REFERENCES

1. Alpatov A.V. Prospects of Using Database Management Systems in Educational Planning Document Development. *Current Issues in Security and Safety*. 2021, pp. 381-384.

2. Gurinov A.V., Voronov A.A. Features of Software Usage in Document Management Systems. *Youth and the 21st Century*. 2022, pp. 46-49.

3. Shadrin P.V. Russian and Open Source Alternatives to Enterprise Software. *Development Vector*. 2022. No. 11, pp. 103-107.

4. Karpov V.V., Ryndin A.A. Import Substitution in Information Technology: A Case Study of Migrating from "Oracle" DBMS to "PostgreSQL" DBMS. *Management of Socio-Economic Development of Regions: Problems and Solutions*. 2023, pp. 425-428.

5. Vanina M.F., Erokhin A.G., Tutova N.V., Frolova E.A. Possibilities of Using Domestic Operating Systems in Modern Higher Economic Education. *Methodological Issues of Teaching Infocommunications in Higher Education*. 2023. No. 1, pp. 15-20.

6. Ustenko G.A. Using the Cross-Platform Package LibreOffice to Optimize Data Handling in Trade Enterprises. *Advances in Science and Technology*. 2018, pp. 74-75.

7. Shamshurin D.A. Comparison of LibreOffice Software Suite with Microsoft Office. *Innovative Processes in the Scientific Environment*. 2021, pp. 92-97.

8. Markelov M.A., Korostelev D.A. Comparative Analysis of Performance between MS SQL Server and PostgreSQL under Equal Workloads. *New Horizons*. 2022, pp. 163-166.

9. Vinogradova M.V., Barashkova E.S., Berezin I.S., Orelikov M.G., Luzin D.S. Overview of a Full-Text Search System in PostgreSQL Post-Relational Database. *E-Scio*. 2020.

10. Zaika I.V., Kumov D.O. Using SQL Queries as a Tool for Interacting with Databases. *Bulletin of Taganrog Institute named after A.P. Chekhov*. 2023. No. 2, pp. 29-34.

11. Makhmutova M.V., Novik O.A. Application of Modeling Tools in the Stages of Database Design. *Wastes and Resources*. 2019. No. 4, pp. 18.

12. Chibysheva V.R. Data Protection System Organization Using PostgreSQL DBMS. *Ensuring Security of Individuals, Society, and the State in the Age of Globalization: Legal and Organizational Issues and Prospects*. 2022, pp. 549-551.

13. Shirokov I.A. Investigating the Capabilities of Storing and Processing Multi-Level Data in PostgreSQL. *Materials of the 10th International Scientific and Practical Conference "Culture, Science, Education: Problems and Prospects*. St. Petersburg, 2022, pp. 451-457.

14. Mantusov A.B., Dorzhinova Z.B. Formation of Computational Competencies for Future Economists through the Use of LibreOffice Calc Software. *Science Vector of Togliatti State University*. 2018. No. 4, pp. 26-33.

15. Pantilimonov M.V., Buchatsky R.A., Zhuikov R.A. Caching Machine Code in a Dynamic Compiler for SQL Queries in PostgreSQL DBMS. *Works of the Institute for System Programming of the RAS*. 2020. No. 32.

16. Nikulina E.Y., Meshcheryakov M.O. Comparative Analysis of Microsoft Office and LibreOffice Software Products. Ohrana, bezopasnost', svyaz' [Security, Safety, Communication]. 2020. No. 5. Pp. 253-256.

17. Smirnov A.A. Query Optimization in PostgreSQL: Strategies and Techniques. *Scientific Leader.* 2023. No. 29, pp. 34-36.

18. Davydovskiy M.A. Finding Shortest Paths in PostgreSQL Database and Displaying Them in the Geoinformation System QGIS. *Digital Transformation of Transportation: Issues and Prospects*. 2021, pp. 85-88.

19. Morgunov E.P., Morgunova O.N. Technology of Using the "Environment Analysis" Method as an Extension for PostgreSQL DBMS. *System Analysis in Design and Management.* 2021, pp. 134-139.

20. Romanov V., Solntseva O., Treskova A. Experience of Transitioning to Alternative Software in the Educational Process at ULSAU. *Free Software in Higher Education*. 2020, pp. 177-180.

21. Markevich D.V., Kharlanova V.V., Khomonenko A.D. Integration of business intelligence systems with database management systems in transport. *H&ES Reserch*. 2023. Vol. 15. No. 2, pp. 41-48. (In Rus)

22. Ermakov S.G., Batalov D.I., Melnikov I.S. Using the Robin RPA Platform in the Process of Digital Transformation of Transport Companies. *Intellectual Technologies on Transport*. 2023. No. 1, pp. 5-14. (In Rus)

INFORMATION ABOUT AUTHORS:

Daniil V. Markevich, *student at the Department of Information and Computing systems of Emperor Alexander I St. Petersburg state transport university, St. Petersburg, Russia*

Anatoly D. Khomonenko, *Grand PhD, Full Professor, Professor of the Department of Information and Computing systems of Emperor Alexander I St. Petersburg state transport university, St. Petersburg, Russia*

Sergey G. Ermakov, *Grand PhD, Full Professor, Head of the Department of Information and Computing systems of Emperor Alexander I St. Petersburg state transport university, St. Petersburg, Russia*

For citation: Markevich D.V., Khomonenko A.D., Ermakov S.G. From Foxpro to PostgreSQL: optimization, effective data management, and report generation. *H&ES Reserch*. 2024. Vol. 16. No. 1. P. 21-30. doi: 10.36724/2409-5419-2024-16-1-21-30 (In Rus)

doi: 10.36724/2409-5419-2024-16-1-31-38

КОНЦЕПЦИЯ ПОДДЕРЖКИ ПРИНЯТИЯ РЕШЕНИЙ ЗАДАЧ СОПРОВОЖДЕНИЯ ЖИЗНЕННОГО ЦИКЛА СЛОЖНЫХ ТЕХНИЧЕСКИХ СИСТЕМ

ГУСЕНИЦА
Ярослав Николаевич[1]

ДОНЧЕНКО
Анатолий Анатольевич [2]

ЛЯСКИН
Антон Сергеевич [3]

АННОТАЦИЯ

Предметом исследования является процесс сопровождения жизненного цикла сложных технических систем. **Целью исследования** является внедрение информационных технологий в процесс сопровождения жизненного цикла. В работе изложена концепция и математическая модель поддержки принятия решений задач сопровождения жизненного цикла сложных технических систем. Отмечено, что поддержка принятия решений должна рассматриваться на следующих взаимосвязанных уровнях: системы информационной поддержки жизненного цикла как источника информации о текущем состоянии образца; цифровой модели как источника информации о прогнозируемом состоянии образца; программного обеспечения компьютерного моделирования, выступающего в роли инструментального средства исследования образца с использованием цифровой модели; информационного обеспечения для обобщения, обработки и хранения промежуточных и итоговых данных об образце; функциональных модулей, обеспечивающих интеллектуальную поддержку и автоматизацию; задач сопровождения жизненного цикла, определяющих целевую функцию системы поддержки принятия решений. **Результаты:** Представлены компоненты системы поддержки принятия решений и приведена их взаимосвязь. Показано, что внедрение системы поддержки принятия решений позволит решить следующие задачи сопровождения жизненного цикла: обоснование необходимости создания сложных технических систем; оценка результатов научно-исследовательских и опытно-конструкторских работ; оценка реализации требований к сложным техническим системам; мониторинг и оценка эксплуатационных характеристик образца системы. Даны рекомендации по дальнейшему развитию систем поддержки принятия решений, касающиеся: внедрения отечественных CALS-систем и их интеграции с программным обеспечением компьютерного моделирования; унификации форматов представления данных и содержания информации о свойствах материалов и веществ для CAE- и CALS-систем; совершенствования нормативной и правовой документации по вопросам проведения цифровых испытаний сложных технических систем; интеграции центров коллективного пользования вычислительных ресурсов и компонентов системы поддержки принятия решений.

КЛЮЧЕВЫЕ СЛОВА: *система поддержки принятия решений, система информационной поддержки жизненного цикла, система управления жизненным циклом, база данных, база знаний, сложная техническая система*

Сведения об авторах:

[1] начальник научно-исследовательского отдела, кандидат технических наук, федеральное государственное автономное учреждение "Военный инновационный технополис "ЭРА", г. Анапа, Россия, era_otd1@mil.ru

[2] главный научный сотрудник, профессор, доктор технических наук, федеральное государственное унитарное предприятие "Российский федеральный ядерный центр - Всероссийский научно-исследовательский институт экспериментальной физики", г. Саров, Россия, press@dc.vniief.ru

[3] кандидат технических наук, доцент, федеральное государственное автономное учреждение "Военный инновационный технополис "ЭРА", г. Анапа, Россия, старший научный сотрудник научно-исследовательского отдела, era_otd1@mil.ru

Для цитирования: *Гусеница Я.Н., Донченко А.А., Ляскин А.С.* Концепция поддержки принятия решений задач сопровождения жизненного цикла сложных технических систем // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16. № 1. С. 31-38. doi: 10.36724/2409-5419-2024-16-1-31-38

Введение

В настоящее время развитие сложных технических систем (СТС) на стадиях жизненного цикла характеризуется высокой степенью неопределенности, которая отражает неполноту информации об образцах, протекающих в них процессах, а также внутренних и внешних воздействий.

Сопровождение жизненного цикла представляет собой скоординированную по целям (задачам), времени и ресурсам совокупность мероприятий, выполняемых на стадиях жизненного цикла образцов СТС, по их анализу, экспертизе, контролю проведения испытаний, а также оценке соответствия получаемых результатов требованиям заказчика.

Достижение целей сопровождения жизненного цикла СТС сопряжено с рядом трудностей, которые обусловлены необходимостью решать плохо структурированные и неструктурированные задачи.

В этих условиях актуальным становится задача внедрения информационных технологий процесс сопровождения жизненного цикла СТС.

Концептуальные основы поддержки принятия решений задач сопровождения жизненного цикла СТС

Основной целью внедрения информационных технологий в процессы сопровождения жизненного цикла является автоматизация оценок реализации требований, предъявляемых к сложным техническим системам, эксплуатационных характеристик образцов СТС, научного-технического уровня СТС и новизны результатов научно-исследовательских (НИР) и опытно-конструкторских работ (ОКР) на стадиях жизненного цикла СТС.

Важным условием автоматизации сопровождения жизненного цикла является наличие и использование разнообразных по своему назначению автоматизированных систем (АС), которые представляют собой совокупность персонала и комплекса средств автоматизации его деятельности, реализующую информационную технологию выполнения установленной функции. Одним из высокоуровневых видов таких систем являются системы поддержки принятия решений.

Под системой поддержки принятия решений (СППР) (англ. Decision Support Systems, DSS) в дальнейшем будем понимать интерактивную АС [1], целью которой является помощь лицу, принимающему решения в сложных условиях.

СППР могут выполнять довольно сложные функции [2], [3, 4], например:

– генерацию возможных решений;
– выявление приоритетов решений;
– оценивание решений;
– выбор наилучшего решения.

Основу СППР составляют разнообразные средства хранения информации, модели принятия решений, развитый человеко-машинный интерфейс, а также информационно-управляющие системы и системы обработки информации. Они, как правило, базируются на использовании следующих технологий:

– технологий представления знаний [5, 6] (онтологии, семантические сети, фреймы, продукции), особенностью которых является отображение статических и динамических свойств предметной области;

– технологий обработки знаний [7, 8] (аналитическое, имитационное, ситуационное, символьное и логическое моделирование, искусственные нейронные сети, эволюционные и мягкие вычисления, агентно-ориентированные технологии и др.), предполагающие поиск решений.

Для достижения целей сопровождения жизненного цикла организация СППР должна рассматриваться как многоуровневая АС, которая позволяет комплексно обеспечить информационное взаимодействие различных программно-технических средств, задействованных на стадиях жизненного цикла, для формирования предложений в требования, программы и методики испытаний, по модернизации образцов СТС (рис. 1).

Системы информационной поддержки жизненного цикла СТС

Верхний уровень занимает система информационной поддержки жизненного цикла СТС (англ. Continuous Acquisition and Lifecycle Support, CALS). CALS-системы или системы управления жизненным циклом изделия (англ. Product Lifecycle Management, PLM) [9, 10] активно применяются при разработке и производстве сложных наукоемких изделий, которые создаются интегрированными промышленными структурами, включающими в себя научно-исследовательские организации, конструкторские бюро, подрядчиков, субподрядчиков, поставщиков, предприятия технического обслуживания, ремонта и утилизации [11, 12].

В настоящее время широкое применение на предприятиях промышленности нашли зарубежные CALS-системы: Siemens Teamcenter (Германия) (www.plm.automation.siemens.com/global/en/products/teamcenter/) и PTC Windchill (США) (www.ptc.com/en/products/windchill). Отечественные системы Лоцман: PLM (разработка АО «Аскон»; ascon.ru/products/889/review/) и T-Flex PLM (разработка ЗАО «ТопСистемы»; www.tflex.ru/plm/) пока еще не нашли широкого применения. Еще одна отечественная CALS-система, ЛОГОС: Данные, сейчас разрабатывается ФГУП «РФЯЦ-ВНИИЭФ» в рамках развития многофункционального программного комплекса ЛОГОС.

Указанные выше программные решения за счет унификации форматов данных, описывающих детали, агрегаты и составные части изделий, а также применения электронного документооборота, обеспечивают единообразные способы управления бизнес-процессами всех участников жизненного цикла, что позволяет существенно сократить потребный объем ресурсов на выполнение работ.

В аспекте решения задачи сопровождения жизненного по оценке реализации требований, предъявляемых к СТС, интерес представляют следующие функциональные возможности CALS-систем [13, 14]:

1. равление требованиями.
2. равление соответствием.
3. У равление системными описаниями.

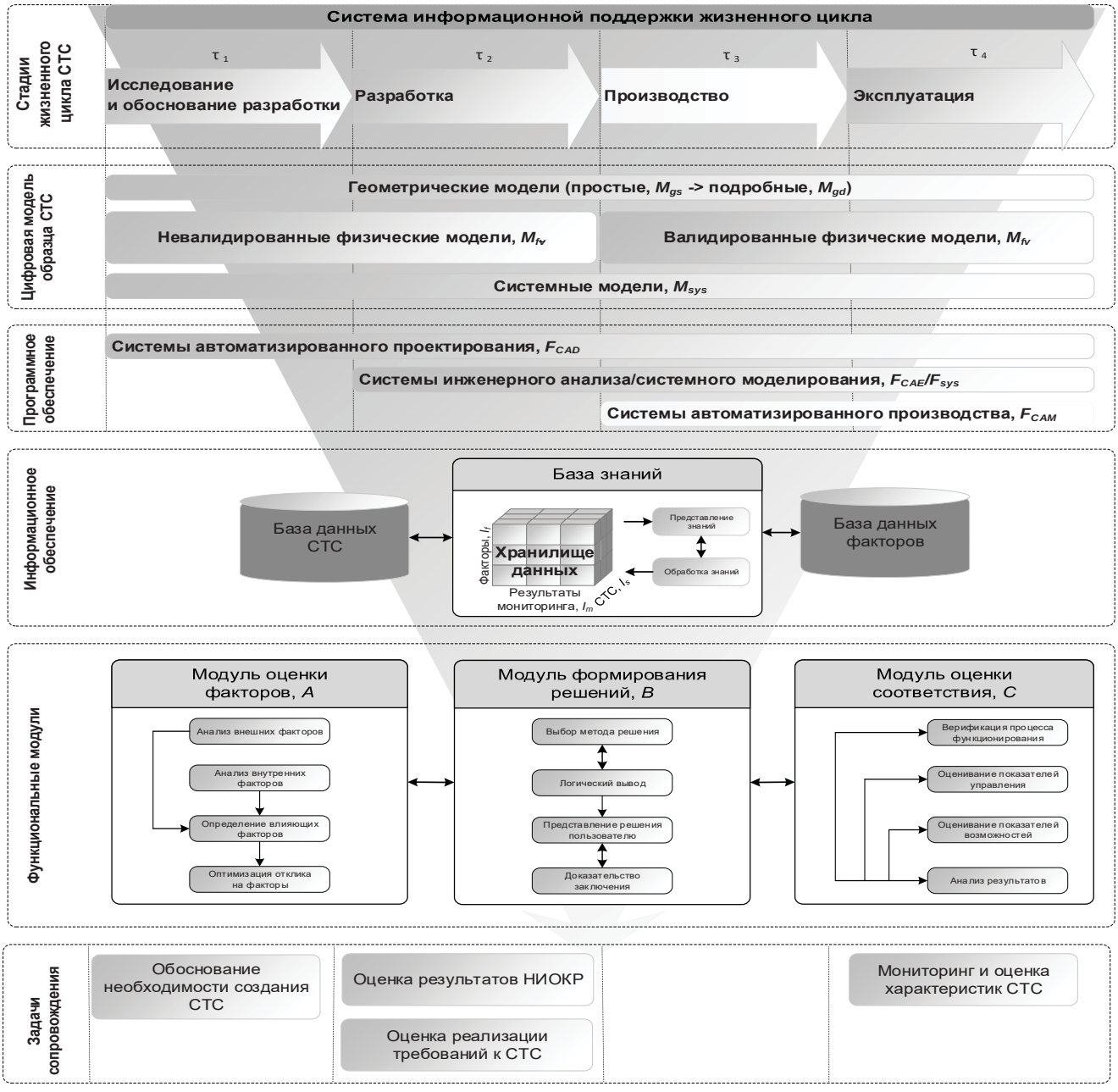


Рис. 1. Структурная схема концепции поддержки принятия решений задач сопровождения жизненного цикла СТС

Первая функциональная возможность позволяет в структурированном виде создавать требования к образцу СТС, а также связывать требования с его системными описаниями, элементами проектного, конструкторского и технологического составов.

Вторая функциональная возможность обеспечивает управление рисками, которые связаны с невыполнением требований, предъявляемых к образцу СТС.

Наконец, третья функциональная возможность позволяет создавать высокоуровневые описания подсистем образца СТС, являющиеся основой для разработки его процессных моделей.

Кроме того, CALS-системы могут выступать в качестве источника информации о геометрических характеристиках образца, свойствах используемых в нем материалов и рабочих сред, внешних воздействиях в ходе эксплуатации и требованиях к образцу СТС. Эта информация используется при построении цифровой модели.

Информационное обеспечение системы поддержки принятия решений задач сопровождения жизненного цикла СТС

Важнейшей составной частью комплекса средств автоматизации СППР является информационное обеспечение, которое в соответствии с ГОСТ 34.003-90 представляет собой совокупность форм документов, классификаторов, норма-

тивной базы и реализованных решений по объемам, размещению и формам существования информации, применяемой в АС при ее функционировании.

В зависимости от уровня иерархии АС информационное обеспечение реализуют с помощью баз данных (БД) [15], хранилищ данных [16, 17] и баз знаний [18] (рис. 2).

Рис. 2. Уровни иерархии информационного обеспечения АС

В интересах решения задач сопровождения жизненного цикла основу информационного обеспечения СППР должны составить:

БД образцов, содержащая информацию I_s о характеристиках отечественных и зарубежных СТС;

БД факторов, содержащая информацию I_f, которая описывает условия функционирования СТС.

Указанные выше БД позволят обеспечить экспертизу и сравнительный анализ отечественных образцов СТС и их зарубежных аналогов.

Кроме того, БД образцов и факторов можно объединить в многомерное хранилище данных $I_f \times I_s \times I_m$, содержащее не только информацию о характеристиках и условиях функционирования, но и о результатах мониторинга СТС I_m. Основным содержанием хранилищ данных являются факты, представляющие собой информацию, достоверность которой доказана. Факты позволяют описывать характеристики образцов СТС в конкретные моменты времени в условиях внешних и внутренних воздействующих факторов. Они получаются из данных путем их обработки, обобщения и представления в виде, удобном и понятном для пользователей.

Хранилище данных о СТС, формируя единый предметно-ориентированный источник фактов, позволит обеспечить решение следующих подзадач сопровождения жизненного цикла:

– анализ остояния и перспектив развития СТС;

– анализ результатов мониторинга эксплуатационных характеристик образцов СТС;

– обобщение опыта эксплуатации и применения СТС;

– оценка требуемого объема и степени воспроизведения эксплуатационных условий при испытаниях в установленном порядке на предприятиях-изготовителях;

– анализ эксплуатационных характеристик СТС на стадиях жизненного цикла;

– анализ информации о ходе эксплуатации образцов (модернизируемых образцов) СТС по исходным данным эксплуатирующих организаций;

– учет и анализ информации об изменении конструкторской документации и рекламаций на образцы СТС.

В свою очередь, хранилище данных о СТС может заложить основу для формирования базы знаний, которая представляет собой совокупность знаний о предметной области, организованных в соответствии с определенной схемой представления. Главным наполнением базы знаний являются знания, которые отражают причинно-следственные зависимости предметной области, полученные в результате практического опыта.

Базы знаний необходимы не только для хранения уже имеющихся знаний о предметной области, но и извлечения новых знаний, необходимых для принятия обоснованных решений. Поэтому база знаний о СТС должна содержать не только факты, но и функции представления и обработки знаний.

Указанная база знаний позволит обеспечить решение следующих подзадач сопровождения жизненного цикла:

– разработка проектов технических требований к перспективным и модернизируемым образцам СТС, технических заданий на НИР и ОКР;

– разработка проектов программ и методик испытаний, апробации СТС (подконтрольной эксплуатации);

– разработка предложений по оценке и повышению эффективности применения существующих образцов СТС и их дальнейшей модернизации;

– разработка предложений по совершенствованию конструкции образцов СТС, их системы технической эксплуатации или эксплуатационно-ремонтной документации.

Цифровая модель образца СТС

Цифровая модель [19] представляет собой систему математических и компьютерных моделей, а также электронных документов, описывающую структуру, функциональность и поведение вновь разрабатываемого или эксплуатируемого образца СТС на различных стадиях жизненного цикла.

В зависимости от исследуемых свойств и стадии жизненного цикла образцов СТС могут использоваться геометрические (M_{gs}, M_{gd}), физические (M_{fv}, M_{fv}) и системные компьютерные модели M_{sys}.

Геометрическая модель содержит необходимые сведения о форме, размерах и свойствах образца СТС, зависящих от его формы и размеров. Степень детализации геометрической модели зависит, в первую очередь, от стадии жизненного цикла образца СТС, а также от ее предназначения:

$$M_{gs} = F_{CAD}\left(F_{CALS}(I_s), \tau_1\right);$$

$$M_{gd} = F_{CAD}\left(F_{CALS}(I_s), \tau_2\right).$$

Физические модели описывают физические свойства образца СТС и его взаимодействие с внешней средой (статика, кинематика, динамика твердого тела, гидро- и газодинамика, деформации, теплопроводность, электромагнитное взаимодействие и др.). Физические модели создаются на основе геометрических:

$$M_{fv} = F_{CAE}\left(M_g, F_{CALS}(I_{\tilde{f}} \times I_s), \tau_2\right);$$

$$M_{fv} = F_{CAE}\left(M_g, F_{CALS}(I_{\tilde{f}} \times I_s) \times I_m, \tau_4\right).$$

Системные модели могут быть функциональными или процессными. При этом функциональные модели формализовано описывают функции отдельных подсистем и составных частей образца СТС, а процессные – технологический процесс его изготовления или технический процесс его эксплуатации:

$$M_{sys} = F_{sys}\left(M_f, I_m, \tau_1, .., \tau_4\right).$$

Результаты моделирования являются одним из важнейших источников информации для принятия решений в задачах сопровождения жизненного цикла образцов СТС. Например, результаты системного моделирования могут использоваться для обоснования технических требований к образцу СТС. Результаты физического моделирования (цифровых испытаний) могут применяться для разработки предложений в программы и методики испытаний, а также предложений по модернизации образца СТС.

Программное обеспечение компьютерного моделирования образцов СТС

Для разработки и исследования образцов СТС с использованием компьютерных моделей требуется специальное программное обеспечение.

Создание геометрических моделей возможно с помощью отечественных систем автоматизированного проектирования (англ. Computer-Aided Design, CAD):

Компас 3D (разработка ООО «Аскон»; kompas.ru/kompas-3d/about/);

T-Flex CAD (разработка ЗАО «ТопСистемы»; www.tflexcad.ru);

nanoCAD 3D (разработка ООО «Нанософт разработка»; www.nanocad.ru/products/platform/).

Для создания физических моделей применяются системы инженерного анализа (англ. Computer-Aided Engineering, CAE), позволяющие решать с помощью различных численных методов системы основных дифференциальных уравнений в конкретных областях физики. Как правило, CAE классифицируют по следующим областям:

– проч ость (механика деформируемого твердого тела);
– динамика и кинематика;
– эрогидродинамика (механика жидкости и газа);
– т рмодинамика;
– эле тромагнетизм.

Кроме того, в последнее время все большую значимость приобретает многодисциплинарное моделирование, учитывающее взаимосвязанные явления из разных областей физики. Примерами таких задач могут служить задачи аэроупругости (аэродинамика + прочность, например, деформация конструкции под действием аэродинамических сил и влияние этой деформации на обтекание конструкции потоком газа) и магнитогидродинамики (аэрогидродинамика + электромагнетизм, например, возникновение электромагнитного поля за счет конвективного переноса носителей заряда потоком жидкости, газа или плазмы и влияние этого электромагнитного поля на поток жидкости). Для решения такого рода задач могут применяться два подхода:

1) создание специальных систем многодисциплинарного моделирования, в которых в основе математической модели лежит сопряженная система уравнений из разных областей физики;

2) использование вспомогательного программного обеспечения для организации обмена данными между системами компьютерного моделирования из разных областей физики.

На сегодняшний день существуют следующие коммерческие отечественные системы физического моделирования, зарегистрированные в реестре российского программного обеспечения:

ЛОГОС (разработка ФГУП «РФЯЦ-ВНИИЭФ»; logos.vniief.ru);

АПМ Multiphysics (разработка НТЦ «Автоматизированное Проектирование Машин»; apm.ru/apm-multiphysic);

Fidesys (разработка ООО «Фидесис»; cae-fidesys.com/products/desktop/);

Euler (разработка ООО «АвтоМеханика»; www.euler.ru);

Universal Mechanism (разработка ООО «Вычислительная механика»; www.umlab.ru);

FlowVision (разработка ООО «ТЕСИС»; flowvision.ru/ru/flowvision-technologies/description).

Для создания системных функциональных моделей может применяться система 1D-моделирования SimInTech (разработка ООО «3В Сервис»; simintech.ru/).

Для разработки системных моделей процесса функционирования образца СТС может применяться отечественная система разработки цифровых двойников CML-Bench (разработка Санкт-Петербургского политехнического университета; cml-bench.ru).

Функциональные модули системы поддержки принятия решений задач сопровождения жизненного цикла СТС

Функционирование СППР представляет собой процесс, который является аналогом мыслительной деятельности человека и заключается в формировании решения в условиях динамически изменяющихся значений параметров окружающей среды.

Данный процесс затрагивает этапы «наблюдение», «ориентация» и «решение» в цикле Бойда [20], каждый из которых реализуется соответствующими функциональными модулями СППР.

В аспекте решения задач сопровождения жизненного цикла СТС в состав СППР должны входить следующие функциональные модули:

– модуль оценки факторов, *A*;
– модуль оценки соответствия, *B*;
– модуль формирования решения, *C*.

Модуль оценки факторов реализует этап «наблюдение» и выполняет анализ условий функционирования СТС, включая внешние и внутренние воздействующие факторы, и определение среди них наиболее влияющих:

$$A: \quad I_{\tilde{f}} \subset I_f.$$

Модуль оценки соответствия реализует этап «ориентация» и выполняет верификацию процесса функционирова-

ния, оценивание показателей возможностей и управления СТС, а также анализ результатов. Данный функциональный модуль СППР позволяет осуществлять проверку соответствия СТС и ее цифровой модели предъявляемым требованиям:

$$C: \quad \begin{matrix} \exists s_i, W_{req} \le W_i \\ W_i = \int\limits_{Q_i} \Phi\left(u_i, v_i, r_i\right) dr \end{matrix},$$

где s_i – вариант СТС, W_{req} – требования к эффективности, W_i – интегральный показатель эффективности для i-го варианта, Ф – модель процесса функционирования СТС, u – вектор показателей управления, v – вектор показателей возможностей, $r \in Q$ – вектор условий функционирования, Q – область допустимых условий функционирования.

Модуль формирования решения реализует этап «решение» и осуществляет выбор методов решения задач сопровождения жизненного цикла, логический вывод, представление решения пользователю и доказательство заключения:

$$B: \quad s_i = \arg\max\left(W_i\right).$$

При этом для обеспечения выбора методов решения, логического вывода и доказательства заключений могут применяться компьютерные модели образца СТС, искусственные нейронные сети, эволюционные и мягкие вычисления, агентно-ориентированные технологии и др. [21], [22]. В свою очередь, для представления решения пользователю могут использоваться онтологии, семантические сети, фреймы и продукции.

Указанные функциональные модули обеспечивают условия реализации процесса функционирования СППР при конкретных значениях параметров окружающей среды для решения следующих задач сопровождения жизненного цикла СТС:

– обоснование необходимости создания СТС;

– оценка результатов НИР и ОКР;

– оценка реализации требований к образцу СТС;

– мониторинг и оценка эксплуатационных характеристик образца СТС.

Заключение

Таким образом, вопрос организации СППР задач сопровождения жизненного цикла СТС является весьма актуальным. Реализация СППР должна рассматриваться на следующих взаимосвязанных уровнях:

– системы информационной поддержки жизненного цикла как источника информации о текущем состоянии СТС на всех этапах работ;

– цифровой модели как источника информации о прогнозируемом состоянии СТС;

– программного обеспечения компьютерного моделирования, выступающего в роли инструментального средства исследования СТС с использованием цифровой модели;

– информационного обеспечения для обобщения, обработки и хранения промежуточных и итоговых данных о СТС;

– функциональных модулей, обеспечивающих интеллектуальную поддержку решения и автоматизацию выполнения задач сопровождения жизненного цикла;

– задач сопровождения жизненного цикла СТС, определяющих целевую функцию СППР.

Дальнейшими направлениями развития СППР задач сопровождения жизненного цикла СТС могут стать:

1. Широкое внедрение отечественных CALS-систем и их интеграция с программным обеспечением компьютерного моделирования.

2. Унификация форматов представления данных и содержания информации о свойствах материалов и веществ для CAE и CALS-систем.

3. овершенствование нормативной и правовой документации по вопросам использования невалидированных компьютерных моделей и проведения цифровых испытаний образцов СТС.

4. Интеграция центров коллективного пользования вычислительных ресурсов организаций изготовителей и эксплуатантов и компонентов СППР задач сопровождения жизненного цикла СТС.

Литература

1. *Балашова И.В., Терещенко Т.А.* Системы поддержки принятия решений // The scientific heritage, No 79, 2021. С. 3-7.

2. *Кондрашов Ю.Н.* Использование аналитические технологии в системах поддержки принятия решений на этапах жизненного цикла аэрокосмических изделий // Научно-технический вестник Поволжья, №5, 2019. С. 37-40.

3. *Cantini A., Peron M., De Carlo F., Sgarbossa F.* A decision support system for configuring spare parts supply chains considering different manufacturing technologies // International Journal of Production Research (Online), 2022. 21 p. URL: https://doi.org/10.1080/00207543.2022.2041757

4. *Wallinga E., Vaneeckhautea C.* Developing Successful Environmental Decision Support Systems: Challenges and Best Practices // Journal of Environmental Management (Online). Vol. 264, 2020, 110513. 42 p. URL: https://doi.org/10.1016/j.jenvman.2020.110513

5. *Жихарев А.Г.* О новой технологии представления знаний для систем поддержки принятия решений // Научные ведомости Белгородского государственного национального исследовательского университета. Серия «История. Политология. Экономика. Информатика». 2011. №19 (114). Выпуск 20/1. С. 158-162.

6. *Chen X., Jia S., Xiang Y.* A review: Knowledge reasoning over knowledge graph // Expert Systems with Applications (Online). Vol. 141, 2020, 112948. 21 p. URL: https://doi.org/10.1016/j.eswa.2019.112948

7. *Di Vaio A., Palladino R., Pezzi A., Kalisz D.E.* The role of digital innovation in knowledge management systems: A systematic literature review // Journal of Business Research, 123, 2021, pp. 220-231.

8. *Черняховская Л.Р., Никулина Н.О., Бармина О.В.* Применение онтологического анализа и обработки знаний для принятия решений в проблемных ситуациях // Материалы XV международной научной конференции «Управление экономикой: методы, модели, технологии». В 2х томах. Том 2. Уфа: Уфимский государственный авиационный технический университет, 2015. С. 181-186.

9. *Яблочников Е.И., Фомина Ю.Н., Саломатина А.А.* Компьютерные технологии в жизненном цикле изделия. СПб: ИТМО, 2010. 188 с.

10. *Liu G., Man R., Wang Y.* A Data Management Approach Based on Product Morphology in Product Lifecycle Management // Processes (Online). Vol. 9, 2021, 1235. 22 p. URL: https://doi.org/10.3390/pr9071235

11. NATO CALS Handbook ATO CALS Office, 2000. 329 p.
12. *Stark J.* Product Lifecycle Management (Volume 2): The Devil is in the Details. Springer, 2015. 634 p.
13. *Есенжол А.М., Куровский А.А., Малахов И.И.* Анализ CALS–технологий при проектировании // Материалы VI международной конференции «Техника и технологии машиностроения». Омск: Омский государственный технический университет, 2017. С. 77-81.
14. *Morshedzadeh I., Ng A.H.C., Jeusfeld M.* Managing manufacturing data and information in product lifecycle management systems considering changes and revisions // International Journal of Product Lifecycle Management. Vol. 13. No. 3, 2021, pp. 244-264.
15. *Новиков Б.А., Горшкова Е.А., Графеева Н.Г.* Основы технологий баз данных. М.: ДМК Пресс, 2020. 582 с.
16. *Кондрашов Ю.Н.* Развитие хранилищ данных для работы с Big Data в системах поддержки принятия решений // Научно-технический вестник Поволжья, №9, 2019. С. 30-32.
17. *Львович Я.Е., Преображенский А.П.* Характеристики хранилища данных в корпоративных системах // Вестник Воронежского института высоких технологий, 2020, № 1. С. 31-33.
18. *Serrano S.A., Santiago E., Martinez-Carranza J. et al.* Knowledge-Based Hierarchical POMDPs for Task Planning // Journal of Intelligent & Robotic Systems (Online), 2021, 101:82. 30 p.
19. *VanDerHorn E., Mahadevan S.* Digital Twin: Generalization, characterization and implementation // Decision Support Systems (Online). Vol. 145, 2021, 113524. 11 p. URL: https://doi.org/10.1016/j.dss.2021.113524
20. *Ивлев А.А.* Основы теории Бойда. Направления развития, применения и реализации. – М.: Знание, 2008. 64 с.
21. *Dudnyk V., Sinenko Yu., Matsyk M. et al.* Development of a method for training artificial neural networks for intelligent decision support systems // Eastern-European Journal of Enterprise Technologies, Vol. 105, No. 3/2, 2020, pp. 37-47.
22. *Merkert J., Mueller M., Hubl M.* A Survey of the Application of Machine Learning in Decision Support Systems // Proceedings of 23rd European conference on information systems. University of Hoffenhaim, 2015. 15 p.

DECISION SUPPORT CONCEPT FOR LIFECYCLE SUPPORT PROBLEMS OF COMPLEX TECHNICAL SYSTEMS

YAROSLAV N. GUSENITZA,
Anapa, Russia, era_otd1@mil.ru

ANATOLIY A. DONCHENKO,
Sarov, Russia, press@dc.vniief.ru

ANTON S. LYASKIN,
Anapa, Russia

KEYWORDS: *lifecycle support, decision support system, continuous acquisition and lifecycle support system, product lifecycle management system, database, knowledge base.*

ABSTRACT

Introduction: The object of the present study is the process of lifecycle support of complex technical systems. The purpose of the present study is application of modern information technologies to this process. We present a concept and mathematical model of decision support for lifecycle support problems. **Methods:** We note that decision support concept must be considered at several interconnected levels: lifecycle information support systems providing data on actual status complex technical system; digital models providing data on expected status of complex technical system; simulation and modeling software as an instrument used for studying complex technical system with the help of its digital model; data processing software used for summary, analysis and storage of intermediate and final data on complex technical system; functional modules providing intellectual support for solution and automation of lifecycle support problems; lifecycle support problems themselves as they decide the goal function of decision support system. We propose components of decision support system and show their interaction. **Results:** We demonstrate that realization of proposed decision support system will allow solving of following lifecycle support problems: justification of necessity for development of a new complex technical system; evaluation of research and development results; evaluation of requirements fulfillment; monitoring and evaluation of system performance. We recommend some steps for further development of decision support system: adoption of domestic CALS systems and their integration with simulation and modeling software; unification of data formats, including data on material properties, between CAE and CALS software; improvement of standards and laws governing digital testing of complex technical systems; integration of decision support system components with industrial computing resources and data centers.

REFERENCES

1. Balashova I.V., Tereschenko T.A. Decision support systems. *The scientific heritage*. No. 79, 2021, pp. 3-7.

2. Kondrashov Yu.N. Usage of analytical technologies in decision support systems at lifecycle stages of aerospace products. *Volga region scientific and technical herald*. No.5, 2019, pp. 37-40.

3. Cantini A., Peron M., De Carlo F., Sgarbossa F. A decision support system for configuring spare parts supply chains considering different manufacturing technologies. *International Journal of Production Research* (Online), 2022. 21 p. URL: https://doi.org/10.1080/00207543.2022.2041757

4. Wallinga E., Vaneeckhautea C. Developing Successful Environmental Decision Support Systems: Challenges and Best Practices. *Journal of Environmental Management* (Online). Vol. 264, 2020, 110513. 42 p. URL: https://doi.org/10.1016/j.jenvman.2020.110513

5. Zhiharev A.G. On a novel technology of knowledge representation for decision support system. *Scientific reports of Belgorod state national research university. Series "History. Political Science. Economics. Informational Science"*. 2011. No.19 (114). Vol. 20/1, pp. 158-162.

6. Chen X., Jia S., Xiang Y. A review: Knowledge reasoning over knowledge graph. *Expert Systems with Applications* (Online). Vol. 141, 2020, 112948. 21 p. URL: https://doi.org/10.1016/j.eswa.2019.112948

7. Di Vaio A., Palladino R., Pezzi A., Kalisz D.E. The role of digital innovation in knowledge management systems: A systematic literature review. *Journal of Business Research*, 123, 2021, pp. 220-231.

8. Chernyahovskaya L.R., Nikulina N.O., Barmina O.V. Usage of ontological analysis and knowledge processing for decision making in problematic situations. *Proceeding of the XV international scientific conference "Economics control: methods, models, technologies"*. Vol. 2 of 2. Ufa: Ufa State Aviation Technical University, 2015, pp. 181-186.

9. Yablochnikov E.I., Fomina Yu.N., Salomatina A.A. Computer technologies in product lifecycle. Saint-Petersburg: ITMO, 2010. 188 p.

10. Liu G., Man R., Wang Y. A Data Management Approach Based on Product Morphology in Product Lifecycle Management. *Processes* (Online). Vol. 9, 2021, 1235. 22 p. URL: https://doi.org/10.3390/pr9071235

11. NATO CALS Handbook. NATO CALS Office, 2000. 329 p.

12. Stark J. Product Lifecycle Management (Volume 2): The Devil is in the Details. Springer, 2015. 634 p.

13. Esenzhol A.M., Kurovskij A.A., Malakhov I.I. Analysis of CALS-technologies in design. *Proceedings of the VI international conference "Technic and technologies in engineering"*. Omsk: Omsk State Technical University, 2017, pp. 77-81.

14. Morshedzadeh I., Ng A.H.C., Jeusfeld M. Managing manufacturing data and information in product lifecycle management systems considering changes and revisions. *International Journal of Product Lifecycle Management*, Vol. 13 No. 3, 2021, pp. 244-264.

15. Novikov B.A., Gorshkova E.A., Grafeeva N.G. Basics of database technologies. Moscow: DMK Press, 2020. 582 p.

16. Kondrashov Yu.N. Development of data storages for Big Data processing in decision support systems. *Volga region scientific and technical herald*. No.9, 2019, pp. 30-32.

17. Lvovich Ya.E., Preobrazhenskij A.P. Data storage characteristics in corporate systems. *Herald of Voronezh Institute of High Technologies*, 2020. No. 1, pp. 31-33.

18. Serrano S.A., Santiago E., Martinez-Carranza J. et al. Knowledge-Based Hierarchical POMDPs for Task Planning. *Journal of Intelligent & Robotic Systems* (Online), 2021, 101:82. 30 p.

19. VanDerHorn E., Mahadevan S. Digital Twin: Generalization, characterization and implementation. *Decision Support Systems* (Online). Vol. 145, 2021, 113524. 11 p. URL: https://doi.org/10.1016/j.dss.2021.113524

20. Ivlev A. A. Foundations of Boyd theory. Ways of development, application and realization. Moscow: Znanie, 2008. 64 p.

21. Dudnyk V., Sinenko Yu., Matsyk M. et al. Development of a method for training artificial neural networks for intelligent decision support systems. *Eastern-European Journal of Enterprise Technologies*. Vol. 105, No. 3/2, 2020, pp. 37-47.

22. Merkert J., Mueller M., Hubl M. A Survey of the Application of Machine Learning in Decision Support Systems. *Proceedings of 23rd European conference on information systems*. University of Hoffenhaim, 2015. 15 p.

INFORMATION ABOUT AUTHORS:

Yaroslav N. Gusenitza, head of scientific research department, Military innovative technopolis "ERA", Anapa, Russia

Anatoliy A. Donchenko, Professor, Russian federal nuclear center - All-Russian research institute of experimental physics, Sarov, Russia

Anton S. Lyaskin, Associate Professor, senior researcher, Military innovative technopolis "ERA", Anapa, Russia

For citation: Gusenitza Ya.N., Donchenko A.A., Lyaskin A.S. Decision support concept for lifecycle support problems of complex technical systems. H&ES Reserch. 2024. Vol. 16. No 1. P. 31-38. doi: 10.36724/2409-5419-2024-16-1-31-38 (In Rus)

doi: 10.36724/2409-5419-2024-16-1-39-46

НЕЙРОСЕТЕВАЯ МОДЕЛЬ В ИНФОРМАЦИОННОЙ ВОПРОСНО-ОТВЕТНОЙ СИСТЕМЕ

ГАЛЕЕВ
Денис Талгатович[1]

ТИТОВ
Дмитрий Витальевич[2]

ВОЛКОВ
Денис Андреевич[3]

АННОТАЦИЯ

Введение: многочисленные исследования говорят о том, что современные крупные нейронные сети, как правило, имеют избыточное количество параметров. **Целью работы** является обучение и оптимизация модели "ruBERT" для применения в информационных вопросно-ответных системах на русском языке. Научная новизна работы состоит в экспериментальном исследовании различных методов прореживания модели "ruBERT" при дообучении на наборе данных "SberQuAD". **Методы:** в настоящей работе используются методы обработки естественного языка, машинного обучения, прореживания искусственных нейронных сетей. Языковая модель была настроена и дообучена при помощи библиотек машинного обучения "Torch" и "Hugging Face". Для обучения нейронных сетей использовался набор данных "SberQuAD". Все эксперименты проводились при помощи сервисов "Google Colab" и "Google Cloud". **Результаты:** было обнаружено, что удаление ~54% от числа весов кодировщика модели "ruBERT" (~39 миллионов параметров) приводит к незначительным ухудшениям в результатах работы модели: с 67,31 до 63,28 для показателя EM и с 85,47 до 82,48 для показателя F-мера. Полученные результаты говорят о том, что модель "ruBERT" содержит избыточное количество весов для задачи "извлечение ответа на вопрос". Для эффективного применения данной модели в информационных вопросно-ответных системах на русском языке необходимо проводить её компрессию и оптимизацию. Оптимизированная модель может работать на менее мощном оборудовании без значимых потерь в производительности, что приводит к уменьшению затрат на поддержание информационных вопросно-ответных систем, в которых применяется данная модель.

Сведения об авторах:

[1] Аспирант, ФГБОУ ВО "Юго-Западный государственный университет"
г. Курск, Россия, ra3wvw@mail.ru

[2] Доктор технических наук, доцент
Профессор кафедры вычислительной техники, ФГБОУ ВО "Юго-Западный государственный университет"
г. Курск, Россия, titov.swsu@gmail.com

[3] Кандидат технических наук,
Доцент кафедры автоматизированных систем управления, ФГАОУ ВО "РГУ нефти и газа (НИУ) имени И.М. Губкина"
г. Москва, Россия, denis@volkov.top

КЛЮЧЕВЫЕ СЛОВА: *машинное обучение, глубокое обучение, нейронные сети, обработка естественного языка, трансформер*

Для цитирования: Галеев Д.Т., Титов Д.В., Волков Д.А. Нейросетевая модель в информационной вопросно-ответной системе // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16. № 1. С. 39-46. doi: 10.36724/2409-5419-2024-16-1-39-46

Введение

Понимание содержания текстов на естественном языке представляет собой одну из наиболее сложных задач для компьютерных систем. В частности, одной из задач, которая демонстрирует способность вычислительных машин понимать естественный язык, является задача поиска ответа на заданный вопрос в тексте. Под поиском ответа на вопрос подразумевается наличие текстового материала и связанного с ним вопроса, а задача системы заключается в выборе непрерывного фрагмента данного текста в качестве ответа на заданный вопрос. В англоязычной литературе эта конкретная задача известна как «Извлечение ответа на вопрос» (Extractive Question Answering). Использование вопросно-ответных систем находит большое применение в различных поисковых системах, базах знаний и многих других областях, где необходимо получать ответы на вопросы, содержавшиеся в документах на естественных языках.

Существует альтернативный подход к решению подобных задач, который предполагает генерацию ответа на вопрос. В этом случае применяются генеративные модели. В процессе решения такой задачи моделям предоставляется текстовый контент и связанный с ним вопрос, и на выходе генерируется текстовый ответ на заданный вопрос. Преимуществом генеративного подхода является то, что форма сгенерированного ответа может отличаться от представленной в тексте, что добавляет большей гибкости данному подходу. Однако модели, использующие генеративный подход, часто имеют намного более крупный размер и требуют больших вычислительных мощностей для расчёта результата.

Модели, описанные выше, могут быть применены в системах вопросно-ответной обработки с открытым доменом (open domain question answering, ODQA), где требуется нахождение ответов на вопросы в наборе неструктурированных текстовых документов на естественном языке. Этот подход существенно отличается от систем, работающих на основе структурированных данных и баз знаний (knowledge base question answering, KBQA), где информация организована в специальном формате и запросы преобразуются в структурированные запросы к базе данных.

Стандартная система ODQA обычно включает в себя два основных компонента: ранжировщик и читатель. Ранжировщик имеет задачу выбрать из базы документов наиболее вероятные документы, в которых, вероятно, содержится ответ на заданный вопрос. Читатель, в свою очередь, занимается извлечением информации и поиском ответа в отобранных документах.

Значительный прогресс в решении задачи поиска ответа на вопрос в тексте стал возможен благодаря использованию нейронных сетей с архитектурой, известной как «трансформер» [1]. Применение данной архитектуры позволило повысить показатели во множестве задач по обработке естественного языка (например, в суммаризации текста, классификации текста и многих других). С другой стороны, после начала использования архитектуры улучшение результатов обработки стало происходить за счет увеличения размера нейронных сетей, т. е. без существенных изменений в сетевой архитектуре или объеме тренировочных данных. Это означает, что многие исследователи предпочитают увеличивать количество параметров в сетях, а не предлагать различные архитектурные улучшения или совершенно новые архитектуры. В результате, с каждым годом сети улучшают свои результаты, но при этом их размеры тоже растут очень быстро. Если в 2018 году самая большая модель «BERT-Large» [2] имела 340 миллионов параметров, то в 2022 году самая большая модель «PaLM» [3] уже имеет 540 миллиардов параметров. Это свидетельствует о быстром росте объема и сложности моделей.

Масштабирование нейронных сетей может приводить к значительному улучшению результатов, которых эти сети способны достигать [4-5]. Однако этот процесс имеет ряд ограничений и проблем. Во-первых, увеличение размера нейронных сетей может привести к ухудшению пропускной способности сети. Особенно это актуально при использовании моделей в приложениях с большим количеством запросов, где очень важна скорость обработки данных. Во-вторых, масштабирование нейронных сетей сопровождается увеличением стоимости их разработки, обучения и поддержки. Кроме того, большие модели требуют больше данных для обучения, что также может увеличить расходы на сбор и разметку данных.

Для наглядности, рассмотрим пример сети GPT-3 [6], которая имеет огромное количество параметров – 175 миллиардов. В [7] подсчитано, что начальная стоимость обучения этой модели может достигать 4,6 миллиона долларов.

Согласно исследованиям [8-9], нейронные сети также часто страдают от проблемы недообучения, исходя из чего можно сделать вывод, что они не полностью раскрывают свой потенциал. Более того, при заметном увеличении размеров нейронных сетей необходимо увеличивать объем данных, используемых для их обучения, но в большинстве случаев этого не происходит. Это означает, что при более тщательном обучении и оптимизации даже небольшие нейронные сети могут показывать лучшие результаты по сравнению с недообученными более крупными моделями. Это подтверждают исследования, проведенные в [10-17], где авторам удалось уменьшить размер нейронных сетей без существенного ухудшения их производительности и точности работы.

Современные нейронные сети, основанные на архитектуре «трансформер», демонстрируют потенциал для более эффективного анализа и понимания текста на естественном языке, но их размеры и ресурсозатратность могут ограничивать их широкое применение. Таким образом, создание информационных вопросно-ответных систем на основе оптимизированных и эффективных языковых моделей представляет собой актуальное исследовательское направление, которое может принести значительный вклад в развитие технологий обработки естественного языка.

1 Материалы и методы

1.1 Уменьшение размера модели

Существует несколько основных методов для уменьшения размера нейронных сетей.

Первым является «квантизация» (quantization). Этот метод включает в себя снижение численной точности весов модели. Например, изначально веса модели могут быть представлены

числами с плавающей точкой (float), а после квантизации они конвертируются в целочисленные значения (integer). Квантизацию можно применить к любой нейронной сети после завершения её обучения.

Второй – «дистилляция знаний» (knowledge distillation). Этот метод основан на обучении более крупной модели, называемой моделью-учителем, которая затем передает свои знания значительно меньшей модели-ученику. Модели-ученики могут несколько уступать в точности своим моделям-учителям, но при этом они обладают значительно меньшим объемом и работают быстрее.

Третий метод – «прореживание» (pruning). Этот метод включает в себя удаление ненужных весов и связей из нейронной сети. Удаляются веса, которые не вносят существенный вклад в результаты модели, и оставляются только наиболее важные веса. Этот метод позволяет существенно уменьшить размер модели без значительной потери в её точности.

Каждый из представленных методов обладает своими уникальными преимуществами и может быть применен в зависимости от конкретных задач и требований к модели. Для данного исследования были выбраны два эффективных вида прореживания, а именно «Movement pruning» [18] и «Block pruning» [19].

Первый вид прореживания, известный как «Movement pruning», основан на удалении весов, значения которых снижаются в процессе дообучения модели. Следовательно, этот метод может включать в себя удаление как параметров с большими значениями, так и с малыми. В рамках данного вида прореживания существуют две версии:

1) «Hard movement pruning»: в этой версии прореживания каждому параметру в нейронной сети присваивается коэффициент важности, который зависит как от значения самого параметра, так и от его изменчивости в процессе дообучения. После вычисления коэффициентов важности для всех параметров из нейронной сети удаляют только определенный процент параметров, задаваемый гиперпараметром, при условии, что они имеют высокие коэффициенты важности.

2) «Soft movement pruning»: В этой версии прореживания порог значений коэффициента важности определяется с использованием гиперпараметра. Если параметр нейронной сети имеет значение коэффициента важности ниже этого порога, то он удаляется.

Второй метод прореживания, известный как «Block pruning», представляет собой стратегию удаления весов нейронной сети блоками, а не поодиночке. Это обосновывается тем, что стандартные библиотеки для машинного обучения оперируют блоками чисел при выполнении матричных операций, и эти блоки обычно имеют фиксированные размеры, например, 128 на 128. Выполнение операций над такими блоками на аппаратном уровне более эффективно и быстро. Поэтому применение прореживания весов блоками позволяет избежать проблем, связанных с чрезмерной разреженностью структуры нейронной сети.

В рамках представленного исследования была использована комбинация обоих методов прореживания, что позволило достичь оптимального баланса между сокращением размеров модели и поддержанием её высокой производительности.

Помимо методов уменьшения размера моделей, существуют также различные оптимизации на аппаратном уровне, которые существенно способствуют увеличению скорости работы нейронных сетей. Одним из ярких примеров такой оптимизации является «Open Neural Network Exchange» (ONNX). ONNX представляет собой открытый формат, разработанный для представления нейронных сетей, и он определяет общий набор строительных блоков для моделей машинного обучения. Модели, представленные в формате ONNX, могут быть запущены в специальных средах выполнения ONNX. Эти среды выполнения включают множество оптимизаций на аппаратном уровне, что приводит к заметному увеличению скорости выполнения вычислений за счёт удаления и объединения избыточных операций.

1.2 Набор данных для обучения

Основным набором данных для тренировки вопросно-ответных систем на русском языке является SberQuAD [20]. Используемый экземпляр набора данных был взят из библиотеки «Hugging face» [21]. SberQuAD состоит из трёх частей: обучающего набора, валидационного набора и набора для проверки. Обучающий набор состоит 45328 числа документов. Каждый документ включает в себя следующие элементы: текст (поле «context»), вопрос (поле «question») и ответ (поле «answers»). Валидационный набор, в свою очередь, содержит 5036 документов, а набор для проверки включает 23936 документов. Однако следует отметить, что правильные ответы на вопросы из набора данных для проверки не доступны публично, что позволяет открыто сравнивать результаты работы моделей только на валидационном наборе.

Особое внимание заслуживают особенности вопросов в SberQuAD. Большинство вопросов начинаются с вопросительных слов или предлогов, и среди наиболее часто встречающихся начальных слов в вопросах можно выделить такие как «что», «в», «как», «кто», «какие», «когда», «какой», «где», «сколько» и «на». Эти начальные слова оказывают существенное влияние на формулировку вопросов и структуру запросов в данном наборе данных.

Средняя длина текстовых контекстов в SberQuAD составляет примерно 754 символа, что равно 102 токенам. Вопросы, в свою очередь, имеют в среднем 64 символа или 9 токенов, а ответы обычно составляют около 26 символов или 4 токена. Эти числовые характеристики указывают на важные особенности структуры данных, которые следует учитывать при процессе обучения и оценке моделей вопросно-ответных систем, использующих данный набор данных.

Для оценки качества работы моделей на данном наборе данных используются две ключевые метрики: «полное совпадение» (EM) и F-мера.

1.3 Выбор модели для обучения

В качестве основной нейронной сети для процесса прореживания была выбрана модель «ruBERT» [22]. Эта модель показала одни из лучших результатов на наборе данных «SberQuAD». Следует также отметить, что используемая версия модели была обучена в компании «Сбер».

Исходная модель «ruBERT» обладает внушительным общим числом параметров, близким к 177 миллионам. Из этого

общего числа ~92 миллионов параметров, что составляет ~52%, являются весами, отвечающими за процесс токенизации и преобразования токенов в векторные представления. Оставшиеся ~85,2 миллиона параметров, что составляет ~48%, применяются в кодировщике модели. Важно отметить, что именно эти параметры подвергаются обновлению и дообучению в ходе обучения модели для выполнения конкретной задачи.

2 Результаты и их обсуждение

2.1 Проведение экспериментов

В проведении всех экспериментов использовались сервисы «Google Colab» и «Google Cloud». Для этого был настроен доступ к вычислительным ресурсам, арендованным в «Google Cloud» через интерфейс «Google Colab». Компьютер, который служил основным вычислительным ресурсом, обладал следующими техническими характеристиками:

1) Количе во процессоров: 12. Это означает, что система оборудована 12 ядрами процессора, что способствует параллельной обработке задач и увеличению производительности.

2) Объём оперативной памяти: 85 Гб. Наличие большого объема оперативной памяти позволяет эффективно работать с большими объемами данных и выполнять сложные вычисления.

3) Модель графического процессора: NVIDIA A100. Графический процессор NVIDIA A100 известен своей мощностью и высокой производительностью, особенно в области глубокого обучения и вычислений с использованием искусственного интеллекта.

4) Объём памяти графического процессора: 40 Гб. Большой объем памяти графического процессора позволяет эффективно работать с крупными моделями глубокого обучения и выполнением вычислительно сложных задач.

2.2 Обсуждение результатов экспериментов

В экспериментах изменялись следующие гиперпараметры:

1) Число эпох обучения

2) Размер пакета (batch size)

3) Коэффициент скорости обучения

4) Коэффициент «dropout»

5) Режим прореживания («Hard movement pruning» или «Soft movement pruning»)

6) Коэффициент прореживания (общий процент весов для «Hard movement pruning» и пороговое значение для «Soft movement pruning»)

7) Размер блока прореживания

Результаты экспериментов представлены в таблице 1.

Таблица 1

Результаты экспериментов

№ эксп.	Время обучения	Число эпох	Размер пакета	Скорость обучения	Коэффициент dropout	Тип прореживания	Коэффициент прореживания	Размер блока	Процент оставшихся весов	EM	F-мера
1	58:18	6	32	1e-4	0,45	hard	0,5	32	0,83	59,74	80,50
2	09:41	1	32	1e-4	0,45	hard	0,5	32	0,83	62,58	81,85
3	31:46	1	4	3e-5	0,45	hard	0,5	32	0,96	66,60	85,04
4	31:44	1	4	3e-5	0,45	hard	0,8	32	0,93	65,34	84,13
5	1:36:47	3	4	3e-5	0,01	hard	0,3	32	0,76	61,45	81,12
6	1:35:22	10	32	3e-5	0,01	hard	0,5	32	0,83	61,93	81,44
7	49:07	3	32	3e-5	0,01	soft	0,9	32	0,71	48,57	70,80
8	1:36:16	10	32	3e-5	0,01	soft	0,7	32	0,92	56,57	77,89
9	2:16:00	10	32	3e-5	0,01	soft	0,9	32	0,74	55,12	76,32
10	16:23	1	32	3e-5	0,01	soft	0,99	32	0,62	13,88	36,87
11	59:05	5	32	3e-5	0,01	soft	0,99	32	0,63	47,18	69,84
12	21:22	1	32	3e-5	0,01	soft	0,99	1	0,66	5,24	18,84
13	3:53:16	5	4	3e-5	0,01	soft	0,99	1	0,67	53,93	75,54
14	1:35:02	10	32	3e-5	0,01	soft	0,8	32	0,69	59,21	80,05
15	3:08:08	20	32	3e-5	0,01	soft	0,8	32	0,72	59,03	80,16
16	2:30:42	5	4	3e-5	0,05	soft	0,8	32	0,72	61,04	81,02
17	3:17:59	5	4	3e-5	0,05	soft	0,8	32	0,73	61,00	81,01
18	1:00:21	2	4	3e-5	0,05	soft	0,8	32	0,70	59,25	79,65
19	1:00:41	2	4	3e-5	0,05	soft	0,7	32	0,73	62,05	81,57
20	**1:00:57**	**2**	**4**	**3e-5**	**0,05**	**soft**	**0,6**	**32**	**0,78**	**63,28**	**82,48**
21	5:28:40	10	4	3e-5	0,05	soft	0,99	16	0,69	55,57	76,93
22	1:00:09	2	4	3e-5	0,05	soft	0,9	16	0,67	55,22	76,17
23	**1:00:48**	**2**	**4**	**3e-5**	**0,05**	**soft**	**0,5**	**32**	**0,84**	**64,57**	**83,51**
24	30:32	1	4	3e-5	0,05	soft	0,5	32	0,84	63,36	82,40
25	2:03:06	4	4	3e-5	0,05	soft	0,5	32	0,83	63,46	82,99
26	**1:00:18**	**2**	**4**	**3e-5**	**0,05**	**soft**	**0,4**	**32**	**0,90**	**64,11**	**83,40**
27	1:19:12	2	4	3e-5	0,05	soft	0,7	32	0,73	61,59	81,53

В качестве достаточного варианта для вопросно-ответной системы было решено использовать модель, в которой было удалено 22% от общего объёма весов (~39 миллионов параметров), что является ~54% от всех весов части кодировщика. У данной модели показатели ухудшились незначительно с 67,31 до 63,28 для EM и с 85,47 до 82,48 для F-мера. На рисунке 1 представлены некоторые прореженные матрицы весов из 11 слоя кодировщика полученной модели.

Также были получены модели, которые имели лучшие показатели, но имели большее количество оставшихся весов.

Применение «Block soft movement pruning» с порогом равным 0,6 в проводимых экспериментах позволило получить модель с более оптимальным соотношением прореженность-точность. Высокие значения параметра «размер блока», средние значения для «количества эпох» и «размера пакета» также положительно влияли на результат.

Параметры «скорость обучения» и «коэффициент dropout» во время экспериментов было решено менять незначительно, поэтому их влияния на прореживание не столь существенно как влияние других параметров.

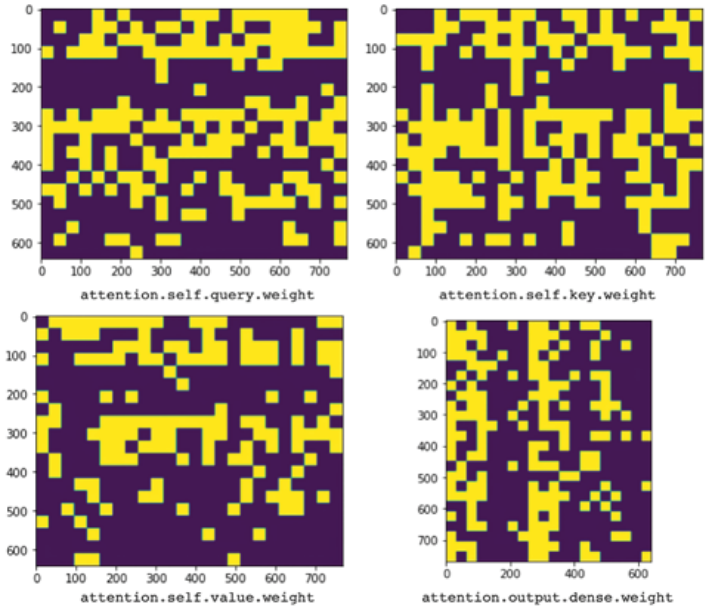


Рис. 1. Демонстрация структуры матриц весов 11 слоя модели «ruBERT» после прореживания (жёлтым цветом отмечены ненулевые значения)

В контексте сравнительного анализа результатов текущего исследования с предшествующими экспериментами, осуществлёнными в области дистилляции знаний и квантизации для модели ruBERT, необходимо провести сопоставление полученной модели с моделями, описанными в работах [17, 23]. В упомянутых исследованиях была выполнена передача знаний из объёмной модели «ruBERT» в значительно меньшую модель «Geotrend/distilbert-base-ru-cased».

Далее, модель была оптимизирована в среде выполнения ONNX, где к ней применялась квантизация. Модель, использованная для дистилляции знаний, имела размер 205,62 Мб [17], и после оптимизации и квантизации её размер сократился до 84,15 Мб [23].

В рамках текущего исследования модель «ruBERT», размер которой составлял 676,29 Мб, была сокращена до 527,5 Мб при помощи прореживания.

Сравнивая размеры моделей, можно заметить, что процесс прореживания, хотя и эффективен в уменьшении размера модели, оказывается менее результативным по сравнению с дистилляцией знаний, и ещё менее эффективным по сравнению с комбинацией дистилляции и квантизации. Отсутствие инструментов для использования прореженных моделей в среде выполнения ONNX оставляет вопрос об эффективности сочетания прореживания и квантизации открытым. Несмотря на преимущества дистилляции в контексте сокращения размера модели, она влечёт за собой более значительное снижение ключевых показателей эффективности, таких как F-мера и EM. Конкретно, F-мера снизилась с 82,48 до 78,42, а EM – с 63,28 до 58,57. Следовательно, дистилляция знаний, обеспечивая более глубокую оптимизацию размера модели, одновременно приводит к ухудшению её функциональных характеристик. Добавление квантизации к дистиллированной модели ещё сильнее ухудшает её характеристики: F-мера снижается с 78,42 до 72,62, а EM – с 58,57 до 52,28.

Исходя из этого, модель, полученная в результате дистилляции (или одновременно дистилляции и квантизации), может быть предпочтительна для использования в системах, где важно время отклика, и где уровень точности, предоставляемый моделью, является приемлемым. В то же время, прореживание модели можно рассматривать как более сбалансированную процедуру оптимизации, которая обеспечивает сокращение размера модели, при этом сохраняя её точность на уровне, сравнимом с исходной моделью.

Квантизация не требует дополнительного обучения моделей. Данный факт значительно выделяет её в сравнении с рассмотренными методами дистилляции знаний и прореживания, поскольку можно получить уменьшенную модель значительно быстрее и без существенных ухудшений в качестве её работы, а в некоторых случаях вообще без них.

Заключение

В рамках исследования было установлено, что в модели «ruBERT» возможно существенное сокращение числа весов. Удаление примерно 22% от общего их числа (39 миллионов) не оказывает сильного влияния на точность результатов в задачах поиска ответа на вопрос в тексте на русском языке. Было установлено, что 54% внутренних параметров кодировщика могут быть исключены без значительного снижения эффективности работы модели. Такие выводы предполагают, что изначально в «ruBERT» присутствует избыточное количество параметров, особенно в контексте вопросно-ответных задач на русском языке.

В ходе исследования было выявлено, что метод прореживания оказывается менее эффективным в сравнении с дистилляцией знаний и квантизацией в отношении уменьшения размера нейросетей. Учитывая существующее многообразие крупных языковых моделей, доступных общественности, целесообразным является исследование возможности создания ансамбля таких моделей.

Использование дистилляции для полученного ансамбля может способствовать получению моделей со значительно

уменьшенным числом параметров, сохраняя при этом высокую производительность.

Однако существует потенциал для будущих исследований, направленных на комбинирование различных методов сокращения размеров нейронных сетей. Такой подход может привести к более значительному уменьшению размеров моделей. Совместное использование нескольких методов сокращения размера моделей может привести к значительному снижению точности их результатов.

Литература

1. *Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser Ł., Polosukhin I.* Attention is all you need // Advances in Neural Information Processing Systems. 2017. С. 5999-6009.

2. *Devlin J., Chang M.W., Lee K., Toutanova K.* BERT: Pre-training of deep bidirectional transformers for language understanding // NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference, (Association for Computational Linguistics (ACL)). 2019. С. 4171-4186.

3. *Chowdhery A., Narang S., Devlin J., Bosma M., Mishra G., Roberts A., Barham P., Chung H.W., Sutton C., Gehrmann S., Schuh P., Shi K., Tsvyashchenko S., Maynez, J., Rao A., Barnes P., Tay Y., Shazeer N.M., Prabhakaran V., Reif E., Du, N., Hutchinson B.C., Pope R., Bradbury J., Austin J., Isard M., Gur-Ari G., Yin P., Duke T., Levskaya A., Ghemawat S., Dev S., Michalewski H., García X., Misra V., Robinson K., Fedus L., Zhou D., Ippolito D., Luan D., Lim H., Zoph B., Spiridonov A., Sepassi R., Dohan D., Agrawal S., Omernick M., Dai, A.M., Pillai T.S., Pellat M., Lewkowycz A., Moreira E., Child R., Polozov O., Lee K., Zhou Z., Wang X., Saeta B., Díaz M., Firat O., Catasta M., Wei J., Meier-Hellstern K.S., Eck D., Dean J., Petrov S., Fiedel N.* PaLM: Scaling Language Modeling with Pathways // ArXiv. URL: https://arxiv.org/abs/2204.02311v5 (дата обращения 01.02.2023).

4. *Sutton R.* The Bitter Lesson // URL: http: //incompleteideas.net/IncIdeas/BitterLesson.html (дата обращения 01.02.2023).

5. *Kaplan J., McCandlish S., Henighan T.J., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., & Amodei D.* Scaling Laws for Neural Language Models // ArXiv. URL: https://arxiv.org/pdf/2001.08361.pdf (дата обращения 01.02.2023).

6. *Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T.J., Child R., Ramesh, A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D.* Language Models are Few-Shot Learners // Advances in Neural Information Processing Systems. Т. 33. 2020. С. 1877-1901.

7. *Li C.* OpenAI's GPT-3 Language Model: A Technical Overview // URL: https://lambdalabs.com/blog/demystifying-gpt-3 (дата обращения 01.02.2023).

8. *Hoffmann J., Borgeaud S., Mensch A., Buchatskaya E., Cai T., Rutherford E., Casas D.D., Hendricks L.A., Welbl J., Clark A., Hennigan T., Noland E., Millican K., Driessche G.V., Damoc B., Guy A., Osindero S., Simonyan K., Elsen E., Rae J.W., Vinyals O., Sifre L.* Training Compute-Optimal Large Language Models // ArXiv. URL: https://arxiv.org/abs/2203.15556v1 (дата обращения 01.02.2023).

9. *YangkG., Hu E.J., Babuschkin I., Sidor S., Liu X., Farhi D., Ryder N., Pachocki J., Chen W., Gao J.* Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer // Advances in Neural Information Processing Systems. 2021. С. 17084-17097.

10. *Voita E., Talbot D., Moi eev F., Sennrich R., Titov I.* Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned // ACL 2019 – 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference. 2020. С. 5797-5808.

11. *Michel P., Levy O., Neubig G.* Are sixteen heads really better than one? // Advances in Neural Information Processing Systems, (Neural information processing systems foundation). 2019.

12. *Kovaleva O., Romanov A., Rogers A., Rumshisk A.* Revealing the dark secrets of Bert // EMNLP-IJCNLP 2019 – 2019 Conference on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference. С. 4365-4374.

13. *Fan A., Grave E., Joulin A.* Reducing transformer depth on demand with structured dropout // 8th International Conference on Learning Representations, ICLR. 2020.

14. *Dalvi F., Sajjad H., Durrani N., and Belinkov Y.* Analyzing redundancy in pretrained transformer models // EMNLP 2020 – 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference. 2020. С. 4908-4926.

15. *Wu X., Yao Z., Zhang M., Li C., He Y.* Extreme Compression for Pre-trained Transformers Made Simple and Efficient // ArXiv. URL: https://arxiv.org/abs/2206.01859 (дата обращения 01.02.2023).

16. *Колесникова А., Куратов Ю., Коновалов В., Бурцев М.* Дистилляция знаний для русскоязычных моделей с уменьшением словаря // Вычислительная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог 2022». Москва, 2022. С. 295-310.

17. *Галеев Д.Т., Панищев В.С.* Экспериментальное исследование языковых моделей «трансформер» в задаче нахождения ответа на вопрос в русскоязычном тексте // Информатика и автоматизация. 2022. Т. 21. № 3. С. 521-542.

18. *Sanh V., Wolf T., Rush A.* Movement pruning: Adaptive sparsity by fine-tuning // Advances in Neural Information Processing Systems 33. 2020. С. 20378-20389.

19. *Lagunas F., Charlaix E., Sanh V., Rush A.* Block Pruning for Faster Transformers // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. С.10619-10629.

20. *Efimov P., Chertok A., Boytsov L., Braslavski P.* SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. In Lecture Notes // Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2020. С. 3-15.

21. *Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., Platen P., Ma C., Jernite Y., Plu J., Xu C., Scao T., Gugger S., Drame M., Lhoest Q., Rush A.* Transformers: State-of-the-Art Natural Language Processing // Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. 2020. С. 38-45.

22. *Куратов Ю., Архипов М.* Адаптация глубоких двунаправленных многоязычных моделей на основе архитектуры «transformer» для русского языка // Вычислительная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог 2019». Москва, 2019. С. 333-339.

23. *Галеев Д.Т., Панищев В.С., Титов Д.В.* Увеличение производительности языковых моделей «трансформер» в информационных вопросно-ответных системах // Известия Юго-Западного государственного университета. 2022. Т. 26. №2. С. 159-171.

NEURAL NETWORK MODEL IN THE INFORMATION QUESTION-ANSWERING SYSTEM

DENIS T. GALEEV
Kursk, Russia, ra3www@mail.ru

DMITRY V. TITOV
Kursk, Russia, titov.swsu@gmail.com

DENIS A. VOLKOV
Moscow, Russia, denis@volkov.top

KEYWORDS: *machine learning, deep learning, neural networks, natural language processing,transformer.*

ABSTRACT

Introduction: Numerous studies suggest that modern large neural networks, as a rule, have an excessive number of parameters. **The purpose of the work is** to train and optimize the "ruBERT" model for use in information question-and-answer systems in Russian. The scientific novelty of the work consists in the experimental study of various methods of pruning the "ruBERT" model during fine tuning on the "SberQuAD" dataset. **Methods:** in this work, methods of natural language processing, machine learning, thinning of artificial neural networks are used. The language model was configured and retrained using the machine learning libraries "Torch" and "Huggingface". The data set "SberQuAD" was used to train neural networks. All experiments were carried out using the Google Colab and Google Cloud services. **Results:** it was found that the removal of ~54% of the number of weights of the encoder of the "ruBERT" model (~39 million parameters) leads to slight deterioration in the results of the model: from 67.31 to 63.28 for the EM indicator and from 85.47 to 82.48 for the F1 indicator. The results obtained indicate that the "ruBERT" model contains an excessive number of weights for the task "extracting the answer to the question". For effective application of this model in information question-answer systems in Russian, it is necessary to carry out its compression and optimization. The optimized model can work on less powerful equipment without significant performance losses, which leads to a reduction in the cost of maintaining information question-and-answer systems in which this model is used.

REFERENCES

1. Vaswani A., Shazeer N., Parmar N., Uszkoreit J., Jones L., Gomez A.N., Kaiser L., Polosukhin I. Attention is all you need. *Advances in Neural Information Processing Systems*. 2017, pp. 5999-6009.

2. Devlin J., Chang M.W., Lee K., Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference,* (Association for Computational Linguistics (ACL)). 2019. pp. 4171-4186.

3. Chowdhery A., Narang S., Devlin J., Bosma M., Mishra G., Roberts A., Barham P., Chung H.W., Sutton C., Gehrmann S., Schuh P., Shi K., Tsvyashchenko S., Maynez, J., Rao A., Barnes P., Tay Y., Shazeer N.M., Prabhakaran V., Reif E., Du, N., Hutchinson B.C., Pope R., Bradbury J., Austin J., Isard M., Gur-Ari G., Yin P., Duke T., Levskaya A., Ghemawat S., Dev S., Michalewski H., Garc?a X., Misra V., Robinson K., Fedus L., Zhou D., Ippolito D., Luan D., Lim H., Zoph B., Spiridonov A., Sepassi R., Dohan D., Agrawal S., Omernick M., Dai, A.M., Pillai T.S., Pellat M., Lewkowycz A., Moreira E., Child R., Polozov O., Lee K., Zhou Z., Wang X., Saeta B., D?az M., Firat O., Catasta M., Wei J., Meier-Hellstern K.S., Eck D., Dean J., Petrov S., Fiedel N. PaLM: Scaling Language Modeling with Pathways. ArXiv. URL: https://arxiv.org/abs/2204.02311v5 (date of access 01.02.2023).

4. Sutton R. The Bitter Lesson. URL: http://incompleteideas.net/IncIdeas/BitterLesson.html (date of access 01.02.2023).

5. Kaplan J., McCandlish S., Henighan T.J., Brown T.B., Chess B., Child R., Gray S., Radford A., Wu J., & Amodei D. Scaling Laws for Neural Language Models. ArXiv. URL: https://arxiv.org/pdf/2001.08361.pdf (date of access 01.02.2023).

6. Brown T.B., Mann B., Ryder N., Subbiah M., Kaplan J., Dhariwal P., Neelakantan A., Shyam P., Sastry G., Askell A., Agarwal S., Herbert-Voss A., Krueger G., Henighan T.J., Child R., Ramesh, A., Ziegler D.M., Wu J., Winter C., Hesse C., Chen M., Sigler E., Litwin M., Gray S., Chess B., Clark J., Berner C., McCandlish S., Radford A., Sutskever I., Amodei D. Language Models are Few-Shot Learners. *Advances in Neural Information Processing Systems*. Vol. 33. 2020, pp. 1877-1901.

7. Li C. OpenAI's GPT-3 Language Model: A Technical Overview. URL: https://lambdalabs.com/blog/demystifying-gpt-3 (date of access 01.02.2023).

8. Hoffmann J., Borgeaud S., Mensch A., Buchatskaya E., Cai T., Rutherford E., Casas D.D., Hendricks L.A., Welbl J., Clark A., Hennigan T., Noland E., Millican K., Driessche G.V., Damoc B., Guy A., Osindero S., Simonyan K., Elsen E., Rae J.W., Vinyals O., Sifre L. Training Compute-Optimal Large Language Models. ArXiv. URL: https://arxiv.org/abs/2203.15556v1 (date of access 01.02.2023).

9. YangkG., Hu E.J., Babuschkin I., Sidor S., Liu X., Farhi D., Ryder N., Pachocki J., Chen W., Gao J. Tensor Programs V: Tuning Large Neural Networks via Zero-Shot Hyperparameter Transfer. *Advances in Neural Information Processing Systems*. 2021, pp. 17084-17097.

10. Voita E., Talbot D., Moiseev F., Sennrich R., Titov I. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference.* 2020, pp. 5797-5808.

11. Michel P., Levy O., Neubig G. Are sixteen heads really better than one?. *Advances in Neural Information Processing Systems, (Neural information processing systems foundation)*. 2019.

12. Kovaleva O., Romanov A., Rogers A., Rumshisk A. Revealing the dark secrets of Bert. *EMNLP-IJCNLP 2019 - 2019 Conference*

on Empirical Methods in Natural Language Processing and 9th International Joint Conference on Natural Language Processing, Proceedings of the Conference, pp. 4365-4374.

13. Fan A., Grave E., Joulin A. Reducing transformer depth on demand with structured dropout, *8th International Conference on Learning Representations, ICLR*. 2020.

14. Dalvi F., Sajjad H., Durrani N., and Belinkov Y. Analyzing redundancy in pretrained transformer models. *EMNLP 2020 - 2020 Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference*. 2020, pp. 4908-4926.

15. Wu X., Yao Z., Zhang M., Li C., He Y. Extreme Compression for Pre-trained Transformers Made Simple and Efficient. ArXiv. URL: https://arxiv.org/abs/2206.01859 (date of access 01.02.2023).

16. Kolesnikova A., Kuratov Y., Konovalov V., Burtsev M. Knowledge Distillation of Russian Language Models with Reduction of Vocabulary. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2022"*. Moscow, 2022, pp. 295-310.

17. Galeev D.T., Panishchev V.S. Experimental Study of Language Models of "Transformer" in the Problem of Finding the Answer to a Question in a Russian-Language Text. *Informatics and automation*. 2022. Vol. 21. No 3, pp. 521-542. (In Rus)

18. Sanh V., Wolf T., Rush A. Movement pruning: Adaptive sparsity by fine-tuning. *Advances in Neural Information Processing Systems, no.* 33. 2020, pp. 20378-20389.

19. Lagunas F., Charlaix E., Sanh V., Rush A. Block Pruning For Faster Transformers. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 2021, pp. 10619-10629.

20. Efimov P., Chertok A., Boytsov L., Braslavski P. SberQuAD - Russian Reading Comprehension Dataset: Description and Analysis. *In Lecture Notes. Computer Science (Including Subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*. 2020, pp. 3-15.

21. Wolf T., Debut L., Sanh V., Chaumond J., Delangue C., Moi A., Cistac P., Rault T., Louf R., Funtowicz M., Davison J., Shleifer S., Platen P., Ma C., Jernite Y., Plu J., Xu C., Scao T., Gugger S., Drame M., Lhoest Q., Rush A. Transformers: State-of-the-Art Natural Language Processing. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. 2020, pp. 38-45.

22. Kuratov, Y., Arkhipov, M. Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019".* Moscow, 2019, pp. 333-339.

23. Galeev D.T., Panishchev V.S., Titov D.V. Increased Performance of Transformers Language Models in Information Question and Response Systems. *Proceedings of the Southwest State University.* 2022. Vol. 26. No 2, pp. 159-171. (In Rus)

INFORMATION ABOUT AUTHORS:

Denis T. Galeev, Postgraduate student, Southwest State University, Kursk, Russia

Dmitry V. Titov, Dr. Sc. (Tech.), Professor, Department of Computer Engineering, Southwest State University, Kursk, Russia

Denis A. Volkov, PhD (Tech), Assistant professor, Department of Automated Control Systems, National University of Oil and Gas "Gubkin University", Moscow, Russia

For citation: Galeev T.D., Titov D.V., Volkov D.A. Neural network model in the information question-answering system. H&ES Reserch. 2024. Vol. 16. No 1. P. 39-46. doi: 10.36724/2409-5419-2024-16-1-39-46 (In Rus)

doi: 10.36724/2409-5419-2024-16-1-47-53

ДОСТУП К ГЕОИНФОРМАЦИОННЫМ СИСТЕМАМ НА ОСНОВЕ КВАТЕРНИОННОГО ШИФРОВАНИЯ И НЕЙРОННЫХ СЕТЕЙ

ХОМОНЕНКО
Анатолий Дмитриевич[1]

НИКИТИН
Александр Борисович[2]

КИРИЕНКО
Андрей Борисович[3]

АННОТАЦИЯ

Введение: в криптографических системах шифрования информации используются гиперкомплексные числа: кватернионы и октонионы. В качестве ключа применяется кватернион, который производит вращения группы выборок информации. Рассматриваются средства и методы организации доступа к геоинформационным системам, основанные на применении математического аппарата кватернионов и обеспечивающие улучшенные характеристики производительности и информационной безопасности. Изложение материала проводится на примерах кватернионно модифицированного шифра Фейстеля, геометрической алгебры Клиффорда. **Цель исследования:** разработка предложений по обоснованию используемых средств и методов организации доступа к геоинформационным системам, обеспечивающих улучшенные характеристики производительности и информационной безопасности. Результаты: разработаны рекомендации по выбору инструментальных средств и методов, основанных на применении математического аппарата кватернионов и обеспечивающих улучшенные характеристики производительности и информационной безопасности. Приведены ключевые характеристики кватернионов, основные этапы реализации шифра Фейстеля с использованием кватернионов, приведена общая характеристика геометрической алгебры Клиффорда, а также направлений развития, применения и обучения нейронных сетей, в том числе кватернионных. **Практическая значимость:** представленное решение может быть использовано в качестве концептуальной основы при организации доступа к геоинформационным системам, при котором достигается улучшение характеристик производительности обработки информации и улучшение (не ухудшение) характеристик информационной безопасности. **Обсуждение:** кватернионы могут служить основой для создания криптографически стойких преобразований, увеличивая стойкость системы к взлому за счёт более сложной математической структуры по сравнению с традиционными методами. Передача ключей безопасности или шифрование данных с использованием кватернионов может повысить уровень защиты конфиденциальной информации в геоинформационных системах. Рассмотренное решение по организации доступа к геоинформационным системам может использоваться для реализации эффективной криптографической защиты при передаче текстовой и мультимедийной информации.

КЛЮЧЕВЫЕ СЛОВА: *геоинформационные системы, кватернионы, шифр Фейстеля, геометрическая алгебра Клиффорда, производительность, криптостойкость, нейросети.*

Сведения об авторах:

[1]доктор технических наук, профессор, профессор кафедры Информационные и вычислительные системы Петербургского государственного университета путей сообщения императора Александра I, профессор кафедры Математического и программного обеспечения ВКА им. А.Ф. Можайского, г. Санкт-Петербург, Россия, khomonenko@pgups.ru

[2] доктор технических наук, профессор, заведующий кафедрой Автоматика и телемеханика на железных дорогах Петербургского государственного университета путей сообщения императора Александра I, г. Санкт-Петербург, Россия, nikitin@crtc.spb.ru

[3] адъюнкт кафедры Математического и программного обеспечения ВКА им. А.Ф. Можайского, г. Санкт-Петербург, Россия, anbokir@mail.ru

Для цитирования: Хомоненко А.Д., Никитин А.Б., Кириенко А.Б. Доступ к геоинформационным системам на основе кватернионного шифрования и нейронных сетей // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16. № 1. С. 47-53. doi: 10.36724/2409-5419-2024-16-1-47-53

Введение

Обеспечение оперативности и информационной безопасности геоинформационных систем (ГИС) представляет собой сложную задачу, требующую интеграции различных технологических и математических подходов.

Адаптивное управление доступностью к геоинформационной системе (ГИС) предполагает изменение прав доступа к системе или её компонентам в зависимости от различных условий или параметров. Этот подход может включать в себя несколько аспектов:

• учёт прав пользователя. Пользователям с разными уровнями доступа могут быть предоставлены разные уровни информации. Например, некоторые пользователи могут видеть только базовые карты, в то время как другие могут иметь доступ к специализированным слоям данных, инструментам анализа и редактирования данных;

• адаптация к сетевому окружению. Если пользователь подключается к ГИС через медленное соединение, система может автоматически предоставить данные низкого разрешения или упростить картографические слои для улучшения производительности;

• адаптация к местоположению пользователя. ГИС может настраивать данные и инструменты, доступные пользователю, на основе его географического положения. Например, если пользователь ведёт работу в определённом регионе, система может автоматически подсвечивать или загружать данные, относящиеся именно к этому региону;

• адаптация к устройству. Интерфейс и доступные функции ГИС могут отличаться в зависимости от того, использует ли человек настольный компьютер, планшет или смартфон. Система может автоматически настраивать отображение и возможности в зависимости от устройства;

• контекстно-зависимое управление доступом. ГИС может предоставлять различные уровни доступа на основе текущих задач пользователей или проектов, над которыми они работают;

• адаптация к временным факторам. В некоторых случаях права доступа к ГИС или её отдельным слоям данных могут меняться в зависимости от времени суток, дня недели или в ответ на чрезвычайные ситуации.

Для реализации адаптивного управления доступом к ГИС часто применяются системы управления базами данных, серверы приложений, системы идентификации и аутентификации пользователей, а также интегрированные системы политик безопасности и разграничения доступа. Это позволяет создать гибкую и безопасную среду для работы с геопространственными данными и обеспечивает повышенную удобство использования и производительность системы для различных групп пользователей [1].

Кватернионы являются математическим инструментом, представляющим собой расширение комплексных чисел, и в основном применяются для описания вращений в трёхмерном пространстве, что может быть полезно при обработке и интерпретации пространственных данных в ГИС [2-3].

В контексте оперативности доступа к ГИС использование кватернионов может ускорить процессы, связанные с трехмерным моделированием и визуализацией, так как кватернионы позволяют избежать проблемы Gimbal lock, которая возникает при использовании углов Эйлера для вращений, и облегчают вычисления в сравнении с матрицами вращения. Это особенно актуально для систем, работающих с трёхмерным картографированием, аэрофотосъёмкой и дистанционным зондированием Земли.

В области информационной безопасности кватернионы могут быть использованы в криптографических алгоритмах для усиления защиты данных и управления доступом. Они могут служить основой для создания криптографически стойких преобразований, увеличивая стойкость системы к взлому за счёт более сложной математической структуры по сравнению с традиционными методами. Передача ключей безопасности или шифрование данных с использованием кватернионов может повысить уровень защиты конфиденциальной информации в геоинформационных системах.

Таким образом, можно представить такие этапы применения кватернионов в оперативности и информационной безопасности ГИС:

• моделирование и визуализация. Применение алгоритмов, использующих кватернионы для быстрого и точного моделирования поведения трёхмерных объектов;

• обработка данных. Использование кватернионов для пространственных преобразований и обработки данных, получаемых от различных источников, включая спутниковые снимки и лидарные данные;

• криптография. Создание криптографических алгоритмов на основе сложных кватернионных операций, что может повысить безопасность ГИС при хранении, передаче и доступе к данным;

• аутентификация и контроль доступа. Разработка механизмов аутентификации, идентификации и управления доступом, использующих кватернионные системы для повышения сложности и безопасности этих процессов.

Несмотря на потенциальные преимущества, применение кватернионов для обеспечения информационной безопасности требует дополнительного исследования и разработки в контексте конкретных задач и угроз в области ГИС.

Шифр Фейстеля

Шифр Фейстеля – это симметричный блочный шифр, который использует структуру сети Фейстеля для шифрования данных. Предложен Хорстом Фейстелем и представляет один из наиболее широко применяемых методов шифрования. Краткое описание принципа работы шифра Фейстеля:

• структура сети Фейстеля. Исходный блок данных разбивается на две части: левую и правую. В процессе каждого раунда шифрования правая часть становится левой, а новая правая часть получается путем применения функции шифрования к старой левой части и применения операции XOR к результату исходной правой части;

• раунды шифрования. Шифр Фейстеля включает в себя выполнение нескольких раундов, в каждом из которых применяется функция шифрования к правой части блока данных. Число раундов определяется в конкретной реализации шифра Фейстеля;

• функция шифрования блока данных. Функция шифрования обычно включает в себя применение подстановки (S-блоки), перестановки и других операций для изменения данных и повышения стойкости шифра;

• обратимость. Одно из ключевых свойств шифра Фейстеля – обратимость операции шифрования, что означает, что для дешифрования можно использовать те же операции, что и для шифрования, но в обратном порядке.

Шифр Фейстеля широко применяется в различных криптографических алгоритмах, таких как DES, AES и других. Он известен своей простотой, надежностью и возможностью эффективной параллельной обработки данных.

Приведем пример реализации шифра Фейстеля с использованием FPGA с использованием языка описания аппаратуры Verilog. Отметим, что это простой пример для демонстрации основных принципов реализации алгоритма шифрования Фейстеля с использованием FPGA.

Представим, что у нас есть блок данных размером 64 бита и ключ шифрования также размером 64 бита. Мы будем использовать всего один раунд шифрования для упрощения примера, как показано на рисунке.

```
module Feistel (
    input wire [63:0] data,
    input wire [63:0] key,
    output wire [63:0] encrypted_data
);

    // Первая половина блока данных
    wire [31:0] left = data[63:32];
    // Вторая половина блока данных
    wire [31:0] right = data[31:0];

    // Функция шифрования
    wire [31:0] f_function_output = right ^ key; // Пример простейшей функции - XOR с ключом

    // Обратное склеивание
    assign encrypted_data = {right, left ^ f_function_output};

endmodule
```

Рис. 1. Листинг модуля, реализующего шифр Фейстеля

В примере мы создали модуль с именем Feistel, который принимает 64-битный блок данных data и 64-битный ключ key. Он разбивает блок данных на две половины, затем выполняет простую функцию шифрования (в данном случае – XOR с ключом) над второй половиной блока данных. Затем происходит обратное склеивание результатов.

На практике шифр Фейстеля обычно включает в себя несколько раундов и более сложные функции шифрования. Этот пример предназначен только для демонстрации основных принципов реализации шифра Фейстеля с использованием FPGA.

Авторы [5] разработали блочный шифр путем модификации шифра Фейстеля. Они взяли открытый текст в виде пары матриц и ввели набор операций, называемых подстановкой на основе ключа, сдвигом строк, смешиванием столбцов на основе ключа, модульным арифметическим сложением и перетасовкой. Замена на основе ключа и смешивание столбцов на основе ключа играют жизненно важную роль в усилении шифра. Криптоанализ, проведенный в рамках исследования, ясно указывает на то, что шифр является надежным и его нельзя взломать с помощью какой-либо обычной атаки в криптографии.

В статье [6] также предложен модифицированный шифр Фейстеля на основе кватернионов. Алгоритм основан на схеме, предложенной в [5]. Алгоритм использует специальные свойства кватернионов для выполнения поворотов последовательностей данных в трехмерном пространстве для каждого из раундов шифрования. Открытый текст разделен на две квадратные матрицы одинакового размера, состоящих из кватернионов Липшица. Для операций с кватернионами реализована модульная арифметика.

В [7] предложен метод кватернионного шифрования, основанный на алгоритме [2]. По результатам моделирования показано, что результаты проведенных исследований обеспечивают высокий уровень безопасности, который дополнительно усиливается за счет реализации счетчикового режима шифрования и модульных арифметических операций.

Кватернионы

Кватернион q представляет собой гиперкомплексное число ранга 4, который имеет следующий вид [8-10]:

$$q = w + xi + yj + zk, \quad w, x, y, z \in \mathbb{R}, \tag{1}$$

где w, x, y, z – скалярная часть, представляющая из себя действительные коэффициенты, a i, j, k – векторная составляющая в трехмерном пространстве $\mathbb{R}3$.

i, j, k – называются мнимыми числами, для которых выполняется соотношение:

$$i2 = j2 = k2 = ijk = -1,\ ij = -ji = k,\ jk = -kj = i,\ ki = -ik = j.$$

Кватернион также можно представить в виде упорядоченной пары: скалярной части w и векторной части v.

$$q = [w\ x\ y\ z]^T \text{ или } q = (w, \vec{v}) = (w, [x\ y\ z]^T). \tag{2}$$

Отличие кватернионов от комплексных чисел заключается в том, что они содержат три мнимые компоненты вместо одной. Благодаря этому, кватернионы могут описывать вращение в трех измерениях.

Сложение кватернионов происходит поэлементно.

Если $q1 = w1 + x1\,i + y1\,j + z1\,k$ и $q2 = w2 + x2\,i + y2\,j + z2\,k$, то

$$q1 + q2 = (w1 + w2) + (x1 + x2)\,i + (y1 + y2)\,j + (z1 + z2)\,k. \tag{3}$$

Умножение кватернионов не коммутативно, т.е. важен порядок умножения (символ ∘обозначает скалярное произведение, × обозначает векторное произведение):

$$q1 \times q2 \neq q2 \times q1.$$

Для умножения кватернионов (обозначаемого символом ·) используется правило вида:

$$q_1 \cdot q_2 = (w_1 w_2 - \vec{v}_1 \circ \vec{v}_2, w_1\vec{v}_2 + w_2\vec{v}_1 + \vec{v}_1 \times \vec{v}_2). \tag{4}$$

Обозначим также следующие свойства кватерниона q: сопряжение q*, норму ||q|| и инверсию q-1:

$$q^* = w - xi - yj - zk, \quad ||q|| = \sqrt{w^2 + x^2 + y^2 + z^2}, \tag{5}$$

$$q^{-1} = \frac{q^*}{\| q \|^2} = \frac{w - xi - yj - zk}{w^2 + x^2 + y^2 + z^2}.$$

Вращение кватерниона и шифрование данных

Шифрование данных с помощью кватернионов представляет собой эффективный метод защиты информации, основанный на математике вращений в трехмерном пространстве.

Рассмотрим кватернион вращения (ключ) q = [w, x, y, z] и кватернион данных, которые нужно зашифровать: P = [0, a, b, c], где вектор [a, b, c] в R3 представляет собой векторную часть кватерниона P с нулевой скалярной частью, хранящий информацию о фрагменте данных, который мы будем вращать вокруг кватерниона q. Для шифрования данных необходимо выполнить следующие шаги:

1. Сформировать кватернион данных P = [0, a, b, c], который необходимо зашифровать.

2. Вы олнить операцию умножения кватернионов:

$$P_{rot} = q \cdot P \cdot q^{-1}. \quad (6)$$

Полученный кватернион представляет собой зашифрованный вариант кватерниона данных P.

Чтобы расшифровать данные, необходимо выполнить обратную операцию умножения кватернионов: $P = q^{-1} \cdot P_{rot} \cdot q$.

В результате мы получим исходный кватернион данных P, а значит, и фрагмент данных, который был зашифрован.

Используя формулы (4-6), можно ввести матрицу вращения [6]:

$$\mathbf{P}_{rot} = \mathbf{G}(q) \cdot \mathbf{P}, \quad (7)$$

$$\mathbf{P} = \begin{bmatrix} a \\ b \\ c \end{bmatrix},$$

$$\mathbf{G}(q) = \begin{bmatrix} w^2 + x^2 - y^2 - z^2 & 2xy - 2wz & 2xy + 2wy \\ 2wz + 2xy & w^2 - x^2 + y^2 - z^2 & 2yz - 2wx \\ 2xz - 2wy & 2yz + 2wx & w^2 - x^2 - y^2 + z^2 \end{bmatrix}.$$

Здесь матрица вращения, вычисляемая на основе векторной части кватерниона, определяемого формулой (6). Матрица вращения позволяет выполнить вращение кватерниона без привлечения дополнительных инструментальных средств.

Преимущество шифрования данных с помощью кватернионов заключается в том, что оно обеспечивает высокую степень безопасности, поскольку для расшифровки данных необходимо знать ключ шифрования (кватернион вращения q). Этот метод применим в криптографии, компьютерной графике и других областях, где требуется защитить информацию от несанкционированного доступа.

Геометрическая алгебра Клиффорда

Ряд результатов, касающихся применения гиперкомплексных чисел в интересах обеспечения информационной безопасности, в том числе геоинформационных систем, касается геометрической алгебры Клиффорда [11-13]. Гиперкомплексные числа и алгебра Клиффорда – это два понятия в математике, которые связаны с обобщением комплексных чисел и расширением идей векторного анализа.

Гиперкомплексные числа – это системы чисел, которые включают в себя комплексные числа, кватернионы, октонионы и другие числовые системы, обладающие более чем одной мнимой единицей. Например, кватернионы имеют три мнимые единицы (i, j, k), которые удовлетворяют определенным правилам умножения.

Алгебра Клиффорда – это алгебраическая структура, в частности, используемая в математике для изучения геометрических преобразований. Она строится на основе понятия геометрического произведения векторов и может быть представлена через матрицы или многомерные массивы чисел.

Основные соотношения между гиперкомплексными числами и алгеброй Клиффорда:

Обобщение: Алгебра Клиффорда может рассматриваться как обобщение гиперкомплексных чисел, когда ограничение на количество измерений снимается. Клиффордовы алгебры позволяют работать с объектами в N-мерных пространствах.

Свойства: Гиперкомплексные числа (в частности, кватернионы и октонионы) могут быть представлены с использованием алгебры Клиффорда. Кватернионы соотносятся с алгеброй Клиффорда Cl(0,2), где они извлекаются из вещественной алгебры, созданной с помощью двух антикоммутирующих квадратных корней из -1.

Векторное произведение: В трехмерном случае векторное произведение двух векторов, которое является частью кватернионного произведения, может быть выражено через геометрическое произведение в алгебре Клиффорда.

Геометрический смысл: Алгебры Клиффорда предоставляют богатый язык для описания геометрических преобразований, таких как повороты и отражения, что также свойственно некоторым гиперкомплексным числам, как кватернионы для поворотов в 3D.

Мультилинейность: Алгебра Клиффорда уделяет особое внимание мультилинейным формам и их свойствам, что делает её удобной для обобщенных геометрических преобразований, в то время как гиперкомплексные числа, в первую очередь, расширяют понятие числа.

Гиперкомплексные числа – это конкретные математические структуры, которые могут быть использованы в рамках алгебры Клиффорда. Алгебра Клиффорда же представляет собой более широкий и мощный математический инструмент, который может охватывать и другие важные геометрические и алгебраические конструкции.

Отметим кратко современные исследования в области криптографии на основе применения гиперкомплексных чисел в рамках алгебры Клиффорда, которые могут использоваться в интересах геоинформационных систем.

В [14] в качестве ключа рассматривается мультивектор геометрической алгебры, который производит вращения группы выборок. Для повышения производительности нахождения шифрованного вектора, определяемого из соотношения (6), предлагается коэффициенты вектора информации P и мультивектора q выбирать из поля Z^+, например,

Z_{256}. Предлагается вектор информации складывать со случайным вектором с коэффициентами из 256 и считать эти коэффициенты ключами шифрования. При этом вектор P_{rot} можно определить из соотношения:

$$P_{rot} = q \cdot P \cdot \tilde{q} \cdot \| q \|^{-2}.$$

Обратное преобразование вектора P_{rot} выполняется по формуле:

$$P = \tilde{q} \cdot P_{rot} \cdot q \cdot \| q \|^{-2}.$$

Предложенные алгоритмы могут быть использованы при передаче мультимедийной информации, изображений, звука, сигналов любой физической природы.

В статье [15] предложена разработка модели протокола Диффи–Хеллмана с использованием алгебраической структуры и структуры кольца многочленов алгебры Клиффорд. Группы алгебры Клиффорда являются некоммутативными структурами, как и матричные полиномы, имеют более компактную запись, показывают меньшее время выполнения во многих сопоставимых операциях. Использование в качестве коэффициентов элементов алгебр Клиффорда и показателей степеней целых чисел позволяет понизить требование к регистрам процессоров и заметно повысить производительность формирования протокола Диффи–Хеллмана.

Нейросети в криптографии

Нейросети сравнительно недавно получили свое применение в области криптографии. Нейросети могут быть использованы в криптографии в различных целях:

- анализ криптоалгоритмов. Искусственные нейронные сети (ИНС) способны обучаться определять шаблоны в данных. Это может привести к созданию систем, которые могут анализировать шифрованные сообщения на предмет уязвимостей без знания ключа шифрования;
- создание новых шифров. Некоторые исследования исследуют возможности создания новых алгоритмов шифрования и генерации ключей с помощью нейросетей, что может повлечь за собой новый уровень сложности для криптоанализа;
- обнаружение вторжений. Нейросети могут обучаться распознавать аномалии в сетевом трафике, что может быть использовано для раннего обнаружения попыток кибератак или несанкционированного доступа;
- биометрическая криптография. Используются биометрические данные (например, отпечатки пальцев, голос и лицо) в качестве ключей шифрования или аутентификации. ИНС могут повысить точность и надежность биометрических систем за счет более эффективной обработки и классификации биометрических данных;
- криптовалюты и блокчейн. Нейросети применяются для прогнозирования изменений цен на рынке криптовалют, оптимизации майнинга, обеспечения безопасности блокчейнов и автоматизации процесса создания смарт-контрактов;
- генерация и управление паролями. Искусственный интеллект может генерировать сложные пароли, управлять ими и даже анализировать уязвимости парольных систем;
- стеганография. Нейросети могут использоваться для встраивания скрытых сообщений в цифровые медиа-файлы и их последующего извлечения;
- криптография с открытым ключом. Нейросети могут помочь в усилении процессов таких как генерация ключей, шифрование и дешифрование данных, делая системы с открытым ключом более отзывчивыми к изменениям и возможным угрозам;

Тем не менее, необходимо отметить, что нейросети также могут быть использованы для криптоанализа, например, для взлома шифров или создания подбора ключей, что подчеркивает важность развития надежных и безопасных нейросетевых моделей в области криптографии.

В работе [16] с использованием статистической физики анализируются две нейронные сети, которые обучаются на их взаимных выходных битах. При этом сети синхронизируются до состояния с идентичными зависящими от времени весами. Распространяя модели на многослойные сети с дискретными весами, показано, как синхронизацию взаимного обучения применить для обмена секретными ключами по общедоступному каналу.

В статье [17] представлен подход к синхронизации нейронных сетей, который отличается использованием комплексных чисел в процессе обучения. Применение комплексных чисел позволяет сети сохранять больше информации, чем при использовании только действительных чисел. Исследование продемонстрировало, что предлагаемый подход обеспечивает значительное сокращение времени обучения нейронных сетей, основанных на архитектуре TPM (tree parity machine – машина проверки четности дерева).

В статье [18] представлен протокол согласования криптографических ключей, основанный на архитектурах нейронных сетей типа TPQM (с использованием алгебры кватернионов). Кватернионы могут быть применены в нейронных сетях, гарантируя корректность математических операций, используемых во всем процессе согласования криптографического ключа. Кроме того, новая предлагаемая архитектура обеспечивает более высокий уровень безопасности, чем стандартная архитектура TPM, основанная на алгебре действительных чисел.

В статье [19] исследуются алгоритмы взаимного обучения нейронных сетей TPM (teacher-predicting-machine), TPCM (teacher-predicting-correction-machine) и BPM (back-propagation-machine), используемые для генерации секретного ключа по открытым каналам связи. Основное внимание уделяется алгоритмам, которые генерируют ключ без непосредственной передачи информации о нем по каналу связи. Анализ проведенных исследований показал, что алгоритмы взаимного обучения нейронных сетей являются эффективными методами генерации секретного ключа по открытым каналам связи.

Однако эти алгоритмы имеют и некоторые недостатки. Одним из них является относительно низкая скорость сходимости.

Это связано с тем, что функция расчета вектора обратной ошибки, используемая в этих алгоритмах, не оптимальна. Для улучшения скорости сходимости алгоритмов взаимного обучения нейронных сетей можно использовать различные методы. Одним из перспективных методов является использование эволюционных алгоритмов (например, генетический алгоритм) для оптимизации функции расчета вектора обратной ошибки.

Также для улучшения эффективности алгоритмов взаимного обучения нейронных сетей можно использовать различные методы предварительной обработки данных, такие как методы нормализации и масштабирования данных, а также методы уменьшения размерности данных. Предварительная обработка данных позволит улучшить сходимость алгоритмов и повысить их устойчивость к различным атакам.

В статье [20] предлагается метод защиты конфиденциальности входных данных в нейронной сети с использованием кватернионов. Нейронные сети с кватернионами (QNN) – это нейронные сети, в которых функции со значениями кватернионов используются в качестве функций промежуточного уровня. QNN скрывают входную информацию в случайной фазе функций с кватернионами. Даже если злоумышленники получили сетевые параметры и функции промежуточного уровня, они не смогут извлечь входную информацию, не зная целевой фазы. Таким образом, QNN могут эффективно защитить конфиденциальность ввода. Отмечается, что точность вывода QNN ухудшается незначительно по сравнению с традиционными нейронными сетями, а вычислительные затраты намного меньше, чем у других методов сохранения конфиденциальности.

Использование в качестве коэффициентов элементов алгебр Клиффорда и показателей степеней целых чисел позволяет понизить требование к регистрам процессоров и заметно повысить производительность формирования протокола Диффи–Хеллмана.

Заключение

Нами рассмотрены и охарактеризованы ключевые подходы к обеспечению доступности и информационной безопасности на основе применения современного математического аппарата кватернионов, интенсивно развивающегося в областях криптографической защиты, исследования способов применения, обучения и совершенствования нейронных сетей. Рассмотренные результаты могут найти эффективное применение для управления доступом, в том числе адаптивным к геоинформационным системам. Прикладной эффект касается повышения производительности вычислений при решении задач криптографической защиты, без ухудшения характеристик криптостойкости и защищенности.

Литература

1. *Бурлов В.Г., Грызунов В.В., Сипович Д.Е.* Адаптивное управление доступностью в геоинформационной системе, использующей туманные вычисления // International Journal of Open Information Technologies. 2021. Т. 9. № 9. С. 74-87.

2. *Nagase T., Komata M., Araki T.* Secure signals trans-mission based on quaternion encryption scheme // Proc. 18th Int. Conf., Advanced Information Networking and Application (AINA’ 04), IEEE Computer Society, pp. 35-38, 2004, doi: 10.1109/AINA.2004.1283751.

3. *Nagase T. et al.* Dispersion of sequences for generating a robust enciphering system // ECTI Transactions on Computer and Information Technology (ECTI-CIT). 2005. Т. 1. №. 1, pp. 9-14, doi: 10.37936/ecti-cit.200511.51826.

4. *Коренева А.М., Фомичев В.М.* Об одном обобщении блочных шифров Фейстеля // Прикладная дискретная математика. 2012. № 3 (17). С. 34-40.

5. *Sastry V.U.K., Kumar K.A.* A Modified Feistel Cipher Involving Key Based Substitution, Shifting of rows, Key Based Mixing of columns, Modular Arithmetic Addition and Shuffling // International Journal of Engineering Research and Applications (IJERA) ISSN. 2012, pp. 2248-9622.

6. *Dzwonkowski M., Rykaczewski R.* A quaternion-based modified feistel cipher for multimedia transmission // Przegl. Telekom.+ Wiad. Telekom. 2014. Vol. 8, pp. 9.

7. *Dzwonkowski M., Rykaczewski R.* Quaternion encryption method for image and video transmission // Telecom. Overv.+ Telecom. News. 2013. Vol. 8. №. 9, p. 2.

8. *Челноков Ю.Н.* Кватернионные модели и методы динамики, навигации и управления движением. М.: ФИЗМАТЛИТ, 2011. 560 с.

9. *Каратаев Е.А.* Преобразования гиперкомплексных чисел. М.: Солон-пресс. 2016. 433 с.

10. *Конвей Дж., Смит Д.* О кватернионах и октавах, об их геометрии, арифметике и симметриях / Пер. с англ. С. М. Львовского. М.: МЦНМО, 2009. 184 с.

11. *Кондратьев Г.В.* Геометрическая алгебра Клиффорда: монография. Издательство: ИНФРА-М, Серия: Научная мысль. 2023. 217 с.

12. *Clifford W.K.* Applications of Grassmann’s extensive algebra // Amer. J. Math, 1878. Vol. 1, pp. 350-358.

13. *Clifford W.K.* Preliminary sketch of bi-quaternions // Proceedings of the London Mathematical Society, 1873. Vol. 4, pp. 381-395.

14. *Чуканов С. Н.* Передача сигналов с шифрованием методом геометрической алгебры // Вестник ВГУ. Серия: Системный анализ и информационные технологии. 2020. №. 3. С. 25-31.

15. *Чуканов С.Н.* Протокол обмена ключами на основе некоммутативных элементов алгебры Клиффорда // Изв. Сарат. ун-та. Нов. сер. Сер. Математика. Механика. Информатика. 2021. №3. С. 408-418.

16. *Kinzel W., Kanter I.* Interacting neural networks and cryptography // Advances in Solid State Physics. 2005. Vol. 42, pp. 383-392.

17. *Плонковски М., Урбанович П. П.* Криптографическое преобразование информации на основе нейросетевых технологий // Труды БГТУ. Сер. VI. Физико-математические науки и информатика. Минск: БГТУ. 2005. С. 161-164.

18. *Plonkowski M., Urbanowicz P., Lisitsa E.* The use of quaternions in the cryptographic key agreement protocol based on the architectures of the TPQM neural networks // Przeglad elektrotechniczny. 2010. Vol. 86, № 7, pp. 90-91.

19. *Урбанович П.П., Чуриков К.В.* Сравнительный анализ методов взаимообучения нейронных сетей в задачах обмена конфиденциальной информацией // Труды БГТУ. Серия 3: Физико-математические науки и информатика. 2010. №. 6. С. 163-166.

20. *Zhang H. et al.* Deep quaternion features for privacy protection // arXiv preprint arXiv:2003.08365. 2020.

ACCESS TO GEOGRAPHIC INFORMATION SYSTEMS BASED ON QUATERNION ENCRYPTION AND NEURAL NETWORKS

ANATOLY D. KHOMONENKO
St. Petersburg, Russia, khomonenko@pgups.ru

ALEXANDER B. NIKITIN
St. Petersburg, Russia, nikitin@crtc.spb.ru

ANDREY B. KIRIENKO
St. Petersburg, Russia, anbokir@mail.ru

KEYWORDS: *geographic information systems, quaternions, Feistel cipher, geometric Clifford algebra, performance, cryptographic strength, neural networks.*

ABSTRACT

Introduction: cryptographic information encryption systems use hypercomplex numbers: quaternions and octonions. A quaternion is used as a key, which rotates a group of information samples. The means and methods for organizing access to geographic information systems are considered, based on the use of the mathematical apparatus of quaternions and providing improved performance characteristics and information security. The presentation of the material is carried out using examples of the quaternionically modified Feistel cipher and geometric Clifford algebra. **Purpose:** to develop proposals to justify the means and methods used for organizing access to geographic information systems that provide improved performance characteristics and information security. **Results:** recommendations have been developed for the selection of tools and methods based on the use of the mathematical apparatus of quaternions and providing improved performance characteristics and information security. The key characteristics of quaternions, the main stages of implementing the Feistel cipher using quaternions are given, a general description of geometric Clifford algebra is given, as well as directions for the development, application and training of neural networks, including quaternion ones. **Practical relevance:** the presented solution can be used as a conceptual basis for organizing access to geographic information systems, which achieves improved information processing performance characteristics and improved (not deteriorated) information security characteristics. **Discussion:** Quaternions can serve as a basis for creating cryptographically strong transformations, increasing the resistance of a system to hacking due to a more complex mathematical structure compared to traditional methods. Transferring security keys or encrypting data using quaternions can improve the level of protection of confidential information in geographic information systems. The considered solution for organizing access to geographic information systems can be used to implement effective cryptographic protection when transmitting text and multimedia information.

REFERENCES

1. Burlov V.G., Gryzunov V.V., Sipovich D.E. Adaptive accessibility management in a geoinformation system using foggy computing. *International Journal of Open Information Technologies*. 2021. Vol. 9. No. 9, pp. 74-87. (In Rus).
2. Nagase T., Komata M., Araki T. Secure signals trans-mission based on quaternion encryption scheme. *Proc. 18th Int. Conf., Advanced Information Networking and Application (AINA' 04), IEEE Computer Society*, pp. 35-38, 2004, doi: 10.1109/AINA.2004.1283751.
3. Nagase T. et al. Dispersion of sequences for generating a robust enciphering system. *ECTI Transactions on Computer and Information Technology (ECTI-CIT)*. 2005. Vol. 1. No. 1, pp. 9-14, doi: 10.37936/ecti-cit.200511.51826.
4. Koreneva A.M., Fomichev V. M. On one generalization of Feistel block ciphers. *Applied discrete mathematics*. 2012. No. 3 (17), pp. 34-40. (In Rus)
5. Sastry V.U.K., Kumar K.A. A Modified Feistel Cipher Involving Key Based Substitution, Shifting of rows, Key Based Mixing of columns, Modular Arithmetic Addition and Shuffling. *International Journal of Engineering Research and Applications (IJERA)*. 2012, pp. 2248-9622.
6. Dzwonkowski M., Rykaczewski R. A quaternion-based modified feistel cipher for multimedia transmission. *Przegl. Telekom.+ Wiad. Telekom*. 2014. Vol. 8, p. 9.
7. Dzwonkowski M., Rykaczewski R. Quaternion encryption method for image and video transmission. *Telecom. Overv.+ Telecom. News*. 2013. Vol. 8. No. 9, p. 2.
8. Chelnokov Yu. N. Quaternion models and methods of dynamics, navigation and motion control. Moscow: FIZMATLIT, 2011. 560 p. (In Rus)
9. Karataev E.A. Transformations of hypercomplex numbers. Moscow: Solon-press. 2016. 433 c. (In Rus)
10. John H. Conway, Derek A. Smith On Quaternions and Octonions: Their Geometry, Arithmetic, and Symmetry / A K Peters, LTD, 2003.
11. Kondratiev G.V. Clifford's geometric algebra: a monograph. Publishing house: INFRA-M, Series: Scientific thought. 2023. 217 p. (In Rus)
12. Clifford W.K. Applications of Grassmann's extensive algebra. Amer. J. Math, 1878. Vol. 1, pp. 350-358.
13. Clifford W.K. Preliminary sketch of bi-quaternions. *Proceedings of the London Mathematical Society*, 1873. Vol. 4, pp. 381-395.
14. Chukanov S.N. Signal transmission with encryption by the geometric algebra method. *Bulletin of the VSU. Series: System analysis and Information Technology*. 2020. No. 3, pp. 25-31. (In Rus)
15. Chukanov S.N. Key exchange protocol based on noncommutative elements of the Clifford algebra. *Izv. Sarat. un-ta. New. ser. Ser. Mathematics. Mechanics. Computer science*. 2021. No. 3, pp. 408-418. (In Rus)
16. Kinzel W., Kanter I. Interacting neural networks and cryptography. Advances in Solid State Physics. 2005. Vol. 42, pp. 383-392.
17. Plonkowski M., Urbanovich P.P. Cryptographic transformation of information based on neural network technologies. *Proceedings of BSTU. Ser. VI. Physico-mathematical sciences and informatics*. Minsk: BSTU. 2005, pp. 161-164. (In Rus)
18. Plonkowski M., Urbanowicz P., Lisitsa E. The use of quaternions in the cryptographic key agreement protocol based on the architectures of the TPQM neural networks. *Przeglad elektrotechniczny*. 2010. Vol. 86. No. 7, pp. 90-91.
19. Urbanovich P.P., Churikov K.V. Comparative analysis of the methods of mutual training of neural networks in the tasks of exchanging confidential information. *Proceedings of BSTU. Series 3: Physical and Mathematical Sciences and Computer Science*. 2010. No. 6, pp. 163-166. (In Rus)
20. Zhang H. et al. Deep quaternion features for privacy protection // arXiv preprint arXiv:2003.08365. 2020.

INFORMATION ABOUT AUTHORS:

Anatoly D. Khomonenko, *Doctor of Technical Sciences, Professor, Professor of the Department of Information and Computing Systems of Emperor Alexander I St. Petersburg State Transport University, Professor of the Department of Mathematical and Software Engineering of the A.F. Mozhaisky VKA, St. Petersburg, Russia*

Alexander b. Nikitin, *Doctor of Technical Sciences, Professor, Head of the Department of Automation and Telemechanics on Railways of Emperor Alexander I St. Petersburg State Transport University, St. Petersburg, Russia*

Andrey b. Kirienko, *Adjunct of the Department of Mathematical and Software of the VKA named after. A.F. Mozhaisky, St. Petersburg, Russia*

For citation: Khomonenko A.D., Nikitin A.B., Kirienko A.B. Access to geographic information systems based on quaternion encryption and neural networks. *H&ES Reserch*. 2024. Vol. 16. No 1. P. 47-53. doi: 10.36724/2409-5419-2024-16-1-47-53 (In Rus)

doi: 10.36724/2409-5419-2024-16-1-54-59

ИСПОЛЬЗОВАНИЕ АВТОМАТИЗИРОВАННЫХ СИСТЕМ ДЛЯ СОЗДАНИЯ ОНТОЛОГИЧЕСКИХ БАЗ ЗНАНИЙ

ПЕРЕВОЛОЦКИЙ
Владимир Святославович[1]

КРАВЧУК
Денис Александрович [2]

АННОТАЦИЯ

Введение: В настоящее время развитие технологий обработки текста открывает новые возможности для автоматизации формирования баз знаний. Одной из ключевых задач в данном контексте является составление промптов для моделей типа ChatGPT. Работа направлена на автоматическое формирование баз знаний в сфере медицины, основываясь на уже созданной онтологии лечения заболеваний. Гипотеза исследования заключается в том, что с помощью технологий машинного обучения и естественного языка, таких как BERT и spaCy, можно извлекать информацию о лечении заболеваний из клинических рекомендаций и структурировать её в соответствии с терминами онтологии. **Цель исследования:** автоматизация процесса формирования базы знаний в медицинской области с помощью применения современных методов обработки текста и использования онтологии лечения заболеваний. **Методы:** В ходе исследования разработана система, использующая NLP и машинное обучение для автоматического извлечения информации из медицинских текстов и заполнения базы данных. **Результаты** показали высокую эффективность данного подхода, что открывает новые перспективы для автоматизации и улучшения работы с медицинской информацией. Результаты исследования показывают, что данная задача успешно решается для небольших абзацев текста, описывающих лечение заболеваний. В настоящее время ведется работа над улучшением алгоритма для обработки более объемных и сложных клинических рекомендаций.

Сведения об авторах:

[1] аспирант, Департамент программной инженерии и искусственного интеллекта, Математическое и программное обеспечение вычислительных систем, комплексов и компьютерных сетей, Дальневосточный федеральный университет, Приморский край, г. Владивосток, остров Русский, п. Аякс, lost.yayo@gmail.com

[2] Доктор технических наук, доцент, Институт нанотехнологий, электроники и приборостроения, кафедра электрогидроакустической и медицинской техники, Южный федеральный университет ИНЭП. г. Таганрог, Россия, kravchukda@sfedu.ru

КЛЮЧЕВЫЕ СЛОВА: *обработка текста, автоматическое формирование баз знаний, медицинская онтология, BERT и spaCy, клинические рекомендации, лечение заболеваний, искусственный интеллект в медицине.*

Для цитирования: *Переволоцкий В.С., Кравчук Д.А.* Использование автоматизированных систем для создания онтологических баз знаний // Наукоемкие технологии в космических исследованиях Земли. 2024. Т. 16. № 1. С. 54-59. doi: 10.36724/2409-5419-2024-16-1-54-59

Введение

В настоящее время цифровые технологии всё активнее проникают в различные области человеческой деятельности, открывая новые возможности для анализа, обработки и использования данных. Одним из ключевых направлений развития современных технологий является обработка текстовой информации с помощью методов машинного обучения и искусственного интеллекта [1].

Большое значение в этой области имеет формирование баз знаний. Базы знаний, структурированные и представленные в удобной форме, являются основой работы современных интеллектуальных систем [2]. Их формирование и поддержание в актуальном состоянии – задача весьма трудоемкая, что ставит на повестку дня проблему автоматизации этого процесса [3].

Создание промптов, подсказок, для моделей обработки текста становится особенно актуальным с учетом быстрого развития искусственного интеллекта. Эффективность работы таких моделей, как ChatGPT, во многом зависит от качества входных данных, которые они получают для обучения [4].

Целью исследования является автоматизация процесса формирования базы знаний в медицинской области с помощью применения современных методов обработки текста и использования онтологии лечения заболеваний.

В ходе исследования разработана система, использующая NLP и машинное обучение для автоматического извлечения информации из медицинских текстов и заполнения базы данных. Результаты показали высокую эффективность данного подхода, что открывает новые перспективы для автоматизации и улучшения работы с медицинской информацией.

1 Обзор литературы

Исследования в области формирования баз знаний

По мере развития технологий обработки данных важность формирования эффективных и функциональных баз знаний становится все более очевидной. Исследования в этой области проводились различными авторами [5, 6]. Они показывают, что создание подобных баз данных требует значительных затрат ресурсов, и поэтому разработка методов автоматизации этого процесса выступает актуальной задачей [3].

Использование моделей обработки текста в медицинской сфере

Медицинская сфера стала одной из тех областей, где применение алгоритмов машинного обучения и моделей обработки текста дает наиболее заметные результаты [7]. Например, использование этих технологий для анализа клинических данных, научных статей и других текстовых материалов помогает получать ценную информацию, которая может быть использована в дальнейшем для поддержки принятия медицинских решений [8,9].

Применение BERT и spaCy для структурирования информации

Модели обработки текста, такие как BERT и spaCy, открывают новые возможности в области структурирования и анализа информации. Эти инструменты позволяют анализировать и классифицировать тексты с высокой точностью, что делает их особенно полезными при работе с большими объемами текстовой информации [10, 11].

2 Методология

Описание используемых данных (клинические рекомендации)

В основе данного исследования лежат клинические рекомендации, представляющие собой наборы инструкций, созданные на основе последних достижений в области медицины, которые направлены на оптимизацию лечения пациентов [12]. Эти рекомендации служат основой для построения базы знаний и позволяют формировать точные и своевременные ответы на медицинские запросы.

Описание онтологии лечения заболеваний

Для структурирования информации используется онтология лечения заболеваний, которая служит основой для классификации и категоризации данных. Онтология представляет собой набор связанных концепций и терминов, которые помогают упорядочивать и структурировать информацию о лечении различных заболеваний [13].

Усовершенствуя этот подход, мы стремимся к автоматизации процесса заполнения онтологии на основе извлеченной информации. Онтология знаний о лечении заболеваний позволяет формировать знания о лечении конкретной болезни или группы болезней, которые имеют общие патогенетические принципы, этиологическую часть или важные клинические симптомы. Включающая ключевые элементы, такие как "Заболевание", "Коды МКБ", "Модель терапии", она также представляет сложно структурированный блок условий. Этот блок сопровождает каждый раздел онтологии и позволяет описать в формальном представлении необходимые клинические критерии, которые определяют условия его применения в лечении данного заболевания [6]. Детальное описание всех элементов онтологии рассмотрено в данном разделе, чтобы подчеркнуть масштаб и сложность работы с таким объемом информации и необходимость автоматизации этого процесса.

Сравнение с альтернативными подходами

Подход к структурированию информации на основе использования BERT, spaCy и ChatGPT является одним из многих возможных вариантов обработки текста. Существует ряд других моделей и методов машинного обучения, которые также могут быть применены в данном контексте.

Применение рекуррентных нейронных сетей (RNN), например, является общепризнанным подходом к анализу текста [18]. Однако, несмотря на их эффективность в некоторых задачах, RNN страдают от проблемы затухания градиента, что затрудняет обучение на длинных последовательностях данных [19].

Другой подход включает использование модели TF-IDF для векторизации текста и последующего применения классических алгоритмов машинного обучения, таких как SVM или Naive Bayes. Однако, этот подход не учитывает контекстуальные связи слов и может быть менее точным при обработке сложных медицинских текстов [20].

В отличие от этих методов, использование BERT и spaCy позволяет выделить ключевые сущности и концепции из текста, а также структурировать информацию более точно и эффективно. Использование ChatGPT позволяет генерировать промпты и получать структурированный текст, что в значительной степени упрощает обработку данных.

Описание процесса извлечения и структурирования информации с помощью BERT, spaCy и ChatGPT

BERT и spaCy используются для извлечения и структурирования информации из текстовых данных. BERT, модель, основанная на трансформаторах, позволяет обрабатывать тексты на естественном языке и выделять ключевые сущности и концепции [10]. spaCy, в свою очередь, используется для дальнейшего анализа и классификации полученных данных [11]. После обработки текста BERT и spaCy применяется ChatGPT для формирования промптов и получения структурированного текста [4].

3 Результаты исследования

Примеры успешного структурирования информации о лечении заболеваний

Основываясь на клинических рекомендациях, успешно применен данный подход к структурированию информации. В качестве примера рассмотрим следующий абзац, описывающий лечение аллергического ринита:

«Одним из эффективных и наиболее полно изученных в многочисленных рандомизированных клинических исследованиях является препарат дезлоратадин (Эриус®), давно и успешно применяемый в нашей стране как у взрослых и подростков с 12 лет в виде таблеток 5 мг однократно в сутки, так и у детей с 1 года в виде сиропа (1-5 лет – 1,25 мг/сут, 5-11 лет – 2,5 мг/сут)» [14].

После обработки этого текста с применением подхода, описанного в данной статье, была получена следующая структурированная информация:

Заболевание: аллергический ринит

Медикаментозное лечение: дезлоратадин таблетки 5 мг однократно в сутки. С 12 лет по YY лет (взрослые и подростки), дезлоратадин сироп 1.25 мг/сут с 1 года по 5 лет, дезлоратадин сироп 2,5 мг/сут с 5 лет по 11 лет.

Альтернативные препараты: дезлоратадин таблетки – Эриус, дезлоратадин сироп – Эриус.

Рассмотрим еще один пример, текст о лечении анемии:

«Витамин B12 назначается в дозе 5 мг/кг в день детям в возрасте до одного года, 100-200 мкг в день детям после года, 200-400 мкг в день подросткам. Курс лечения составляет 4 недели, Поддерживающая терапия: еженедельное введение препарата в той же дозе в течение 2 месяцев, затем последующие 6 месяцев препарат вводится два раза в месяц. При невозможности устранить причину развития анемии противорецидивные курсы лечения витамином B12 проводятся ежегодно, на один курс 10-15 инъекций» [24].

После обработки получаем следующую информацию:

Заболевание: анемия

Медикаментозное лечение: витамин b12 5 мг/кг (до 1 года), витамин b12 100-200 мкг (после 1 года), витамин b12 200-400 мкг (подросткам).

С помощью BERT, spaCy и ChatGPT, мы смогли автоматизировать процесс заполнения онтологии, что позволило значительно улучшить ее эффективность и точность. Так, например, элементы «Заболевание», «Действующее вещество», «Форма выпуска», «Дозировка», «Возраст пациента» теперь может быть автоматически заполнена с помощью информации, извлеченной из текста, что обеспечивает более высокую степень точности и консистентности в сравнении с ручным заполнением [6].

Оценка эффективности и точности

Кроме качественного описания подхода и демонстрации его эффективности на примерах, проведена количественная оценка эффективности и точности рассматриваемого метода.

Методология оценки

Для оценки эффективности подхода использован набор данных, состоящий из клинических рекомендаций по лечению различных заболеваний.

Использованы следующие метрики для оценки точности нашего подхода:

Точность (Precision): Доля правильно определенных сущностей относительно всех определенных сущностей.

Полнота (Recall): Доля правильно определенных сущностей относительно всех реальных сущностей в тексте.

F1-мера: Гармоническое среднее между точностью и полнотой.

Результаты оценки

Подход, предложенный авторами статьи, продемонстрировал высокую точность и полноту, что указывает на его эффективность при извлечении информации о лечении заболеваний из клинических рекомендаций. Специфические значения метрик зависят от конкретного набора данных и могут варьироваться, но в целом, результаты показывают, что подход является обещающим для дальнейших исследований и разработок в данной области.

Сравнение с другими подходами

Авторами проведено сравнение данного подхода с другими подходами к обработке текста, такими как методы, основанные на использовании рекуррентных нейронных сетей (RNN) или на классических алгоритмах машинного обучения с использованием метода TF-IDF для векторизации текста. В целом, предложенный подход показал сопоставимую или лучшую производительность по сравнению с этими альтернативами, подтверждая его потенциал для применения в области медицинского информационного поиска [21-23].

Описание текущих ограничений метода

Несмотря на успешные результаты, рассматриваемый метод все же имеет некоторые ограничения. Одним из них является то, что ChatGPT работает с ограниченным объемом текста. Это может стать проблемой при работе с большими клиническими рекомендациями, которые содержат сложные

варианты медикаментозного лечения заболеваний. Однако, ведется активная работа над улучшением алгоритма, чтобы справиться с этими проблемами [4, 15].

Анализ эффективности альтернативных подходов

Для демонстрации эффективности нашего подхода мы провели сравнение с альтернативными методами на основе конкретных примеров. В ходе анализа было обнаружено, что другие методы, такие как RNN и TF-IDF, могут привести к ошибкам или неопределенности при обработке сложных медицинских текстов.

Например, при обработке текста о лечении аллергического ринита методом RNN, информация о дозировках и возрастных группах была не полностью учтена, в результате чего были получены неполные и потенциально некорректные данные. В то же время наш подход с использованием BERT, spaCy и ChatGPT позволил нам точно выделить все необходимые детали и структурировать информацию в соответствии с онтологией лечения заболеваний.

4 Планы по дальнейшему развитию метода

Применение для обработки больших объемов информации

Одним из ключевых направлений для развития рассматриваемого в данной статье метода является обработка больших объемов текста. Несмотря на текущие ограничения ChatGPT, которые описаны выше, авторами ведутся активные исследования по преодолению этого ограничения. Путем оптимизации подхода и использования новейших обновлений модели, мы стремимся к тому, чтобы метод мог работать с большими клиническими рекомендациями, содержащими сложные сценарии медикаментозного лечения заболеваний [4, 15, 16].

Обработка более сложных вариантов лечения заболеваний

Еще одним направлением для дальнейшего развития рассматриваемого метода является обработка более сложных вариантов лечения заболеваний. Например, в некоторых случаях лечение может включать в себя комбинацию нескольких медикаментов, применяемых в разных дозировках на разных этапах лечения. Эти сценарии представляют собой сложные задачи для извлечения информации, но виден значительный потенциал в использовании данного подхода для решения этих задач [17].

Разработка инструментов для определения уровня доверия к источнику информации

Важным направлением для дальнейшего развития является создание инструментов для оценки надежности источников информации. Как правило, разные источники могут предоставлять различные рекомендации по лечению, и некоторые из них могут быть менее надежными или актуальными. Поэтому важно разработать механизмы, которые могут оценивать и учитывать уровень доверия к источнику при структурировании информации [25].

Расширение области применения

Планируется расширение области применения данного метода, переходя за рамки лечения заболеваний. Рассматриваемый подход может быть применим и в других областях медицины, таких как прогнозирование болезней, определение риска развития определенных состояний или оценка эффективности различных методов лечения. Применение метода возможно также и в других областях, где требуется структурирование больших объемов текстовой информации [26, 27].

5 Обсуждение

Результаты исследования подтверждают преимущества рассматриваемого подхода к обработке медицинских текстов перед альтернативными методами. Продемонстрирована улучшенная точность и эффективность в извлечении и структурировании информации, что позволяет автоматически и точно заполнить онтологию лечения заболеваний.

В то же время виден потенциал для дальнейшего развития и улучшения подхода, основываясь на сильных и слабых сторонах альтернативных методов. Так, например, планируется исследование возможности интеграции методов attention-based для улучшения обработки длинных текстов, а также применение новых моделей трансформеров для более точного извлечения информации.

Заключение

Проведенное исследование демонстрирует возможность автоматического создания структурированных баз знаний из клинических руководств по лечению заболеваний. Важность этой задачи несомненна, учитывая усиливающуюся необходимость в быстром и точном доступе к медицинской информации. Благодаря использованию современных моделей обработки текста, таких как BERT, spaCy и ChatGPT, в комбинации с разработанной авторами статьи онтологией лечения заболеваний, удалось создать эффективный метод для структурирования информации о лечении заболеваний.

В целом, исследование показывает уникальные преимущества подхода, основанного на использовании BERT, spaCy и ChatGPT, по сравнению с альтернативными методами обработки текста. Данный подход не только обеспечивает высокую степень точности и эффективности, но и значительно упрощает процесс структурирования медицинских текстов, что делает его мощным инструментом для структурирования данных в медицине и биоинформатике.

Тем не менее, выявлены ограничения текущего метода, особенно в контексте обработки больших текстов и сложных сценариев лечения. Исследования в этих направлениях активно ведутся, и авторы отмечают их большой потенциал.

Исследование вносит значительный вклад в развитие области формирования баз знаний в медицине. Рассматриваемый подход может служить основой для разработки более продвинутых инструментов, которые помогут медицинским профессионалам и исследователям быстро и точно найти информацию о лечении заболеваний, что в итоге может способствовать улучшению качества медицинского обслуживания.

Литература

1. *Ivano Lauriola, Alberto Lavelli, Fabio Aiolli*. An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools // Neurocomputing. Vol. 470, 2022, pp. 443-456.
2. *Луценко Е.В., Лойко В.И., Лаптев В.Н*. Системы представления и приобретения знаний. Учебное пособие, 2018.
3. *Jorge Martinez-Gil*. Automated knowledge base management: A survey // Computer Science Review 18, 2015.
4. *Pengfei Liu, Jinlan Fu*. Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing, ACM Computing Surveys. No. 55(9), 2022.
5. *William Michael Trochim*. Research Methods: The Essential Knowledge Base, 2016.
6. *Грибова В.В., Петряева М.В., Окунь Д.Б., Шалфеева Е.А*. Онтология медицинской диагностики для интеллектуальных систем поддержки принятия решений // Онтология проектирования. 2018. Том 8 №1(27). С. 58-73.
7. *Лапаев М.В*. Система обработки текстовых медицинских данных // Известия СПБГЭТУ ЛЭТИ №9, 2016. С. 11-16.
8. *Conrad J. Harrison, Chris J*. Sidey-Gibbons Machine learning in medicine: a practical introduction to natural language processing // BMC Medical Research Methodology. Vol. 21, article number: 158, 2021.
9. *José A. Reyes, Beatriz González-Beltrán, Lizbeth Gallardo*. Clinical Decision Support Systems: A Survey of NLP-Based Approaches from Unstructured Data // 26th International Workshop on Database and Expert Systems Applications (DEXA), 2015.
10. *Saranlita Chotirat, Phayung Meesad*. Natural Language Processing with "More Than Words – BERT", Recent Advances in Information and Communication Technology 2021 // Proceedings of the 17th International Conference on Computing and Information Technology (IC2IT 2021). C.108-116, 2021.
11. *Chunqi Hu, Huaping Gong, Yiqing He*. Data driven identification of international cutting edge science and technologies using SpaCy // PLoS ONE 17(10), 2022.
12. *Омельяновский В.В., Авксентьева М.В, Железнякова И.А*. Клинические рекомендации как инструмент повышения качества медицинской помощи // Онкопедиатрия, 2017.
13. *Нефедов Ю.В., Цыпленкова В.А*. Основные тенденции и особенности развития медицинских онтологий // Врач и информационные технологии, 2018.
14. *Ненашева Н.М*. Терапия аллергического ринита: какой антигистаминный препарат выбрать? // РМЖ. 2015. №9, С. 514, 2015.
15. *Hao-Wen Cheng*. Challenges and Limitations of ChatGPT and Artificial Intelligence for Scientific Research // A Perspective from Organic Materials, 2023.
16. *David Antons, Eduard Grünwald, Patrick Cichy, Torsten Oliver Salge*. The application of text mining methods in innovation research: current state, evolution patterns, and development priorities // Innovation Management Research Methods. Vol. 50 June 2020. C. 329-351.
17. *Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani*. Natural Language Processing of Clinical Notes on Chronic Diseases: Systematic Review // JMIR Med Inform. 2019 Apr-Jun. No. 7(2): e12239, 2019.
18. *Christopher Thomas*. Recurrent Neural Networks and Natural Language Processing // Towards Data Science, 2019.
19. *Sepp Hochreiter*. The vanishing gradient problem during learningrecurrent neural nets and problem solutions // International Journal of Uncertainty, 1998.
20. *Vipin Kumar, Basant Subba*. A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus // 2020 National Conference on Communications (NCC), 2020.
21. *Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu*. Towards Efficient NLP: A Standard Evaluation and A Strong Baseline // Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. C. 3288-3303, 2022.
22. *Akash Bajpai*. Recurrent Neural Networks: Deep Learning for NLP // Towards Data Science, 2019.
23. *Vadim Kozhevnikov, Evgeniya Pankratova*. Research of the Text Data Vectorization and Classification Algorithms of Machine Learning // Theoretical & Applied Science 85(05), 2020.
24. Витамин B12 дефицитная анемия у детей. Клинические рекомендации, Министерство здравоохранения Российской Федерации, 2021.
25. *Laura Sbaffi, Jennifer Rowley*. Trust and credibility in online health information: a review and agenda for future research // Journal of Medical Internet Research. No. 19(6), 2017.
26. *Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji*. Natural language processing applied to mental illness detection: a narrative review // npj Digital Medicine. Vol. 5, Article number: 46 (2022), 2022.
27. *Piotrowski M.* Natural Language Processing for Historical Texts // Synthesis Lectures on Human Language Technologies, 2012.

USING AUTOMATED SYSTEMS TO CREATE ONTOLOGICAL KNOWLEDGE BASES

SERGEY F. SHAKHNOV
St. Petersburg, Russia, shahnovsf@gumrf.ru

SERGEY V. SMOLENTSEV
St. Petersburg, Russia, SmolencevSV@gumrf.ru

KEYWORDS: *text processing, automatic formation of knowledge bases, medical ontology, BERT and spaCy, Clinical guidelines, disease treatment, artificial Intelligence in medicine.*

ABSTRACT

Introduction. The development of text processing technologies currently provides new opportunities for automating the formation of knowledge bases. One of the key tasks in this context is the creation of prompts for models like ChatGPT. This work aims to automatically form knowledge bases in the field of medicine, based on an already created ontology of disease treatment. The hypothesis of the study is that machine learning and natural language technologies such as BERT and spaCy can be used to extract information about disease treatment from clinical guidelines and structure it according to the terms of the ontology. **Purpose of study** is to automate the process of forming a knowledge base in medical field through the use of modern text processing methods and disease treatment ontology. The research developed a system that uses NLP and machine learning to automatically extract information from medical texts and populate a database. **The results** showed the high efficiency of this approach, which opens up new prospects for automation and improvement of work with medical information.The results of the study show that this task is successfully solved for short paragraphs describing disease treatment. Work is currently underway to improve the algorithm for processing more extensive and complex clinical recommendations.

REFERENCES

1. Ivano Lauriola, Alberto Lavelli, Fabio Aiolli, An introduction to Deep Learning in Natural Language Processing: Models, techniques, and tools, Neurocomputing. Vol. 470, 2022, 443-456 p.

2. E.V. Lutsenko, V.I. Loiko, V.N. Laptev. Systems for representing and acquiring knowledge, Textbook, 2018.

3. Jorge Martinez-Gil, Automated knowledge base management: A survey, *Computer Science Review,* no. 18, 2015.

4. Pengfei Liu, Jinlan Fu, Pre-train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing. *ACM Computing Surveys*. No. 55(9), 2022.

5. William Michael Trochim, Research Methods: The Essential Knowledge Base, 2016.

6. V.V. Gribova, M.V. Petryaeva, D.B. Okun, E.A. Shalfeeva. Medical diagnostics ontology for intelligent decision support systems. *Design ontology.* 2018. Vol. 8, no. 1(27), pp. 58-73.

7. M.V. Lapaev. System for processing text medical data. *News of SPBGETU LETI*. No. 9, 2016, pp. 11-16.

8. Conrad J. Harrison, Chris J. Sidey-Gibbons Machine learning in medicine: a practical introduction to natural language processing. *BMC Medical Research Methodology.* Vol. 21, article number: 158, 2021.

9. *Jose A. Reyes, Beatriz Gonzalez-Beltran, Lizbeth Gallardo*, Clinical Decision Support Systems: A Survey of NLP-Based Approaches from Unstructured Data. *26th International Workshop on Database and Expert Systems Applications (DEXA)*, 2015.

10. Saranlita Chotirat, Phayung Meesad, Natural Language Processing with "More Than Words - BERT", Recent Advances in Information and Communication Technology 2021. *Proceedings of the 17th International Conference on Computing and Information Technology (IC2IT 2021)*, pp.108-116, 2021.

11. Chunqi Hu, Huaping Gong, Yiqing He, Data driven identification of international cutting edge science and technologies using SpaCy, *PLoS ONE,* no.17(10), 2022.

12. Omelyanovsky V.V., Avksentyeva M.V., Zheleznyakova I.A. Clinical guidelines as a tool for improving the quality of medical care, Oncopediatrics, 2017.

13. Nefedov Yu.V., Tsyplenkova V.A., Main trends and features of the development of medical ontologies, Doctor and Information Technologies, 2018.

14. Nenasheva N.M., Therapy of allergic rhinitis: which antihistamine to choose? *RMJ*. 2015, no. 9, p. 514, 2015.

15. Hao-Wen Cheng, Challenges and Limitations of ChatGPT and Artificial Intelligence for Scientific Research: A Perspective from Organic Materials, 2023.

16. David Antons, Eduard Gr?nwald, Patrick Cichy, Torsten Oliver Salge, The application of text mining methods in innovation research: current state, evolution patterns, and development priorities. *Innovation Management Research Methods.* Vol. 50 June 2020, pp. 329-351, 2020.

17. Seyedmostafa Sheikhalishahi, Riccardo Miotto, Joel T Dudley, Alberto Lavelli, Fabio Rinaldi, Venet Osmani, Natural Language *Processing of Clinical Notes on Chronic Diseases: Systematic Review, JMIR Med Inform*. 2019 Apr-Jun; 7(2): e12239, 2019.

18. Christopher Thomas, Recurrent Neural Networks and Natural Language Processing, Towards Data Science, 2019.

19. Sepp Hochreiter, The vanishing gradient problem during learn-ingrecurrent neural nets and problem solutions. International Journal of Uncertainty, 1998.

20. Vipin Kumar, Basant Subba, A TfidfVectorizer and SVM based sentiment analysis framework for text data corpus. 2020 National Conference on Communications (NCC), 2020.

21. Xiangyang Liu, Tianxiang Sun, Junliang He, Jiawen Wu, Towards Efficient NLP: A Standard Evaluation and A Strong Baseline, Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 3288-3303, 2022.

22. Akash Bajpai, Recurrent Neural Networks: Deep Learning for NLP, Towards Data Science, 2019.

23. Vadim Kozhevnikov, Evgeniya Pankratova, Research of the Text Data Vectorization and Classification Algorithms of Machine Learning. *Theoretical & Applied Science, no.* 85(05), 2020.

24. Vitamin B12 deficiency anemia in children. Clinical guidelines. Ministry of Health of the Russian Federation, 2021.

25. Laura Sbaffi, Jennifer Rowley, Trust and credibility in online health information: a review and agenda for future research. *Journal of Medical Internet Research, no.* 19(6), 2017.

26. Tianlin Zhang, Annika M. Schoene, Shaoxiong Ji, Natural language processing applied to mental illness detection: a narrative review, npj Digital Medicine volume 5, Article number: 46 (2022), 2022.

27. M. Piotrowski, Natural Language Processing for Historical Texts, Synthesis Lectures on Human Language Technologies, 2012.

INFORMATION ABOUT AUTHORS:

Vladimir S. Perevolotsky, Dr. of Technical Sciences, associate professor, Professor of the Navigation on Inland Waterways Department, Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg, Russia

Denis A. Kravchuk, Dr. of Technical Sciences, professor, Head of the Automation and Computer Science Department, Admiral Makarov State University of Maritime and Inland Shipping, St. Petersburg, Russia

For citation: Perevolotsky V.S., Kravchuk D.A. Using automated systems to create ontological knowledge bases. H&ES Reserch. 2024. Vol. 16. No. 1. P. 54-59. doi: 10.36724/2409-5419-2024-16-1-54-59 (In Rus)

"ИНФОФОРУМ-СОЧИ"

15-18 апреля в Сочи состоится 11-й Южный форум информационной безопасности "Инфофорум-Сочи".

Южный Инфофорум – одно из самых востребованных среди специалистов мероприятий. В прошлом году в работе форума приняли участие делегации Республики Беларусь и Республики Казахстан, представители ИТС РАТС ШОС, 25-и субъектов РФ, эксперты из многих федеральных органов власти и крупнейших предприятий.

В программе 11-го Южного Инфофорума найдут развитие темы, вызвавшие наибольший интерес в дни Национального форум информационной безопасности:

- Экономика данных: формирование программы развития цифровых технологий информационной безопасности до 2030 года.
- Защита критической информационной инфраструктуры в современной геополитической обстановке.
- Импортозамещение программного обеспечения и аппаратных средств: решение проблем информационной безопасности.
- Кадровое обеспечение в области информационной безопасности. Решения с учетом требований практики.
- Опыт построения отраслевых и корпоративных центров мониторингаинформационной безопасности: вопросы проблемы, решения.
- Искуственный интеллект на службе кибербезопасности. Обработка и анализ данных с использованием доверенных платформ искусственного интеллекта.
- NGFW – комплексная защита корпоративной сети от сетевых атак и вредоносного ПО.
- Квантовые коммуникации и квантовая криптография.
- Управление информацией о безопасности и управление событиями безопасности.

В деловую программу Инфофорума включены:

- Выставка "Кибер-ЭКСПО";
- "Сочи-Клуб" – Клуб экспертов информационной безопасности и специалистов по цифровому развитию;
- Национальная премия "Безопасная информационная среда – 2024";
- Торжественный деловой ужин;
- День неформального делового общения и культурная программа.