ISSN 2518-1092

НАУЧНЫЙ РЕЗУЛЬТАТ

ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

RESEARCH RESULT. INFORMATION TECHNOLOGY

8(3) 2023

16+

Сайт журнала:
rrinformation.ru
сетевой научный рецензируемый журнал
online scholarly peer-reviewed journal

НАУЧНЫЙ РЕЗУЛЬТАТ
RESEARCH RESULT

НАУЧНЫЙ РЕЗУЛЬТАТ. ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ
RESEARCH RESULT. INFORMATION TECHNOLOGIES

Журнал зарегистрирован в Федеральной службе по надзору в сфере связи, информационных технологий и массовых коммуникаций (Роскомнадзор)
Свидетельство о регистрации средства массовой информации Эл. № ФС77-69101 от 14 марта 2017 г.

The journal has been registered at the Federal service for supervision of communications information technology and mass media (Roskomnadzor)
Mass media registration certificate El. № FS 77-69101 of March 14, 2017

Том 8, № 3. 2023

СЕТЕВОЙ НАУЧНО-ПРАКТИЧЕСКИЙ ЖУРНАЛ
Издается с 2016 г.
ISSN 2518-1092

Volume 8, № 3. 2023

ONLINESCHOLARLYPEER-REVIEWEDJOURNAL
First published online: 2016
ISSN 2518-1092

РЕДАКЦИОННАЯ КОЛЛЕГИЯ:

ГЛАВНЫЙ РЕДАКТОР: ***Черноморец А.А.,*** *доктор технических наук, профессор кафедры прикладной информатики и информационных технологий Белгородского государственного национального исследовательского университета.*

ЗАМЕСТИТЕЛЬ ГЛАВНОГО РЕДАКТОРА: ***Жихарев А.Г.,*** *доктор технических наук, доцент кафедры информационных и робототехнических систем Белгородского государственного национального исследовательского университета.*

ОТВЕТСТВЕННЫЙ СЕКРЕТАРЬ: ***Болгова Е.В.,*** *кандидат технических наук, доцент кафедры прикладной информатики и информационных технологий Белгородского государственного национального исследовательского университета.*

РЕДАКТОР АНГЛИЙСКИХ ТЕКСТОВ СЕРИИ: ***Ляшенко И.В.,*** *кандидат филологических наук, доцент*

ЧЛЕНЫ РЕДАКЦИОННОЙ КОЛЛЕГИИ:

Басов О.О., *доктор технических наук (Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО), г. Санкт-Петербург)*

Белов С.П., *доктор технических наук, профессор (Белгородский государственный национальный исследовательский университет, г. Белгород)*

Волчков В.П., *доктор технических наук, профессор (Московский технический университет связи и информатики, г. Москва)*

Дмитриенко В.Д., *доктор технических наук, профессор (Харьковский национальный технический университет «ХПИ», г. Харьков, Украина)*

Иващук О.А., *доктор технических наук, профессор (Белгородский государственный национальный исследовательский университет, г. Белгород)*

Калмыков И.А., *доктор технических наук, профессор (Северо-Кавказский федеральный университет, г. Ставрополь)*

Корсунов Н.И., *заслуженный деятель науки РФ, доктор технических наук, профессор (Белгородский государственный национальный исследовательский университет, г. Белгород)*

Коськин А.В., *доктор технических наук, профессор (Орловский государственный университет им. И. С. Тургенева, г. Орел)*

Ломазов В.А., *доктор физико-математических наук, профессор (Белгородский государственный аграрный университет им. В.Я. Горина, г. Белгород)*

Маторин С.И., *доктор технических наук, профессор (Белгородский государственный национальный исследовательский университет, г. Белгород)*

Таранчук В.Б., *доктор физико-математических наук, профессор, (Белорусский государственный университет, г. Минск, Республика Беларусь)*

EDITORIAL TEAM:

EDITOR-IN-CHIEF: ***Andrey A. Chernomorets,*** *Doctor of Technical Sciences, Associate Professor, Professor, Belgorod State National Research University*

DEPUTY EDITOR-IN-CHIEF: ***Alexander G. Zhikharev,*** *Doctor of Technical Sciences, Associate Professor, Belgorod State National Research University*

EXECUTIVE SECRETARY: ***Evgeniya V. Bolgova,*** *Candidate of Technical Sciences, Associate Professor, Belgorod State National Research University*

ENGLISH TEXT EDITOR: ***Igor V. Lyashenko,*** *Ph.D. in Philology, Associate Professor*

EDITORIAL BOARD:

Oleg O. Basov, *Doctor of Technical Sciences, Professor (Russia)*

Sergey P. Belov, *Doctor of Technical Sciences, Professor (Russia)*

Valery P. Volchkov, *Doctor of Technical Sciences, Professor (Russia)*

Valery D. Dmitrienko, *Doctor of Technical Sciences, Professor (Ukraine)*

Olga A. Ivaschuk, *Doctor of Technical Sciences, Professor (Russia)*

Igor A. Kalmykov, *Doctor of Technical Sciences, Professor (Russia)*

Nikolay I. Korsunov, *Honoured Science Worker of Russian Federation, Doctor of Technical Sciences, Professor (Russia)*

Alexander V. Koskin, *Doctor of Technical Sciences, Professor (Russia)*

Vadim A. Lomazov, *Doctor of Physico-mathematical Sciences, Professor (Russia)*

Sergey I. Matorin, *Doctor of Technical Sciences, Professor (Russia)*

Valery B. Taranchuk, *Doctor of Physico-mathematical Sciences, Professor (Belarus)*

Учредитель: Федеральное государственное автономное образовательное учреждение высшего образования
«Белгородский государственный национальный исследовательский университет»
Издатель: НИУ «БелГУ». Адрес издателя: 308015 г. Белгород, ул. Победы, 85.
Журнал выходит 4 раза в год

Founder: Federal state autonomous educational establishment of higher education
«Belgorod State National Research University»
Publisher: Belgorod State National Research University
Address of publisher: 85 Pobeda St., Belgorod, 308015, Russia
Publication frequency: 4 /year

© Белгородский государственный национальный исследовательский университет, 2023
© Belgorod State National Research University, 2023

СОДЕРЖАНИЕ

CONTENTS

ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ

INFORMATION SYSTEM AND TECHNOLOGIES

Кузьминых Е.С., Ильина С.П., Маслова М.А. Анализ сходства кода и поиска его заимствований	**3**
Урсол Д.В., Тихонский Н.А., Черноморец Д.А., Болгова Е.В., Черноморец А.А. О применении информативного частотного портрета при локализации отрезков эмпирических данных	**11**
Kuzminykh E.S., Ilina S.P., Maslova M.A. Analysis code similarity and search for its borrowings	**3**
Ursol D.V., Tikhonsky N.A., Chernomorets D.A., Bolgova E.V., Chernomorets A.A. On the informative frequency portrait application for the empirical data segments localization	**11**

АВТОМАТИЗАЦИЯ И УПРАВЛЕНИЕ

AUTOMATION AND CONTROL

Аль-Обайди А.М.Ж. О планировании активности сенсорных узлов в беспроводной сенсорной сети	**19**
Al-Obaidi A.M.Zh. About planning the activity of sensor nodes in a wireless sensor network	**19**

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРИНЯТИЕ РЕШЕНИЙ

ARTIFICIAL INTELLIGENCE AND DECISION MAKING

Бузов П.А., Игитян Е.В., Польщиков К.А., Лебедева А.С. Оценивание вероятностно-временных характеристик человеко-машинного диалога	**27**
Жихарев А.Г., Черных В.С. Классификация речевых данных по эмоциональному фону	**34**
Соколова Е.В. Интеллектуальный анализ наукометрических показателей российских научных журналов по информатике	**45**
Гармашев М.А., Резников Н.Г. Технология NoSQL: технические особенности, перспективы развития и сравнительные преимущества перед SQL	**56**
Buzov P.A., Igityan E.V., Polshchikov K.A., Lebedeva A.S. Estimation of probabilistic-temporal characteristics of human-machine dialogue	**27**
Zhikharev A.G., Chernykh V.S. Classification of speech data by emotional background	**34**
Sokolova E.V. Intellectual analysis of scientometric indicators of russian computer science journals	**45**
Garmashev M.A., Reznikov N.G. NoSQL technology: technical features, development prospects and comparative advantages over SQL	**56**

ИНФОРМАЦИОННЫЕ СИСТЕМЫ И ТЕХНОЛОГИИ
INFORMATION SYSTEM AND TECHNOLOGIES

УДК 004.05 **DOI: 10.18413/2518-1092-2022-8-3-0-1**

Кузьминых Е.С.
Ильина С.П.
Маслова М.А.

АНАЛИЗ СХОДСТВА КОДА И ПОИСКА ЕГО ЗАИМСТВОВАНИЙ

Севастопольский государственный университет, ул. Университетская, д. 33, г. Севастополь, 299053, Россия

e-mail: egor2014ru@mail.ru, sofi.ilina@mail.ru, mashechka-81@mail.ru

Аннотация
В настоящей статье рассматривается актуальная проблема в сфере программирования – заимствование кода и возможности его анализа. С увеличением числа программистов и обилием программных решений, появляется необходимость определения зависимостей и уникальности используемого кода. Вместе с возрастающим распространением кода через библиотеки и открытые источники, становится актуальным выявление случаев плагиата или незаконного использования. Данная статья исследует эффективные методы и инструменты для поиска заимствованного кода. Через представление существующих алгоритмов и подходов, авторы статьи предлагают оценку эффективности существующих решений и представляют примеры алгоритмов, способствующих анализу кода на схожесть.
Ключевые слова: информационная безопасность; ИБ; безопасность; кибербезопасность; код; программирование; анализ кода; поиск заимствований; заимствование кода; оценка сходства кода; дубликаты кода; машинное обучение; сортировка
Для цитирования: Кузьминых Е.С., Ильина С.П., Маслова М.А. Анализ сходства кода и поиска его заимствований // Научный результат. Информационные технологии. – Т.8, №3, 2023. – С. 3-10. DOI: 10.18413/2518-1092-2022-8-3-0-1

Kuzminykh E.S.
Ilina S.P.
Maslova M.A.

ANALYSIS CODE SIMILARITY AND SEARCH FOR ITS BORROWINGS

Sevastopol state University, 33 Universitetskaya St., Sevastopol, 299053, Russia

e-mail: egor2014ru@mail.ru, sofi.ilina@mail.ru, mashechka-81@mail.ru

Abstract
This article discusses an actual problem in the field of programming — the borrowing of code and the possibility of its analysis. With the increase in the number of programmers and the abundance of software solutions, there is a need to determine the dependencies and uniqueness of the code used. Along with the increasing distribution of code through libraries and open sources, it becomes relevant to identify cases of plagiarism or illegal use. This article explores effective methods and tools for finding borrowed code. Through the presentation of existing algorithms and approaches, the authors of the article offer an assessment of the effectiveness of existing solutions and provide examples of algorithms that contribute to the analysis of the code for similarity.
Keywords: information security; information security; security; cybersecurity; code; programming; code analysis; search for borrowings; code borrowing; code similarity assessment; code duplicates; machine learning; sorting

For citation: Kuzminykh E.S., Ilina S.P., Maslova M.A. Analysis code similarity and search for its borrowings // Research result. Information technologies. – T.8, №3, 2023. – P. 3-10. DOI: 10.18413/2518-1092-2022-8-3-0-1

ВВЕДЕНИЕ

В век развития информационных технологий появляется всё больше программистов, которые пишут свои уникальные программы или улучшают уже созданные, что приводит к огромному количеству созданных программ. Существует единая библиотека для любителей поделиться своими работами с другими. Такое распространение кода приводит к его воровству с небольшой доработкой или без. Известно, что существенная часть кодовой базы продуктов является наследованной от различных библиотек, фреймворков и пакетов. При анализе кода хочется уделять внимание уникальному коду и определять источники зависимостей. Для этого требуется реализовать эффективный алгоритм поиска заимствованного кода.

ОСНОВНАЯ ЧАСТЬ

В основном рассматриваются 4 типа клонов кода:

– Полные дубликаты — это копии исходного кода, которые полностью повторяют друг друга. Такой тип клонирования возникает при копировании и вставке кода без каких-либо изменений.

– Функциональные дубликаты — это копии исходного кода, которые выполняют ту же функцию, что и оригинал, но с некоторыми изменениями в структуре, например, с использованием других имен переменных или других алгоритмов решения задачи.

– Ключевые фрагменты — это небольшие участки кода, которые используются в разных местах программы. Это может быть как полная строка кода, так и отдельный оператор или функция.

– Структурные дубликаты — это копии исходного кода, которые имеют похожую структуру, но различаются по содержанию. Например, две функции могут иметь одинаковую структуру, но различаться в том, какие данные они обрабатывают и какие значения возвращают.

Существует несколько подходов к оценке сходства участков кода:

– Метрический подход: используются метрики для измерения сходства, такие как расстояние Левенштейна и коэффициент Жаккара.

Для поиска необходимо воспользоваться следующими пунктами:

1) Дизассемблирование исполняемого файла с использованием IDA Pro;

2) Деление кода ассемблера на перекрывающиеся куски кода, состоящие из дальнейших действий функции;

3) Упорядочение ряда команд при помощи абстракции информации о ячейках памяти и регистрах;

4) Подсчет метрик для нормализованных последовательностей;

5) Дальнейшее сравнение метрик, с результатом – получением групп клонов.

– Синтаксический подход: основан на анализе структуры и синтаксисе кода. В этом подходе сначала происходит разбор кода для получения его абстрактного синтаксического дерева (AST), а затем происходит сравнение AST-деревьев.

– Семантический подход: основан на анализе смысла кода. В этом подходе используется анализ контекста и вычисления зависимостей между переменными и функциями в коде.

– Гибридный подход: комбинирует два или более подхода для достижения лучших результатов [2, 3].

Каждый из этих подходов имеет свои преимущества и недостатки, и выбор определенного подхода зависит от конкретной задачи.

Рассматриваемые алгоритмы могут использовать различные свойства и признаки для решения задач.

Например, в алгоритмах машинного обучения для классификации объектов часто используются следующие свойства и признаки:

– числовые признаки (например, размер, вес, цена);
– категориальные признаки (например, цвет, форма, материал);
– текстовые данные (например, название товара, описание продукта);
– изображения и видео (например, фотографии или видеозаписи объектов).

В зависимости от задачи и типа данных, которые необходимо обработать, могут использоваться и другие свойства, и признаки. Например, для анализа временных рядов будут использоваться временные признаки, а для голосового управления — звуковые характеристики голоса пользователя.

Существует несколько методов преобразования категориальных признаков:

– кодирование меток — способ замены категориальных признаков на числовые значения. Любое свойство категориального признака заменяется на подходящее ему число. Данный способ подходит для признаков с несколькими значениями, например, уровень образования;

– однократное кодирование — способ замены категориальных признаков на бинарные значения. Каждая переменная категориального признака преобразуется в столбец бинарных значений, где 1 указывает на присутствие значения признака, а 0 – на его отсутствие. Данный способ подходит для признаков без порядка значений, таких как цвет либо тип объекта;

– двоичное кодирование — способ замены категориальных признаков в бинарные значения методом кодирования Грея. В данном методе каждое значение категориального признака заменяется на бинарное значение. Далее каждый последующий столбец имеет величину, отличную на одно значение от предшествующей. Данный метод можно использовать и для признаков без порядка значений.

Числовыми признаками считаются: масштабирование, нормализация, дискретизация, стандартизация [4].

Проанализированные подходы могут столкнуться со следующими проблемами:

– недостаточное количество данных — это может привести к низкой точности моделей и переобучению. Решение: сбор большего количества данных или использование техник аугментации данных;

– несбалансированные классы — это может привести к низкой точности моделей для меньшего класса объектов. Решение: использование взвешивания классов, увеличение обучающих данных для меньшего класса;

– переобучение — это происходит, когда модель слишком сложная для задачи или когда она обучается на выбросах. Решение: использование регуляризации, выбор упрощенной модели, использование случайных выборок (dropout) [5].

– недообучение — это происходит, когда модель слишком простая для задачи или когда она не имеет достаточно данных для обучения. Решение: использование более сложной модели, использование большего количества данных;

– скрытые переменные — это может быть вызвано отсутствием информации о состоянии системы. Решение: использование вероятностных моделей или методов инверсии;

– необходимость интерпретации результатов — это может быть вызвано необходимостью объяснения принятого решения или причины ошибки. Решение: использование методов интерпретации, таких как анализ важности функций, сводные таблицы и графические представления.

Решение каждой из этих проблем зависит от конкретной задачи и используемых данных. Важно подобрать наиболее подходящие методы для решения каждой из проблем.

Одним из главных преимуществ методов, использующих машинное обучение (ML), является автоматическое извлечение закономерностей из больших объемов данных. Это позволяет решать задачи, которые трудно или невозможно решить с помощью традиционных методов.

Например, ML-алгоритмы могут использоваться для построения моделей, которые предсказывают будущие события на основе прошлых данных. Это может применяться в различных отраслях, таких как финансы, медицина, производство и т.д.

Еще одним преимуществом методов, использующих машинное обучение, является возможность создания универсальных моделей, которые могут быть применены к различным задачам. Например, нейронные сети могут быть обучены распознавать образы, а также делать предсказания на основе временных рядов.

Кроме того, ML-алгоритмы могут автоматически адаптироваться к изменениям в данных, что делает их более гибкими и универсальными. Они также могут работать с данными, которые трудно интерпретировать или имеют сложную структуру.

Наконец, использование ML-методов может помочь снизить затраты на ручную обработку и анализ данных, что может быть особенно полезно в случае больших объемов информации.

Для проверки работы исследуемых алгоритмов могут использоваться различные тестовые данные в зависимости от конкретной задачи. Например, для задач классификации можно использовать наборы данных, содержащие изображения или тексты с определенными метками классов. Для задач регрессии — наборы данных со значениями целевых переменных.

Кроме того, для оценки качества работы алгоритмов могут использоваться различные метрики, такие как точность (precision), полнота (recall), F-мера и другие, которые могут быть выбраны в зависимости от конкретной задачи.

Также можно использовать кросс-валидацию, чтобы оценить обобщающую способность модели на новых данных. При кросс-валидации данные делятся на несколько частей, и каждый раз обучение и тестирование проводятся на разных подмножествах данных.

Кросс-валидация — используют для оценки качества работы модели. Он обширно применяется в машинном обучении, дает возможность сравнивать между собой различные модели и определять наилучшую для конкретной задачи [6].

В любом случае, для выбора тестовых данных следует принимать во внимание особенности конкретной задачи и возможные источники ошибок при работе алгоритмов на этих данных.

Точность — это доля ответов, которые алгоритм верно классифицировал относительно общего количества ответов, которые он предсказал. То есть, если алгоритм предсказал 10 объектов, из которых 8 оказались верными, то его точность составит 80%.

Полнота — это доля верно предсказанных ответов от общего числа правильных ответов. Если, например, всего было 20 правильных ответов, но алгоритм угадал лишь 16, то его полнота будет равна 80%.

Применяется несколько способов объединения precision и recall в агрегированный критерий качества. F-мера (в общем случае F_β) — среднее гармоническое precision и recall (1):

$$F_\beta = (1 + \beta^2) \cdot \frac{precision \cdot recall}{(\beta^2 \cdot precision) + recall} \quad (1)$$

где

β – определяет вес точности в метрике, и при $\beta = 1$ – это среднее гармоническое (с множителем 2, чтобы в случае precision = 1 и recall = 1 иметь $F_1 = 1$);

F – может достигнуть максимума при полноте и точности, равными единице, также близка к нулю, когда один из аргументов близок к нулю [7].

Оценка сходства кода может быть использована для решения следующих прикладных задач информационной безопасности:

– обнаружение вредоносного кода: эффективный алгоритм оценки сходства кода может позволить обнаружить вредоносный код, который был изменен или скрыт за шифрованием или другими методами маскировки;

– анализ утечек данных: алгоритм оценки сходства кода может помочь выявить случаи кражи информации и утечки конфиденциальных данных. Также алгоритм может показать, какие файлы были скопированы или перемещены на другой носитель;

– выявление уязвимостей: алгоритм оценки сходства кода может использоваться для анализа кода на наличие уязвимостей. Например, он может выявить похожий уязвимый код, который уже был использован в атаках или которые может стать целью будущих атак;

– проверка подлинности: алгоритм оценки сходства кода может использоваться для проверки подлинности программного обеспечения. Например, он может помочь определить, является ли программа настоящей или фальсифицированной;

– анализ соблюдения правил безопасности: алгоритм оценки сходства кода может помочь анализировать соблюдение правил безопасности и стандартов в программном обеспечении. Например, он может выявить нарушение политик безопасности, таких как использование небезопасных функций или открытие портов без разрешения [8].

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Далее продемонстрирована работа нескольких базовых алгоритмов и показаны их недостатки.

Листинг 1 — Алгоритм SequenceMatcher [9]

```
      import difflib

      def compare_code(file1, file2):
          with open(file1, 'r') as f1, open(file2, 'r') as f2:
              code1 = f1.readlines()
              code2 = f2.readlines()

          matcher = difflib.SequenceMatcher(None, code1, code2)
          similarity_ratio = matcher.ratio()

          return similarity_ratio

      file1 = "file1.txt"
      file2 = "file2.txt"

      similarity = compare_code(file1, file2)
print(f"Similarity ratio: {similarity}")
```

Метод SequenceMatcher — это класс пакета difflib Python, который предоставляет функции и классы для сравнения последовательностей. Используется для сравнения различных файлов на схожесть.

В этом примере программа открывает и считывает содержимое файлов «file1.txt» и «file2.txt». Затем создает экземпляр класса SequenceMatcher, передавая в него строки кода из обоих файлов. Метод ratio() возвращает отношение схожести между этими строками, где 0 означает полное отсутствие сходства, а 1 — полное совпадение. Программа выводит коэффициент сходства между двумя файлами. Чем ближе значение к 1, тем более похожи коды в файлах. Однако, следует учесть, что такой алгоритм будет работать только на уровне строк и не учитывает семантику кода. Если требуется более точный анализ сходства кода, то могут потребоваться более сложные методы и библиотеки, такие как алгоритмы на основе графового анализа или машинное обучение.

Следующий алгоритм похож на первый, но чуть сложнее, хотя использует ту же самую библиотеку difflib.

Листинг 2 — Вторая версия difflib

```
import difflib

def compare_code(file1, file2):
    with open(file1, 'r') as f1, open(file2, 'r') as f2:
        code1 = f1.read()
        code2 = f2.read()

    matcher = difflib.SequenceMatcher(None, code1, code2)
    blocks = matcher.get_matching_blocks()

    similarities = []
    for block in blocks:
        if block.size > 0:
            similarity = code1[block.a:block.a+block.size]
            similarities.append(similarity.strip())

    return similarities

file1 = "file1.txt"
file2 = "file2.txt"

similar_lines = compare_code(file1, file2)
if similar_lines:
    print("Similar lines found:")
    for line in similar_lines:
        print(line)
else:
print("No similar lines found.")
```

В этом примере программа также открывает и считывает содержимое файлов «file1.txt» и «file2.txt». Затем создает экземпляр класса SequenceMatcher, чтобы получить сравнение между строками кода. Функция get_matching_blocks() возвращает список обнаруженных блоков, включающих совпадения и различия между файлами. Программа проходится по каждому блоку и, если размер блока больше 0, добавляет схожую строку в список similarities. Затем выводит найденные схожие строки. Этот пример программы позволяет найти не только коэффициент сходства между файлами кода, но и конкретные строки, которые совпадают или похожи друг на друга.

Сравнив эти 2 алгоритма, можно сделать вывод, что для базовой проверки кода на схожесть достаточно программы из 20 строк, которая сможет выявить явную схожесть кода, но для более глубокой проверки необходимо учитывать семантику кода и разрабатывать более сложные алгоритмы проверки [10].

ЗАКЛЮЧЕНИЕ

В заключении можно подчеркнуть важность анализа сходства кода и поиска его заимствований в современной разработке программного обеспечения. Этот метод помогает обнаружить повторяющиеся фрагменты кода и выявить возможные проблемы с плагиатом или небезопасными практиками программирования. Анализ сходства кода предоставляет разработчикам возможность оптимизировать процесс разработки, повысить эффективность и надежность кода. Поиск заимствований кода помогает командам разработчиков обнаружить повторное использование кода из различных источников, включая открытые источники, легаси-код или даже внутри компании. Это позволяет оптимизировать использование ресурсов и избежать проблем, связанных с авторскими правами или патентами.

Современные инструменты анализа сходства кода предлагают широкий набор функций, позволяющих проводить глубокий и точный анализ кода, учитывая его синтаксис, структуру и

логику. Такие инструменты помогают разработчикам быстро обнаруживать дублированный код, проводить рефакторинг и бороться с потенциальными уязвимостями. В итоге, осознанный подход к анализу сходства кода и поиску его заимствований является неотъемлемой частью современной разработки ПО. Это помогает повысить качество и надежность кода, оптимизировать его использование и обеспечить соблюдение авторских прав [11].

Список литературы

1. Бабкина А.А. Инструментальная поддержка поиска клонов в программном коде: Магистерская диссертация – Сибирский федеральный университет, 2017.
2. Саргсян С.С. Методы поиска клонов кода и семантических ошибок на основе семантического анализа программы: дис. канд. физико-математических Наук – Ин-т систем. программирования, 2016.
3. Выявление клонов [Электронный ресурс]. URL: https://newtechaudit.ru/clone-code/
4. Обработка данных для машинного обучения [Электронный ресурс]. URL: https://4brain.ru/aibasics/data.php
5. Какие бывают типы ошибок в моделях машинного обучения? [Электронный ресурс]. URL: https://qaa-engineer.ru/kakie-byvayut-tipy-oshibok-v-modelyah-mashinnogo-obucheniya/
6. Кросс-валидация [Электронный ресурс]. URL: https://academy.yandex.ru/handbook/ml/article/kross-validaciya
7. Чутка ликбеза: что такое f мера в машинном обучении? [Электронный ресурс]. URL: https://dzen.ru/a/ZIHJQbB6HGcQVC6i
8. Ищем уязвимости в коде: теория, практика и перспективы SAST [Электронный ресурс]. URL: https://www.securitylab.ru/analytics/483063.php
9. Алгоритм SequenceMatcher [Электронный ресурс]. URL: https://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher
10. Вовченко Н.Г., Кузнецов Н.Г., Макаренко Е.Н. и др. Реализация ESG-принципов в стратегии устойчивого развития экономики России. – Ростов-на-Дону: Ростовский государственный экономический университет "РИНХ", 2022. – 508 с.
11. Нестеренко В.Р., Маслова М.А. Современные вызовы и угрозы информационной безопасности публичных облачных решений и способы работы с ними // Научный результат. Информационные технологии. – 2021. – Т. 6, № 1. – С. 48-54.

References

1. Babkina A.A. Instrumental support for the search for clones in the program code: Master's thesis – Siberian Federal University, 2017.
2. Sargsyan S.S. Methods of searching for code clones and semantic errors based on semantic analysis of the program: dis. Candidate of Physical and Mathematical Sciences – In-t Systems. programming, 2016.
3. Identification of clones [Electronic resource]. URL: https://newtechaudit.ru/clone-code/
4. Data processing for machine learning [Electronic resource]. URL: https://4brain.ru/aibasics/data.php
5. What are the types of errors in machine learning models? [electronic resource]. URL: https://qaa-engineer.ru/kakie-byvayut-tipy-oshibok-v-modelyah-mashinnogo-obucheniya/
6. Cross-validation [Electronic resource]. URL: https://academy.yandex.ru/handbook/ml/article/kross-validaciya
7. A little educational program: what is an f measure in machine learning? [electronic resource]. URL: https://dzen.ru/a/ZIHJQbB6HGcQVC6i
8. Looking for vulnerabilities in the code: theory, practice and prospects of SAST [Electronic resource]. URL: https://www.securitylab.ru/analytics/483063.php
9. Algoritm SequenceMatcher [Elektronnyj resurs]. URL: https://docs.python.org/3/library/difflib.html#difflib.SequenceMatcher

10. Vovchenko N.G., Kuznetsov N.G., Makarenko E.N. and others. Implementation of ESG principles in the strategy for sustainable development of the Russian economy. – Rostov-on-Don: Rostov State Economic University "RINH", 2022. – 508 p.

11. Nesterenko R.V., Maslova M.A. Modern challenges and threats information security public cloud making and methods of work with them // Research result. Information technologies. – T.6, №1, 2021. – P. 48-54. DOI: 10.18413/2518-1092-2021-6-1-0-6

Кузьминых Егор Сергеевич, студент четвертого курса кафедры Информационная безопасность Института информационных технологий
Ильина София Павловна, студент четвертого курса кафедры Информационная безопасность Института информационных технологий
Маслова Мария Александровна, старший преподаватель кафедры Информационная безопасность Института информационных технологий

Kuzminykh Egor Sergeevich, fourth-year Student of the Department Information security, Institute of Information Technologies
Ilina Sofia Pavlovna, fourth-year Student of the Department Information security, Institute of Information Technologies
Maslova Maria Alexandrovna, Senior Lecturer of the Department Information security Institute of Information Technologies

УДК 004.93 **DOI: 10.18413/2518-1092-2022-8-3-0-2**

Урсол Д.В.[1]
Тихонский Н.А.[2]
Черноморец Д.А.[2]
Болгова Е.В.[2]
Черноморец А.А.[2]

О ПРИМЕНЕНИИ ИНФОРМАТИВНОГО ЧАСТОТНОГО ПОРТРЕТА ПРИ ЛОКАЛИЗАЦИИ ОТРЕЗКОВ ЭМПИРИЧЕСКИХ ДАННЫХ

[1)] ООО «Промышленные электронные системы», ул. Михайловское шоссе 121а, г. Белгород, 308000, Россия
[2)] Белгородский государственный национальный исследовательский университет, ул. победы, 85, Белгород, 308015, Россия

e-mail: chernomorets@bsu.edu.ru

Аннотация
Задача локализации (определения местоположения) в зарегистрированном числовом ряде заданного ряда меньшей размерности зачастую возникает при обработке эмпирических данных различной природы. В данной работе разработан алгоритм локализации отрезков эмпирических данных, которые представлены числовыми рядами. Разработанный алгоритм основан на вычислении меры сходства признаков анализируемых числовых рядов. В качестве множества признаков, характеризующих числовой ряд, предложено применять информативный частотный портрет отрезка числовых данных, для вычисления которого в работе приведены соответствующие соотношения на основе дискретного косинус-преобразования. В работе приведены примеры вычисления информативного частотного портрета для различных числовых рядов (прецедентов). Приведенные в работе результаты вычислительных экспериментов демонстрируют работоспособность разработанного алгоритма по оцениванию локализации отрезков числовых рядов на основе применения информативного частотного портрета при дискретном косинус-преобразовании.

Ключевые слова: числовой ряд; прецедент; локализация; дискретное косинус-преобразование; информативный частотный портрет; погрешность локализации
Для цитирования: Урсол Д.В., Тихонский Н.А., Черноморец Д.А., Болгова Е.В., Черноморец А.А. О применении информативного частотного портрета при локализации отрезков эмпирических данных // Научный результат. Информационные технологии. – Т.8, №3, 2023. – С. 11-18. DOI: 10.18413/2518-1092-2022-8-3-0-2

Ursol D.V.[1]
Tikhonsky N.A.[2]
Chernomorets D.A.[2]
Bolgova E.V.[2]
Chernomorets A.A.[2]

ON THE INFORMATIVE FREQUENCY PORTRAIT APPLICATION FOR THE EMPIRICAL DATA SEGMENTS LOCALIZATION

[1)] Industrial Electronic Systems LLC, 121a Mikhailovskoe shosse, Belgorod, 308000, Russia
[2)] Belgorod State National Research University, 85 Pobedy St., Belgorod, 308015, Russia

e-mail: chernomorets@bsu.edu.ru

Abstract
The problem of localization (location determination) in a registered numerical series of a given series of smaller dimensions often arises when processing empirical data of various nature. In this paper, an algorithm for the empirical data segments localization, which are represented by numerical series, is developed. The developed algorithm is based on the calculation of the similarity measure of the analyzed numerical series features. As a set of features characterizing a numerical series, it is proposed to use an informative frequency portrait of a segment of numerical data, for the calculation of which the corresponding relations based on a discrete cosine transformation are given in the work. The paper provides examples of calculating an informative frequency portrait

for various numerical series (precedents). The results of computational experiments presented in this paper demonstrate the operability of the developed algorithm for estimating the localization of segments of numerical series based on the use of an informative frequency portrait with discrete cosine transformation.

Keywords: numerical series; precedent; localization; discrete cosine transform; informative frequency portrait; localization error

For citation: Ursol D.V., Tikhonsky N.A., Chernomorets D.A., Bolgova E.V., Chernomorets A.A. On the informative frequency portrait application for the empirical data segments localization // Research result. Information technologies. – T. 8, №3, 2023. – P. 11-18. DOI: 10.18413/2518-1092-2022-8-3-0-2

ВВЕДЕНИЕ

При обработке эмпирических данных зачастую возникает необходимость локализации (определения местоположения) в зарегистрированном числовом ряде заданного ряда меньшей размерности. Данные задачи возникают, например, при обработке звукозаписей с целью обнаружения в них отдельных звуков, при обработке радиолокационных сигналов и др. [1-8]. В настоящее время известно значительное количество методов решения данных задач, в которых анализируются отрезки некоторого числового ряда на соответствие эталонному числовому ряду. В большинстве случаев данные методы основаны на вычислении меры сходства множеств различных признаков, соответствующих анализируемым числовым рядам. В данной работе в качестве множества признаков, характеризующих числовой ряд, предлагается применять так называемый информативный частотный портрет отрезка числовых данных.

Эмпирические (зарегистрированные) данные во многих случаях представляют собой ряд вещественных чисел, при обработке которых целесообразно применять вещественные преобразования, такие как дискретное косинус-преобразование (ДКП), дискретное синус-преобразование, преобразование Уолша-Адамара, преобразование Хаара и др., в отличие от широко применяемого комплексного дискретного преобразования Фурье сигналов различной природы [9-20]. В данной работе для решения задачи построения информативного частотного портрета отрезка вещественного числового ряда предлагается применять дискретное косинус-преобразование, для которого известны алгоритмы быстрого преобразования.

ОСНОВНАЯ ЧАСТЬ

Рассмотрим алгоритм построения информативного частотного портрета отрезка числовых данных на основе дискретного косинус-преобразования.

Обозначим, $F=\{f_n\}$, $n=1,2,...,N$, – числовой ряд, содержащий N элементов. Коэффициенты $D=\{d_k^F\}$, $k=1,2,...,N$, ДКП данного ряда вычисляются на основании следующего соотношения:

$$d_k^F=\sqrt{2/N}\sum_{n=1}^{N} f_n\cos(\frac{\pi}{N}(n-1)(k-\frac{1}{2})),\ k=1,2,...,N. \quad (1)$$

С целью снижения размерности анализируемых данных предлагается применить так называемое субполосное представление результатов ДКП. Обозначим, M – заданное количество субполос ДКП. Тогда, количество J коэффициентов ДКП, соответствующих отдельной субполосе, определяется следующим образом:

$$J=[N/M], \quad (2)$$

где [] – операция вычисления целой части числа.

Элементы субполосного представления ДКП $S^F=\{s_m^F\}$, $m=1,2,...,M$, определим следующим образом:

$$s_m^F=\sum_{j=1}^{J}(d_{(m-1)J+j}^F)^2,\ m=1,2,...,M. \quad (3)$$

Частотным портретом числового ряда $F=\{f_n\}$, $n=1,2,...,N$, назовем вектор $G^{*F}=\{g_m^{*F}\}$, $m=1,2,...,M$, элементы которого имеют вид:

$$g_m^{*F}=s_m^F/\sum_{i=1}^{M} s_i^F,\ m=1,2,...,M. \quad (4)$$

Информативным частотным портретом (ИЧП) числового ряда $F=\{f_n\}$, $n=1,2,...,N$, назовем вектор $G^F=\{g_m^F\}$, $m=1,2,...,M$, элементы которого имеют вид:

$$g_m^F=\begin{cases} g_m^{*F}, & \text{если } g_m^{*F}>1/M, \\ 0, & \text{в противном случае.} \end{cases},\ m=1,2,...,M. \quad (5)$$

На рисунке 1 приведен пример вычисления ИЧП для числового ряда F_1, размерности 512 элементов, значения которого в виде графика представлены на рисунке 1а. На рисунке 1 также приведены результаты вычисления коэффициентов ДКП D^{F_1}, а также частотный портрет G^{*F_1} и информативный частотный портрет G^{F_1}, содержащие 32 элемента (на основании соотношений (1)-(5)).

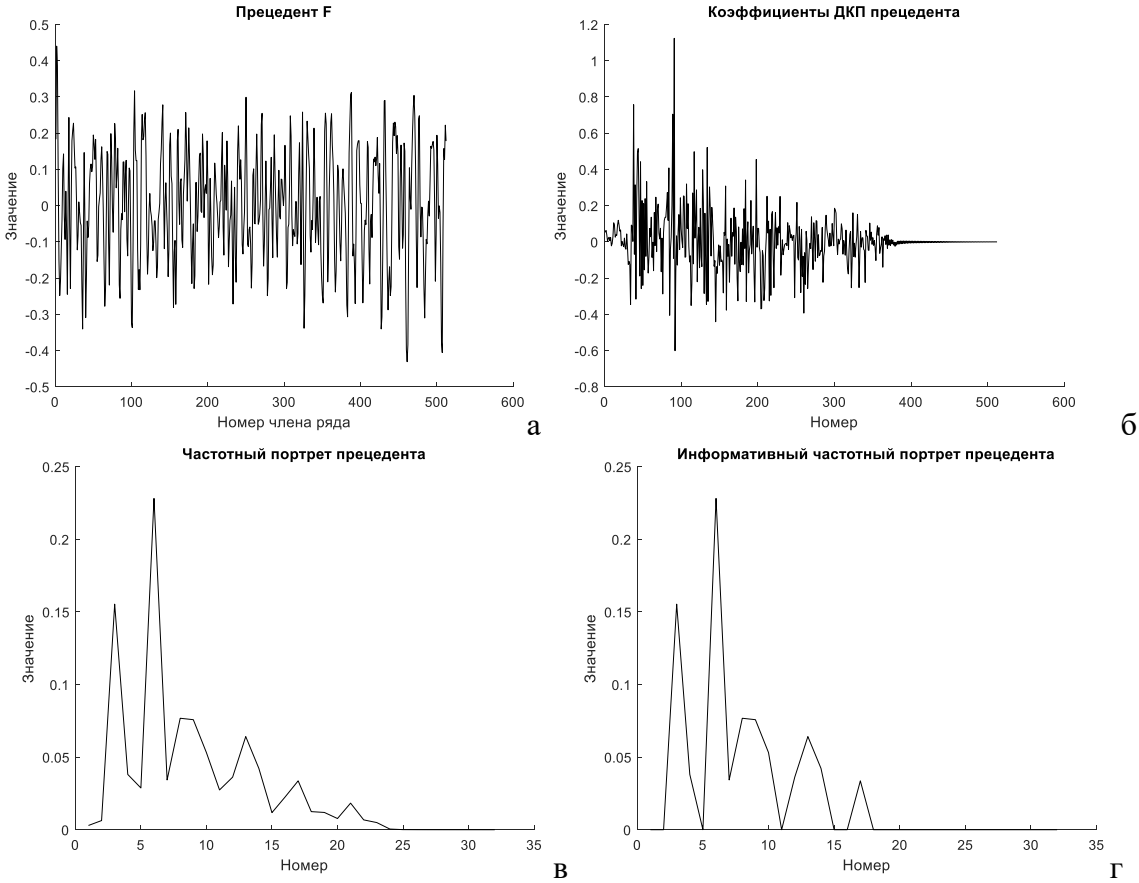


Рис. 1. Пример вычисления информативного частотного портрета:
а – числовой ряд F_1, б – коэффициенты ДКП D^{F_1},
в – частотный портрет G^{*F_1}, г – информативный частотный портрет G^{F_1}
Fig. 1. An example of calculating an informative frequency portrait:
a – numerical series F_1, b – DCT coefficients D^{F_1},
c – frequency portrait G^{*F_1}, d – informative frequency portrait G^{F_1}

Для оценивания местоположения (локализации) заданного прецедента F в числовом ряде W разработан следующий алгоритм.

Исходные данные:

N – размерность прецедента F,

$F = \{f_n\}$, $n = 1,2,...,N$, – прецедент (числовой ряд),

M – размерность информативного частотного портрета,

$G^F = \{g_m^F\}$, $m = 1,2,...,M$, – информативный частотный портрет,

L – размерность числового ряда, в котором необходимо определить местоположение прецедента,

$W = \{w_l\}$, $l = 1,2,...,L$, – анализируемый числовой ряд.

Шаг 1. Вычислить информативный частотный портрет G^F прецедента F.

Шаг 2. Выбрать очередное значение i – местоположение в числовом ряде W первого элемента (координата полюса) анализируемого отрезка данных, размерности N,

$$1 \le i \le L - N + 1. \quad (6)$$

Шаг 3. Создать анализируемый отрезок данных $V = \{v_n\}$, $n = 1,2,...,N$, размерности N, на основе числового ряда W,

$$v_n = w_{i+n-1}, \; n = 1,2,...,N. \quad (7)$$

Шаг 4. Вычислить информативный частотный портрет G^V для вектора V (на основании соотношений (1)-(5)).

Шаг 5. Вычислить среднеквадратическое отклонение σ_i информативного частотного портрета G^V относительно информативного частотного портрета G^F прецедента F,

$$\sigma_i = \sqrt{\frac{\sum_{m=1}^{M}(g_m^V - g_m^F)^2}{\sum_{m=1}^{M}(g_m^F)^2}}. \quad (8)$$

Шаг 6. Перейти к шагу 2, если выполнено условие:

$$i < L - N + 1.$$

Шаг 7. Вычислить номер i^*, соответствующий минимальному значению σ_i, $i = 1,2,...,L-N+1$,

$$i^* = \arg \min_{i=1,2,...,L-N+1} \sigma_i. \quad (9)$$

Шаг 8. Конец.

В качестве оценки предполагаемого местоположения прецедента F в числовом ряде W предлагается выбрать значение i^* (9).

При проверке работоспособности разработанного алгоритма предлагается оценивать погрешность δ^F локализации (определения местоположения) прецедента F в числовом ряде W,

$$\delta^F = |i^* - i^F|. \quad (10)$$

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ И ИХ ОБСУЖДЕНИЕ

Следующие вычислительные эксперименты демонстрируют результаты применения ИЧП при локализации на числовом ряде отрезков данных, соответствующих заданному прецеденту.

На рисунке 2 приведен пример прецедента F (рисунок 2а), размерности N=256 элементов, местоположение которого необходимо оценить в заданном числовом ряде W (рисунок 2в), размерности L=2048 элементов. Значения элементов прецедента выбраны так, чтобы они совпадали со значениями элементов числового ряда W, с номерами от 200 до 455. Для прецедента F на основании соотношений (1)-(5) построен информативный частотный портрет G^F, содержащий 32 элемента (рисунок 2б).

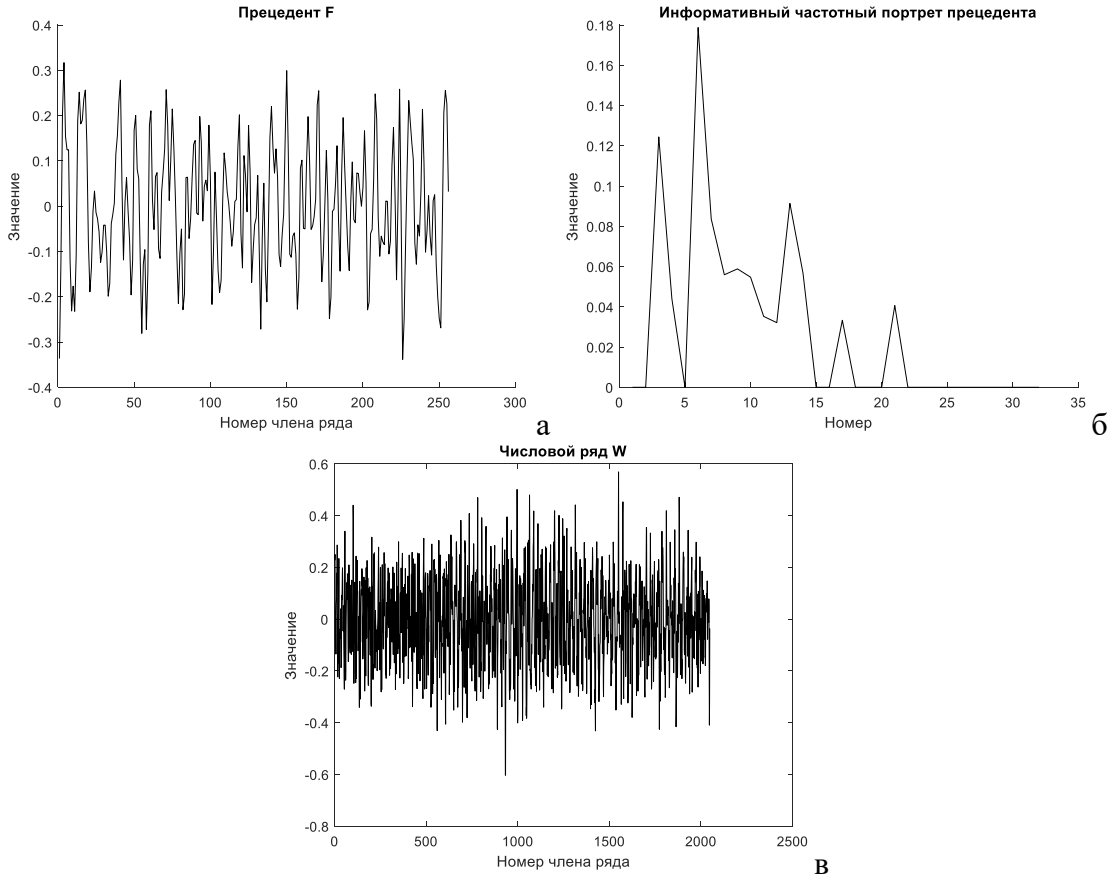


Рис. 2. Прецедент и анализируемый числовой ряд: а – прецедент F , б – информативный частотный портрет G^F прецедента, в – анализируемый числовой ряд W

Fig. 2. Precedent and analyzed numerical series: a – precedent F , b – informative frequency portrait G^F , c – analyzed numerical series W

При применении разработанного алгоритма оценивания местоположения заданного прецедента F в числовом ряде W получены следующие результаты.

На рисунке 3 приведен график значений среднеквадратического отклонения (8) информативного частотного портрета G^V отрезка данных V, размерности 256 элементов, относительно информативного частотного портрета G^F прецедента F, вычисленных для всевозможных положений полюса анализируемого отрезка данных V в заданном числовом ряде W.

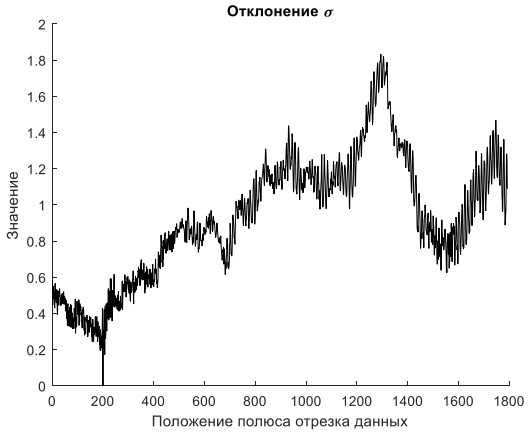


Рис. 3. Значения среднеквадратического отклонения (8) для всевозможных положений полюса анализируемого отрезка данных

Fig. 3. Values of the square deviation (8) for all possible pole positions of the analyzed data segment

При проведении вычислительного эксперимента в качестве оценки (9) местоположения прецедента F в числовом ряде W выбран 200-й элемент числового ряда W (минимальное значение на графике на рисунке 3), что совпадает с модельным положением прецедента. Следовательно, погрешность δ^F (10) определения местоположения прецедента F в числовом ряде W равна 0, что демонстрирует работоспособность разработанного алгоритма.

В таблице 1 приведены результаты вычислительных экспериментов по оцениванию на основании разработанного алгоритма местоположений полюса (первого элемента) прецедентов, размерности 128, 256 и 512 элементов в числовом ряде W. При проведении вычислительных экспериментов была задана размерность информативного частотного портрета 16, 32 и 64 элемента.

Таблица 1

Результаты вычислительных экспериментов по оцениванию местоположений полюса прецедентов

Table 1

Results of computational experiments on estimating the precedents pole locations

Размерность прецедента (элементы)	Размерность информативного частотного портрета (элементы)	Заданное местоположение полюса прецедента (номер элемента)	Оценка местоположения полюса прецедента (номер элемента)
128	16	100	100
128	32	100	100
128	64	100	100
256	16	300	300
256	32	300	300
256	64	300	300
512	16	500	500
512	32	500	500
512	64	500	500

Данные, приведенные в таблице 1, демонстрируют работоспособность разработанного алгоритма по оцениванию локализации отрезков числовых рядов на основе применения информативного частотного портрета при ДКП – полученные оценки местоположения полюса прецедента совпадают с заданными положениями полюса. Следует отметить, что в вычислительных экспериментах, в которых использованы размерности прецедента и ИЧП, отличные от приведенных в таблице 1, также были получены соответствующие результаты оценивания местоположений полюса прецедентов.

ЗАКЛЮЧЕНИЕ

Таким образом, в работе приведено построение информативного частотного портрета в рамках косинус преобразования отрезка эмпирических данных, а также разработан алгоритм оценивания местоположения заданного прецедента в числовом ряде. Проведенные вычислительные эксперименты показали работоспособность алгоритма выделения прецедента в числовом ряде на основе применения информативного частотного портрета.

Список литературы

1. Абрамов Г.В., Коробова Л.А., Ивашин А.Л., Матыцина И.А. Анализ и использование математических методов для распознавания звуковых сигналов // Вестник воронежского государственного университета инженерных технологий. – 2015. – №2(64). – С. 61-65.

2. Гапочкин А.В. Нейронные сети в системах распознавания речи // Science time. – 2014. – №1. – С. 29-36.

3. Пузанов А.Д., Нефедов Д.С. Синтез алгоритма обнаружения беспилотных летательных аппаратов по акустическим шумам // Доклады белорусского государственного университета информатики и радиоэлектроники. – 2021. – №2. – С. 65-73.

4. Зинченко Ю.Е., Азаров А. Б., Константинов В.С., Зинченко Т.А. Разработка алгоритма и архитектуры вкс распознавания звуковых сигналов на базе FPGA // Современные тенденции развития и перспективы внедрения инновационных технологий в машиностроении, образовании и экономике. – 2019. – №1 (4). – С. 75-80.

5. Алпатов Б.А., Бабаян П.В., Балашов О.Е., Степашкин А.И. Методы автоматического обнаружения и сопровождения объектов. Обработка изображений и управление. – М.: Радиотехника, 2008 г., – 176 с.

6. Рысин Ю.С., Терехов А.Н., Яблочников С.Л., Иевлев О.П. Теоретические аспекты идентификации объектов-источников звуковых сигналов // Системы синхронизации, формирования и обработки сигналов. – 2018. – Т. 9, № 4. – С. 101-105.

7. Калугин Ю.А., Рудаков И.В. Методы выделения одного заданного источника звука по одной зашумленной записи // Modern Science. – 2020. – № 3-2. – С. 254-264.

8. Плекин В.Я. Алгоритмы и цифровые устройства селекции движущихся целей. – М.: МАИ, 2002.

9. Жиляков Е.Г. Вариационные методы анализа и построения функций по эмпирическим данным на основе частотных представлений. – Белгород: Изд-во БелГУ, 2007. – 160 с.

10. Черноморец А.А., Болгова Е.В., Черноморец Д.А. Обобщенный субполосный анализ на основе унитарных преобразований // Научные ведомости БелГУ. Сер. Экономика. Информатика. – 2015. – № 7(204). – Вып. 34/1. – С. 97-104.

11. Черноморец А.А., Болгова Е.В. Об анализе данных на основе косинусного преобразования // Научные ведомости БелГУ. Сер. История. Политология. Экономика. Информатика. – 2015. – № 1(198). – Вып. 33/1. – С. 68-73.

12. Ахмед Н., Рао К. Ортогональные преобразования при обработке цифровых сигналов. – М.: Связь, 1980. – 248 с.

13. Залманзон Л.А. Преобразования Фурье, Уолша, Хаара и их применение в управлении, связи и других областях. – М.: Наука, 1989. – 496 с.

14. Яне Б. Цифровая обработка изображений. – М.: Техносфера, 2007. – 584 с.

15. Гонсалес Р., Вудс Р. Цифровая обработка изображений. Издание 3-е, исправленное и дополненное. – М.: Техносфера, 2012. – 1104 с.

16. Рабинер Л., Голд Г. Теория и применение цифровой обработки сигналов. – М.: Мир, 1988. – 512 с.

17. Сергиенко А.Б. Цифровая обработка сигналов: учеб. пособие для студ. вузов. – СПб.: Питер, 2002. – 603с.: ил.

18. Голд Б., Рейдер Ч. Цифровая обработка сигналов: пер. с англ. / Б. Голд– М.: Сов. радио, 1973.

19. Нуссбаумер Г. Быстрое преобразование Фурье и алгоритмы вычисления сверток / Г. Нуссбаумер; ред.: В.М. Амербаев, Т.Э. Кренкель; пер. с англ.: Ю.Ф. Касимов, И.П. Пчелинцев. – М.: Радио и связь, 1985. – 248 с.

20. Прэтт У. Цифровая обработка изображений. – М.: Мир, 1982. – 312 с.

References

1. Abramov G.V., Korobova L.A., Ivashin A.L., Matytsina I.A. Analysis and use of mathematical methods for recognizing sound signals // Bulletin of the Voronezh State University of Engineering Technologies. – 2015. – No. 2(64). – pp. 61-65.

2. Gapochkin A.V. Neural networks in speech recognition systems // Science time. – 2014. – No. 1. – pp. 29-36.

3. Puzanov A.D., Nefedov D.S. Synthesis of an algorithm for detecting unmanned aerial vehicles based on acoustic noise // Reports of the Belarusian State University of Informatics and Radioelectronics. – 2021. – No. 2. – P. 65-73.

4. Zinchenko Yu.E., Azarov A.B., Konstantinov V.S., Zinchenko T.A. Development of an algorithm and architecture for audio signal recognition based on FPGA // Modern development trends and prospects for the implementation of innovative technologies in mechanical engineering, education and economics. – 2019. – No. 1(4). – P. 75-80.

5. Alpatov B.A., Babayan P.V., Balashov O.E., Stepashkin A.I. Methods for automatic detection and tracking of objects. Image processing and control. – M.: Radio engineering, 2008, – 176 p.

6. Rysin Yu.S., Terekhov A.N., Yablochnikov S.L., Ievlev O.P. Theoretical aspects of identification of objects-sources of sound signals // Systems of synchronization, generation and processing of signals. – 2018. – T. 9, No. 4. – P. 101-105.

7. Kalugin Yu.A., Rudakov I.V. Methods for identifying one given sound source from one noisy recording // Modern Science. – 2020. – No. 3-2. – pp. 254-264.

8. Plekin V.Ya. Algorithms and digital devices for selecting moving targets. – M.: MAI, 2002.

9. Zhilyakov E.G. Variational methods of analysis and construction of functions from empirical data based on frequency representations. – Belgorod: BelSU Publishing House, 2007. – 160 p.

10. Chernomorets A.A., Bolgova E.V., Chernomorets D.A. The generalized subband analysis on the basis of unitary transformations // Belgorod State University Scientific Bulletin. Economics. Information technologies. – 2015. – No. 7(204). – Vol. 34/1. – pp. 97-104.

11. Chernomorets A.A., Bolgova E.V. On the analysis of data based on the cosine transformation // Belgorod State University Scientific Bulletin. History. Political science. Economics. Information technologies. – 2015. – No. 1(198). – Vol. 33/1. – pp. 68-73.

12. Ahmed N., Rao K. Orthogonal transformations in digital signal processing. – M.: Communication, 1980. – 248 p.

13. Zalmanzon L.A. Fourier, Walsh, Haar transforms and their application in control, communications and other areas. – M.: Nauka, 1989. – 496 p.

14. Jane B. Digital image processing. – M.: Tekhnosphere, 2007. – 584 p.

15. Gonzalez R., Woods R. Digital image processing. 3rd edition, corrected and expanded. – M.: Tekhnosphere, 2012. – 1104 p.

16. Rabiner L., Gold G. Theory and application of digital signal processing. – M.: Mir, 1988. – 512 p.

17. Sergienko A.B. Digital signal processing: textbook for students universities. – St. Petersburg: Peter, 2002. – 603 p.: ill.

18. Gold B., Rader Ch. Digital signal processing: trans. from English / B. Gold – M.: Sov. radio, 1973.

19. Nussbaumer G. Fast Fourier transform and algorithms for calculating convolutions / G. Nussbaumer; ed.: V.M. Amerbaev, T.E. Krenkel; lane from English: Yu.F. Kasimov, I.P. Pchelintsev. – M.: Radio and Communications, 1985. – 248 p.

20. Pratt W. Digital image processing. – M.: Mir, 1982. – 312 p.

Урсол Денис Владимирович, кандидат технических наук, инженер-программист
Тихонский Николай Антонович, магистрант кафедры прикладной информатики и информационных технологий
Черноморец Дарья Андреевна, аспирант кафедры информационно-телекоммуникационных систем и технологий
Болгова Евгения Витальевна, кандидат технических наук, доцент кафедры прикладной информатики и информационных технологий
Черноморец Андрей Алексеевич, доктор технических наук, доцент, профессор кафедры прикладной информатики и информационных технологий

Ursol Denis Vladimirovich, Candidate of Technical Sciences, Software Engineer
Tikhonsky Nikolay Antonovich, master's student of the Department of Applied Informatics and Information Technologies
Chernomorets Darya Andreevna, postgraduate student of the Department of Information and Telecommunications Systems and Technologies
Bolgova Evgeniya Vitalievna, Candidate of Technical Sciences, Associate Professor of the Department of Applied Informatics and Information Technologies
Chernomorets Andrey Alekseevich, Doctor of Technical Sciences, Associate Professor, Professor of the Department of Applied Informatics and Information Technologies

АВТОМАТИЗАЦИЯ И УПРАВЛЕНИЕ
AUTOMATION AND CONTROL

УДК 004.896 **DOI: 10.18413/2518-1092-2022-8-3-0-3**

Аль-Обайди А.М.Ж.

О ПЛАНИРОВАНИИ АКТИВНОСТИ СЕНСОРНЫХ УЗЛОВ В БЕСПРОВОДНОЙ СЕНСОРНОЙ СЕТИ

Компания zuhoor Al_Ameer general contracting, Джадрия, Багдад 10070, Ирак

e-mail: 1229004@bsu.edu.ru

Аннотация
В данной статье представлен новый алгоритм для оптимизации планирования активности сенсорных узлов в беспроводных сенсорных сетях. Алгоритм основан на использовании обучаемого автомата, который позволяет определить оптимальное количество активных сенсорных узлов, необходимых для обеспечения контроля над всеми объектами в заданной области. Обучение автомата производится с использованием модели беспроводной сенсорной сети, которая позволяет учесть различные параметры сети, такие как плотность распределения узлов и объектов, размер зоны покрытия и другие факторы.
Эффективность предложенного алгоритма оценивается путем сравнения его результатов с результатами других алгоритмов, используемых для планирования активности узлов. Результаты показывают, что предложенный алгоритм позволяет существенно снизить энергопотребление сенсорной сети за счет уменьшения числа активных узлов, что в свою очередь позволяет увеличить время автономной работы сети и улучшить ее производительность.
Таким образом, данный алгоритм может быть использован для оптимизации работы беспроводных сенсорных сетей и повышения их эффективности в условиях изменяющихся параметров окружающей среды и требований к контролю объектов.
Ключевые слова: сенсорные сети; самоорганизующиеся сети; обучающийся автомат; машинное обучение; энергоэффективность
Для цитирования: Аль-Обайди А.М.Ж. О планировании активности сенсорных узлов в беспроводной сенсорной сети // Научный результат. Информационные технологии. – Т.8, №3, 2023. – С. 19-26. DOI: 10.18413/2518-1092-2022-8-3-0-3

Al-Obaidi A.M.Zh.

ABOUT PLANNING THE ACTIVITY OF SENSOR NODES IN A WIRELESS SENSOR NETWORK

zuhoor Al_Ameer general contracting company, Jadriyah, Baghdad 10070, Iraq

e-mail: 1229004@bsu.edu.ru

Abstract
This article presents a new algorithm for optimizing the activity planning of sensor nodes in wireless sensor networks. The algorithm is based on the use of a trainable automaton, which allows you to determine the optimal number of active sensor nodes necessary to ensure control over all objects in a given area. The machine is trained using a wireless sensor network model, which allows you to take into account various network parameters, such as the density of distribution of nodes and objects, the size of the coverage area and other factors.
The effectiveness of the proposed algorithm is evaluated by comparing its results with the results of other algorithms used for node activity planning. The results show that the proposed algorithm can significantly reduce the power consumption of the sensor network by reducing the number of active nodes, which in turn allows you to increase the battery life of the network and improve its performance.

Thus, this algorithm can be used to optimize the operation of wireless sensor networks and increase their efficiency in conditions of changing environmental parameters and requirements for monitoring objects.
Keywords: sensor networks; self-organizing networks; learning automaton; machine learning; energy efficiency
For citation: Al-Obaidi A.M.Zh. About planning the activity of sensor nodes in a wireless sensor network // Research result. Information technologies. – T. 8, №3, 2023. – P. 19-26. DOI: 10.18413/2518-1092-2022-8-3-0-3

ВВЕДЕНИЕ

Беспроводные сенсорные сети (БСС) являются одним из перспективных направлений развития цифровых технологий в XXI веке. Одной из сфер применения БСС является автоматический сбор информации об окружающей среде и управляемых объектах. Особенно БСС полезны там, где присутствие человека из-за внешних факторов невозможно, либо сбор информации необходимо производить продолжительное время.

Использование подобных сетей ориентированно на применение в таких сферах, как информационная инфраструктура, цифровая трансформация промышленности, системы управления, государственное управление, «умный город» и цифровое здравоохранение.

Как правило, узлом сети является сенсор, содержащий набор датчиков, процессор с внутренней памятью, приемо-передатчик и автономный элемент питания. Сенсорный узел сети потребляет электроэнергию необходимую для сбора, обработки и передачи информации, также, в зависимости от топологии, электроэнергия расходуется и на маршрутизацию пакетов, получаемых от других узлов. Время работы сенсорного узла БСС напрямую зависит от ресурса элемента питания.

Элемент питания имеет ограниченный энергоресурс, при исчерпании которого узел сети прекращает функционировать, что может крайне негативно сказаться на работе БСС в целом. В результате большая часть исследований по увеличению срока работы узла сети связана с уменьшением энергопотребления.

Выбор энергоэффективных методов и алгоритмов сбора и передачи информации между является одной из основных актуальных научных проблем при проектировании БСС.

В рамках данной работы рассмотрен алгоритм адаптивного планирования работы узлов БСС на основе обучающихся автоматов.

ОБУЧАЮЩИЕСЯ АВТОМАТЫ

Обучающиеся автоматы – это модель машинного обучения. Эта модель является абстрактной моделью, которая выбирает подходящее действие из конечного набора действий и выполняет его в случайной среде [2]. Затем с помощью сигнала подкрепления окружающая среда оценивает выбранное действие и реагирует на автоматы. Чтобы выбрать следующее действие, автоматы сначала обновляют свое внутреннее состояние на основе выбранного действия, и сигнал принимается. Таким образом, следуя определенному правилу, автоматы находят наилучшее выходное решение, и для этого автомат непрерывно взаимодействует с окружающей средой, чтобы принять соответствующее решение для соответствующего действия.

ПОДХОД К ПОСТРОЕНИЮ ЭНЕРГОЭФФЕКТИВНОЙ БСС

В настоящее время наблюдается стремительный прогресс в области беспроводной связи. Эта беспроводная сенсорная сеть включает в себя сенсорные устройства, которые оснащены одним или несколькими датчиками, одним или несколькими приемопередатчиками, ресурсами хранения данных для обработки и возможными исполнительными механизмами. Это устройство собирает, хранит и обрабатывает информацию об окружающей среде, в которой оно развернуто. Таким образом, сенсорная сеть плотно связана с ограниченными ресурсами и динамической топологией. Проблема нехватки электроэнергии заключается в ограничении размера и веса датчиков. Энергопотребление сенсорной сети напрямую связано с увеличением срока службы и

эксплуатационными характеристиками сенсорной сети. Таким образом, необходимо оптимизировать потребление энергии при работе сенсорной сети. В последнее время было проведено много исследовательских работ в области оптимизации энергопотребления беспроводных сенсорных сетей [3].

Чтобы сделать сенсорную сеть энергоэффективной, управление питанием сенсорных узлов может осуществляться следующим образом:

1. Планирование узлов с использованием активного и спящего режимов.
2. Управление дальностью передачи между беспроводными узлами.
3. Использование энергоэффективных методов маршрутизации и сбора данных.

В данной работе основное внимание было уделено методу планирования активности узла для решения проблемы мониторинга контролируемых объектов. Этот метод обеспечивает эффективный результат при случайном размещении датчиков, которые отслеживают фиксированный список контролируемых объектов (целей). Существуют некоторые требования к производительности, такие как подключение к маршрутизации, покрытие сети, требования к резервированию и т.д. которые используются для планирования работы узлов. Что касается проблемы обслуживания целей, то здесь было сделано предположение о том, что каждый узел радиодиапазона способен поддерживать свой маршрут. Таким образом, каждая цель в сети обслуживается более чем одним сенсорным узлом, где избыточные узлы переводятся в спящий режим, не влияя на покрытие.

В данной работе предлагается использовать подход на основе обучающихся автоматов (ОА) для планирования активности сенсорных узлов. Данный метод позволяет найти минимальный набор активных сенсорных узлов, которые отслеживают максимальное количество контролируемых объектов (КО) в любой момент времени.

Данный алгоритм является итерационным. При этом в качестве исходных данных на каждом сенсорном узле задается вероятность выбора ОА начального состояния узла, «активное», либо «сон» со значением 0,5. От куда следует, что оба состояния равновероятны.

Все сенсорные узлы имеют фиксированную дальность действия R. Количество контролируемых объектов задается значением M, а количество сенсорных узлов значением N. При этом $1 \le j \le M$, $1 \le i \le N$.

На этапе обучения сенсорный узел выбирается случайным образом. Используя ОА, каждый сенсорный узел выбирает свое состояние. Затем он передает пакет сообщения, включая всю необходимую информацию, остальным сенсорным узлам.

Вектор вероятности состояния автомата в узле i в момент времени t равен

$$P_i(t) = [p_{(i,1)}(t), p_{(i,2)}(t)] \quad (1)$$

Функция обратной связи – это двоичная функция, которая выдает вознаграждение всякий раз, когда зона покрытия сети улучшается. Проще говоря, если совокупное состояние сенсорных узлов, выбранных командой из N обучающихся автоматов, приводит к улучшению зоны покрытия сети, тогда увеличивается вероятность состояний ОА, которые сформировали такое решение [6].

Пусть $J = \{j_1(t), j_2(t), \ldots, j_N(t)\}$ обозначает действие команды ОА. При этом $J^* = \{j_1^*(t), j_2^*(t), \ldots j_N^*(t)\}$ - лучшее совокупное действие команды ОА на данный момент, которое обеспечивает наибольшее покрытие. Таким образом, основная идея заключается в использовании обучения с подкреплением, то есть наградить ОА, совокупное действие которых обеспечивает получение максимально возможного результата на данный момент, т.е. до момента времени t.

Обновление ОА в узле i задается соотношением:

$$p_{(i,j)}(t+1) = (1-\lambda)\delta_{(i,j)} + \lambda p_{(i,j)}(t) \quad (2)$$

где

$$\delta_{(i,j)} = \begin{cases} 1 \text{ если } j = j_i^*(t) \\ 0 \text{ в других случаях} \end{cases} \quad (3)$$

λ – является параметром обновления и не зависит от времени.

Если $j \neq j_i^*(t)$, то $p_{(i,j)}(t+1)$ уменьшается путем умножения на λ, λ <1.

$$p_{(i,j)}(t+1) = \lambda p_{(i,j)}(t) \quad (4)$$

Если $j = j_i^*(t)$, тогда $p_{(i,j)}(t+1)$ увеличивается на

$$p_{(i,j)}(t+1) - p_{(i,j)}(t) = \left[(1-\lambda) + \lambda p_{(i,j)}(t)\right] - p_{(i,j)}(t) \quad (5)$$

$$p_{(i,j)}(t+1) - p_{(i,j)}(t) = (1-\lambda) + p_{(i,j)}(t)(\lambda - 1) \quad (6)$$

$$p_{(i,j)}(t+1) - p_{(i,j)}(t) = (1-\lambda)(1 - p_{(i,j)}(t)) \geq 0 \quad (7)$$

СЛОВЕСНОЕ ОПИСАНИЕ АЛГОРИТМА ПЛАНИРОВАНИЯ РАБОТЫ СЕНСОРНЫХ УЗЛОВ

Предлагаемый алгоритмом в данной работе применяется для планирования работы сенсорных узлов. Этот алгоритм помогает в поиске наилучшего набора активных датчиков, которые отслеживают максимальное количество контролируемых объектов. Процесс работы алгоритма разделен на три фазы, которые включают в себя начальную стадию, фазу обучения и фазу мониторинга контролируемых объектов. Начальный этап начинается с широковещательной передачи сообщения, содержащего информацию о сенсорном узле, его соседям. Этот этап заканчивается получением ответной информации от соседей и информацией об отслеживаемых контролируемых объектов. Затем фаза обучения начинается с выбора состояния узла с помощью обучающих автоматов и заканчивается предоставлением соответствующего значения вектора вероятности для выбора соответствующего состояния. Заключительным этапом является фаза мониторинга контролируемых объектов — она заключается в выборе наилучшего состояния датчика с использованием обучающихся автоматов

ВЫЧИСЛИТЕЛЬНЫЕ ЭКСПЕРИМЕНТЫ

Оценка предлагаемого алгоритма адаптивного планирования (ААП) производится путем проведения компьютерного моделирования сети с сенсорными узлами и объектами контроля, размещенными случайным образом в пространстве размером 600 ×600 метров.

Для всех экспериментов используются следующие параметры:

• N – количество сенсорных узлов, развернутых случайным образом. Его значение изменяется в интервале от 4 до 90;

• M – количество случайно развернутых целей. Его значение изменяется в интервале от 3 до 60;

• R – дальность работы датчика. Его значение изменяется в интервале от 50 м до 600 м.

• Параметр обучения;

– Лямбда(λ), его значение изменяется в интервале 0,0001 и 0,4;

– Эпсилон(ε), его значение остается постоянным, которое составляет 0,01 для всех экспериментов.

Влияние количества сенсоров

Цель первого эксперимента являлось исследование зависимости между количеством сенсорных узлов и получаемым минимальным количеством активных сенсорных узлов в сети. Для этого число сенсорных узлов берется между 20 и 30 с шагом 1. Полученные результаты приведены в таблицах. В таблице представлены результаты, полученные при размещении 10 и 15 объектов контроля, дальность обнаружения сенсорного узла составляла 300 м и 400 м.

Таблица 1

Результаты эксперимента, полученные при размещении 10 объектов контроля, дальности обнаружения сенсорного узла 300 м, количество сенсорных узлов изменялось в пределах от 20 до 30

Table 1

The experimental results obtained by placing 10 control objects, the detection range of the sensor node is 300 m, the number of sensor nodes varied from 20 to 30

Количество сенсорных узлов (N)	Среднее количество активных сенсоров
20	2,136
21	2,132
22	2,130
23	2,125
24	2,109
25	1,879
26	1,768
27	1,520
28	1,280
29	1,113
30	1,023

Таблица 2

Результаты эксперимента, полученные при размещении 15 объектов контроля, дальности обнаружения сенсорного узла 150 м, количество сенсорных узлов изменялось в пределах от 20 до 30

Table 2

The experimental results obtained by placing 15 control objects, the detection range of the sensor node is 150 m, the number of sensor nodes varied from 20 to 30

Количество сенсорных узлов (N)	Среднее количество активных сенсоров
20	7,168
21	7,120
22	6,972
23	6,832
24	6,721
25	6,287
26	6,176
27	6,143
28	6,111
29	5,835
30	5,439

На рисунке 1 показано, что при малой дальности обнаружения требуется больше сенсорных узлов для обслуживания объектов контроля. Но увеличение дальности обнаружения приводит к уменьшению количества активных сенсорных узлов, используемых для наблюдения за объектами

контроля, что показывает ту же тенденцию результатов, что и при больших значениях дальности обнаружения, показанных на рисунке 2.

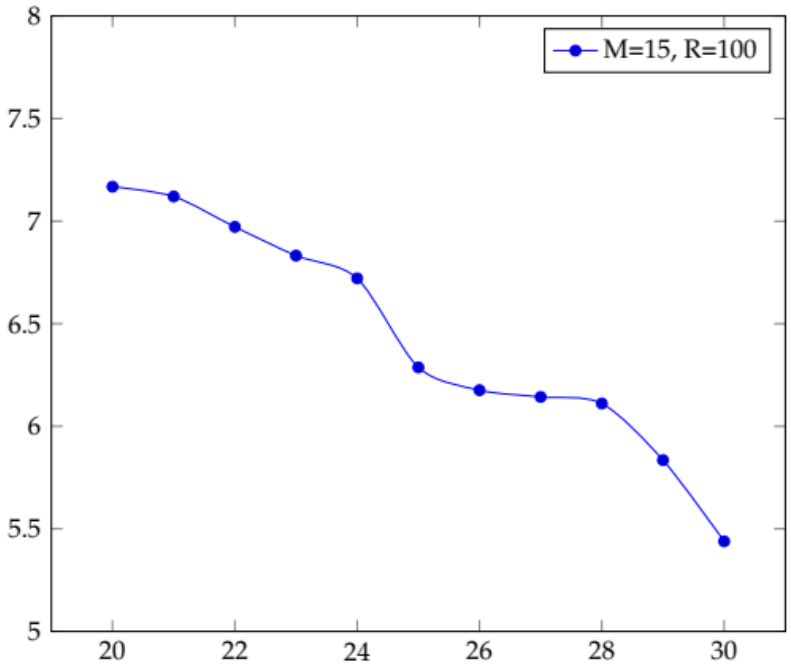


Рис. 1. График, показывающий влияние увеличения числа сенсорных узлов при обслуживании 15 объектов контроля и дальности обнаружения 100 м на среднее количество активных сенсорных узлов

Fig. 1. A graph showing the effect of increasing the number of sensor nodes when servicing 15 monitoring objects and a detection range of 100 m on the average number of active sensor nodes

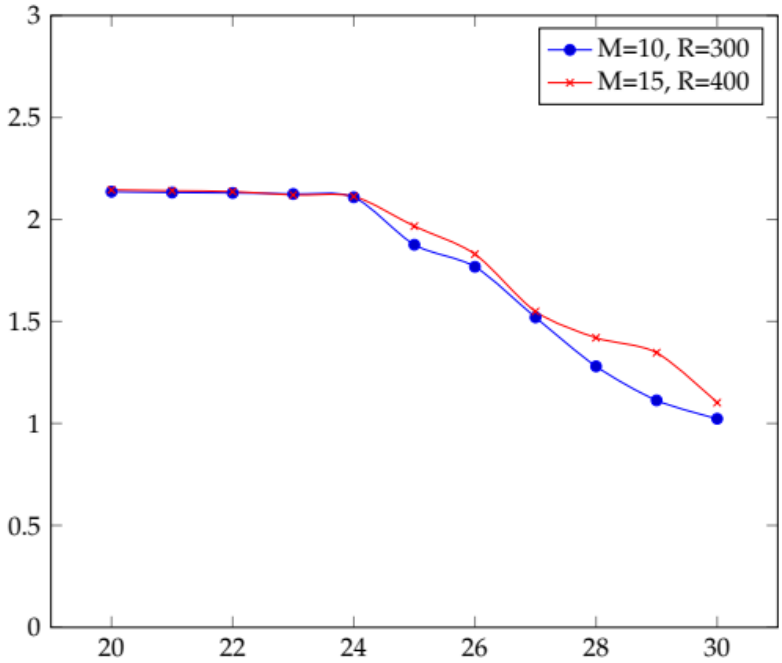


Рис. 2. График, показывающий влияние увеличения числа сенсорных узлов на среднее количество активных сенсорных узлов

Fig. 2. A graph showing the effect of increasing the number of sensor nodes on the average number of active sensor nodes

Эксперимент по сравнению разработанного алгоритма с другими алгоритмами планирования проводится для определения эффективности предложенного алгоритма обучения. Для этого взят алгоритм планирования LADSC из статьи [7] и алгоритм прямого перебора. Эти алгоритмы сравниваются на основе изменения плотности сенсорных узлов. Алгоритмы оцениваются по их свойству использовать минимальное количество активных сенсорных узлов для обслуживания всех объектов контроля.

Таблица 3

Сравнение предложенного алгоритма с алгоритмом LADSC, и проверка правильности результатов с использованием алгоритма прямого перебора при 9 объектах контроля, количество сенсорных узлов изменялось в пределах 9 до 25, дальность действия 100 м

Table 3

Comparison of the proposed algorithm with the LADSC algorithm, and verification of the correctness of the results using the direct search algorithm with 9 control objects, the number of sensor nodes varied from 9 to 25, the range of 100 m

Количество сенсорных узлов (N)	**Предлагаемый алгоритм**		**LADSC**		**Прямой перебор**	
	Количество активных сенсоров	Количество обслуживаемых объектов	Количество активных сенсоров	Количество обслуживаемых объектов	Количество активных сенсоров	Количество обслуживаемых объектов
10	6,0	9	7,0	9	6	9
12	5,0	9	7,0	9	5	9
14	5,0	9	6,0	8	5	9
16	5,0	9	6,0	9	5	9
18	5,0	9	4,0	7	5	9
20	5,0	9	6,0	9	4	9
22	4,0	9	5,0	9	4	9
24	4,0	9	5,0	9	4	9

Из полученных результатов можно заметить, что предложенный алгоритм планирования обеспечивает лучшие результаты по сравнению с алгоритмом LADSC для получения меньшего количества активных сенсорных узлов для мониторинга всех объектов контроля. Даже в некоторых случаях алгоритм LADSC дает эффект использования меньшего количества активных сенсорных узлов, но он не в состоянии охватить все объекты контроля. Таким образом, он менее эффективен, чем предлагаемый алгоритм. Это обеспечивает тот факт, что предлагаемый алгоритм потребляет меньшее количество энергии, чем алгоритм LANDS, и поэтому помогает максимально увеличить срок службы сенсорной сети. И получать оптимизированные результаты.

ЗАКЛЮЧЕНИЕ

Предлагаемый алгоритм для экономии энергии позволяет производить настройку сенсорных узлов, что представляет собой концепцию обучения. Таким образом, они могут в автономном режиме выбрать свое состояние, быть активными или находится в ожидании. Предполагается, что в активном состоянии сенсорные узлы обслуживают доступные контролируемые объекты, затрачивая некоторое количество энергии. Предложенный алгоритм обеспечивает методологию поиска минимально количества активных сенсорных узлов, которые обслуживают все контролируемые объекты.

Другими словами, это можно определить так: при использовании минимального количества активных сенсорных узлов используется минимальная энергия сети, и, таким образом, эта концепция помогает экономить энергию в беспроводной сенсорной сети.

Этот алгоритм применим как к малым, так и к большим беспроводным сенсорным сетям. Но при проведении экспериментов в сети такого типа следует учитывать значение параметра обучения, чтобы получить хорошие результаты. Дальность действия сенсорного узла оказывает одинаковое влияние на оба типа сетей. И если провести сравнение, при одинаковых условиях, предложенный алгоритм показывает более высокие результаты, чем выбранные для сравнения алгоритмы.

Список литературы

1. Butun I., Morgera S.D., Sankar R.A. Survey of Intrusion Detection Systems in Wireless Sensor Networks // IEEE communications surveys & tutorials – 2013, P. 266-282.
2. Rassam M.A, Maarof M.A., Zainal A. An Efficient Distributed Anomaly Detection Model for Wireless Sensor Networks // Knowledge-Based Systems – 2014, 60 P. 44-57.
3. Кучерявый А.Е., Прокопьев А.В., Кучерявый Е.А. Самоорганизующиеся сети. СПб.: Любавич. 2011. 312 с.
4. Najim, K., Poznyak, A.S.: Learning Automata: Theory and Applications. Pergamon Press, Oxford, 1994, 238 pages.
5. Thathachar, M.A.L., Sastry, P.S.: Networks of Learning Automata: Techniques for Online Stochastic Optimization. Kluwer Academic Publishers 2004, Springer New York, NY 288 pages.
6. Mostafaei H., Meybodi M.R.: Maximizing lifetime of target coverage in wireless sensor networks using learning automata. Wireless Personal Communications 71(2), 2013, P. 1461–1477.
7. Habib Mostafaei1, Mehdi Esnaashari, Mohammad Reza Meybodi. [Электронный ресурс] ‘A Coverage Monitoring algorithm based on Learning Automata for Wireless Sensor Networks’. In: Cornell University Library (Sept. 2014). URL: https://arxiv.org/ftp/arxiv/papers/1409/1409.1515.pdf. (дата обращения: 01.06.2023)

References

1. Butun I., Morgera S.D., Sankar R.A. Survey of Intrusion Detection Systems in Wireless Sensor Networks // IEEE communications surveys & tutorials – 2013, P. 266-282.
2. Rassam M.A, Maarof M.A., Zainal A. An Efficient Distributed Anomaly Detection Model for Wireless Sensor Networks // Knowledge-Based Systems – 2014, 60 P. 44-57.
3. Kucheryavy A.E., Prokopyev A.V., Kucheryavy E.A. Self-organizing networks. St. Petersburg: Lyubavich. 2011. 312 p.
4. Najim, K., Poznyak, A.S.: Learning Automata: Theory and Applications. Pergamon Press, Oxford, 1994, 238 pages.
5. Thathachar, M.A.L., Sastry, P.S.: Networks of Learning Automata: Techniques for Online Stochastic Optimization. Kluwer Academic Publishers 2004, Springer New York, NY 288 pages.
6. Mostafaei H., Meybodi M.R.: Maximizing lifetime of target coverage in wireless sensor networks using learning automata. Wireless Personal Communications 71(2), 2013, P. 1461–1477.
7. Habib Mostafaei1, Mehdi Esnaashari, Mohammad Reza Meybodi. [Electronic resource] ‘A Coverage Monitoring algorithm based on Learning Automata for Wireless Sensor Networks’. In: Cornell University Library (Sept. 2014). URL: https://arxiv.org/ftp/arxiv/papers/1409/1409.1515.pdf. (data access: 01.06.2023)

Аль-Обайди Амир Мохаммед Жасим, руководитель проектов компании Al_Ameer general contracting company
Al-Obaidi Amir Mohammed Jasim, project manager, zuhoor Al_Ameer general contracting company

ИСКУССТВЕННЫЙ ИНТЕЛЛЕКТ И ПРИНЯТИЕ РЕШЕНИЙ
ARTIFICIAL INTELLIGENCE AND DECISION MAKING

УДК 004.03 **DOI: 10.18413/2518-1092-2022-8-3-0-4**

Бузов П.А.[1]
Игитян Е.В.[2]
Польщиков К.А.[2]
Лебедева А.С.[2]

ОЦЕНИВАНИЕ ВЕРОЯТНОСТНО-ВРЕМЕННЫХ ХАРАКТЕРИСТИК ЧЕЛОВЕКО-МАШИННОГО ДИАЛОГА

[1)] АО «СОФТКОННЕКТ», ул. Рабочая, 14, офис 1 Белгород, 308013, Россия
[2)] Белгородский государственный национальный исследовательский университет, ул. Победы, 85, Белгород, 308015, Россия

e-mail: medvedeva_e@bsu.edu.ru

Аннотация
Работа посвящена оцениванию вероятностно-временных характеристик человеко-машинного диалога. Авторами проведен ряд экспериментов по исследованию характеристик диалоговой системы «Алиса» и оцениванию показателей, характеризующих особенности её пользователей. Результат показал, что отклонения характеристик человеко-машинных диалогов, полученных в результате экспериментов, от характеристик, вычисленных в результате моделирования, не превышают ±3,8 % при доверительной вероятности 0,95. В частности, авторами разработана модель человеко-машинного диалога, позволяющая учитывать характеристики диалоговой системы.
Ключевые слова: диалоговые системы; эффективность чат-бот-приложений; человеко-машинный диалог
Для цитирования: Бузов П.А., Игитян Е.В., Польщиков К.А., Лебедева А.С. Оценивание вероятностно-временных характеристик человеко-машинного диалога // Научный результат. Информационные технологии. – Т.8, №3, 2023. С. 27-33. DOI: 10.18413/2518-1092-2022-8-3-0-4

Buzov P.A.[1]
Igityan E.V.[2]
Polshchikov K.A.[2]
Lebedeva A.S.[2]

ESTIMATION OF PROBABILISTIC-TEMPORAL CHARACTERISTICS OF HUMAN-MACHINE DIALOGUE

[1)] JSC "SOFTCONNECT", 14 Rabochaya str.,office 1 Belgorod, 308013, Russia
[2)] Belgorod State National Research University, 85 Pobedy Str., Belgorod, 308015, Russia

e-mail: medvedeva_e@bsu.edu.ru

Abstract
The paper is devoted to the evaluation of the probabilistic-temporal characteristics of the human-machine dialogue. The authors conducted a number of experiments to study the characteristics of the Alice dialog system and evaluate the indicators characterizing the characteristics of its users. The result showed that the deviations of the characteristics of human-machine dialogues obtained as a result of experiments from the characteristics calculated as a result of modeling do not exceed ± 3.8% with a confidence probability of 0.95. In particular, the authors have developed a human-machine dialogue model that allows taking into account the characteristics of the dialog system.
Keywords: dialog systems; efficiency of chatbot applications; human-machine dialogue
For citation: Buzov P.A., Igityan E.V., Polshchikov K.A., Lebedeva A.S. Estimation of probabilistic-temporal characteristics of human-machine dialogue // Research result. Information technologies. – T.8, №3, 2023. – P. 27-33. DOI: 10.18413/2518-1092-2022-8-3-0-4

ВВЕДЕНИЕ

Одним из важнейших направлений автоматической обработки естественно-языковых данных является разработка и совершенствование интеллектуальных диалоговых систем и их упрощенных версий – чат-ботов. Эти системы стали все чаще применяться в коммерческих проектах, где они используются в общении с клиентами для помощи в покупке товаров, технической поддержки, навигации по сайтам и т.д. Диалоговые системы используются в качестве интеллектуальных модулей общения социальных роботов, которые ухаживают за больными, престарелыми людьми.

Задача диалоговых систем – не только продемонстрировать пользователю свои коммуникативные способности, но и предоставить максимально точный ответ на вопрос, заданный на естественном языке [5]. Диалоговые системы, основанные на использовании программных средств обработки естественно-языковых данных и предназначенные для выдачи ответов на заданные пользователями вопросы, будем именовать вопросно-ответными системами. Современные вопросно-ответные системы обладают модулями контентной аналитики, позволяющими собирать и упорядочивать информацию, а также использовать машинное обучение на основе нейронных сетей. Известными примерами таких систем являются вопросно-ответные сервисы «ChatGPT» (OpenAI), «Watson» (IBM), виртуальные ассистенты «Алиса» (Яндекс), «Siri» (Apple), «Google Assistant», «Amazon Alexa», «Cortana» (Microsoft). Специалисты утверждают, что существует потребность в адекватных средствах оценивания характеристик диалоговых систем. Недостает общепризнанных методов, моделей, метрик, критериев и количественных показателей, на основе которых можно было бы обеспечить обоснованность принимаемых решений о качестве функционирования диалоговых систем, объективно выявить их недостатки и определить направления дальнейшего совершенствования.

Существующие инструментарии оценки характеристик диалоговых систем позволяют оценить обобщенный показатель выполнения диалоговой системой вопросно-ответных функций без привязки к особенностям тех или иных пользователей или их групп [2]. При этом в таких средствах оценки параметры нечеткого вывода настроены на «среднего пользователя» и не учитывают такие человеческие особенности, как, например, терпеливость в задавании вопросов или сообразительность при получении недостаточного точного и полного ответа и т.п. Чтобы обеспечить учет подобных пользовательских особенностей потребуется многократное усложнение предложенных алгоритмов оценивания, основанных на применении нечеткого вывода и нейросетевой настройки, т.к. для каждой группы пользователей будет необходимо:

- вычислить частные показатели точности, лаконичности и полноты ответов;
- сформировать отдельные обучающие выборки;
- создать отдельные системы нечеткого вывода;
- настроить параметры каждой системы нечеткого вывода с помощью нейросетевого обучения;
- оценить обобщенный показатель выполнения диалоговой системой вопросно-ответных функций, реализовав процедуры нечеткого вывода.

Такое усложнение процесса вычисления характеристик диалоговой системы неприемлемо, поэтому для оценивания соответствия характеристик вопросно-ответной системы особенностям её пользователей требуется модель человеко-машинного диалога, результаты разработки которой представлены ниже.

РАЗРАБОТКА МОДЕЛИ ЧЕЛОВЕКО-МАШИННОГО ДИАЛОГА

Пусть у пользователя имеется потребность в получении определенной информации. Для удовлетворения этой потребности пользователь вступает в человеко-машинный диалог. Следовательно, целью этого диалога является удовлетворение определенной информационной потребности пользователя. На рисунке 1 представлен вероятностно-временной граф, моделирующий процесс человеко-машинного диалога.

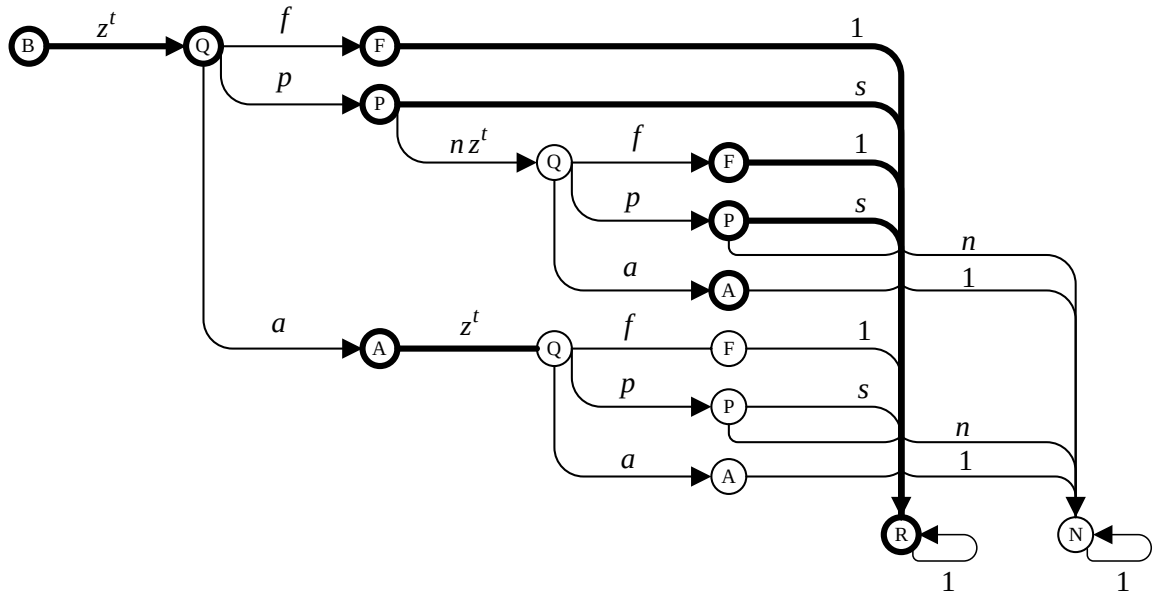


Рис. 1. Вероятностно-временной граф, моделирующий процесс человеко-машинного диалога
Fig. 1. Probabilistic-time graph modeling the process of human-machine dialogue

Начальное состояние моделируемого процесса обозначено «B». В ходе диалога человеком задается вопрос, при этом процесс моделирования переходит в состояние «Q». Далее на заданный вопрос система выдает ответ. Если полученный пользователем ответ в полной мере содержит требуемую информацию (такой вариант соответствует переходу из состояния «Q» в состояние «F»), то считается, что цель диалога достигнута, и моделируемый процесс переходит из состояния «F» в терминальное состояние «R». Если в полученном ответе практически отсутствует требуемая информация (переход из состояния «Q» в состояние «A»), пользователь может прекратить диалог без достижения цели (переход из состояния «P» в терминальное состояние «N»), посчитав систему бесполезной для удовлетворения его информационной потребности, а может задать системе уточняющий вопрос (переход из состояния «P» в состояние «Q»).

В случае получения ответа, в котором частично содержится требуемая информация, моделируемый процесс переходит из состояния «Q» в состояние «P». При этом возможны следующие варианты:

1) пользователь удовлетворяется полученным ответом, цель диалога достигается, а моделируемый процесс переходит из состояния «P» в терминальное состояние «R»;

2) пользователь не удовлетворяется полученным ответом и задает уточняющий вопрос, моделируемый процесс переходит из состояния «P» в состояние «Q»;

3) пользователь не удовлетворяется полученным ответом и прекращает задавать вопросы, диалог заканчивается, не достигнув цели, что соответствует переходу из состояния «P» в терминальное состояние «N».

В результате выполнения конечного числа указанных выше переходов диалог завершается, т.е. моделируемый процесс оказывается в терминальном состоянии «R» или в терминальном состоянии «N». Граф, изображенный на рисунке 1, построен для случая, при котором число задаваемых пользователем уточняющих вопросов, ограничивается величиной $m=1$. Переход по стрелке из какой-либо вершины графа в соседнюю вершину моделируется с помощью некоторой функции:

$$\Omega(z)=\omega\cdot z^{\tau}, \qquad (1)$$

где ω – вероятность осуществления перехода; τ – время, в течение которого осуществляется переход; z – вспомогательный параметр, используемый в теории производящих функций.

Переход из состояния «N» в состояние «Q» моделируется функцией z^t, где t – средняя продолжительность по времени формулировки пользователем вопроса и выдачи системой ответа.

Переходу из состояния «Q» в состояние «F» соответствует величина f – вероятность того, что выданный системой ответ в полной мере содержит требуемую пользователю информацию. Переходу из состояния «Q» в состояние «P» соответствует величина p – вероятность выдачи системой ответа, в котором частично содержится требуемая информация. Из состояния «Q» возможен также переход в состояние «A». Этому событию соответствует величина a – вероятность отсутствия в выданном системой ответе информации, требуемой пользователю. По условию нормировки вероятность a может быть вычислена с помощью выражения:

$$a = 1 - f - p.$$

Из состояния «F» в терминальное состояние «R» осуществляется переход с вероятностью, равной 1. Переходу из состояния «P» в состояние «R» соответствует величина s – вероятность удовлетворенности пользователя полученным ответом, в котором частично содержится требуемая информация. Переход из состояния «P» в состояние «Q» моделируется функцией nz^t, где n – вероятность неудовлетворенности пользователя полученным ответом, в котором частично содержится требуемая информация. По условию нормировки вероятностей величину n можно вычислить с помощью выражения:

$$n = 1 - s.$$

Переход из состояния «A» в состояние «Q» моделируется функцией z^t. Переходу из состояния «P» в состояние «N» соответствует вероятность n. Наконец, переход из состояния «A» в терминальное состояние «N» осуществляется с вероятностью, равной 1. В результате эквивалентных преобразований графа, получены выражения для вычисления $\Omega_{BR}^1(z)$ – функции перехода из состояния «B» в состояние «R» при $m=1$, $\Omega_{BN}^1(z)$ – функции перехода из состояния «B» в состояние «N» при $m=1$, $\Omega_{BE}^1(z)$ – функции перехода из состояния «B» в состояние «E» при $m=1$. Удалось вывести соответствующие выражения для $m=2$, $m=3$ и $m=4$, выявить в них обобщающие закономерности, на основе которых получены формулы для вычисления функций $\Omega_{BR}^m(z)$, $\Omega_{BN}^m(z)$ и $\Omega_{BE}^m(z)$ при любом натуральном m:

$$\Omega_{BR}^m(z) = (f + ps) z^t \sum_{\alpha=0}^{m} \sum_{\beta=0}^{m-\alpha} \left(pnz^t\right)^{\alpha} \left(\alpha z^t\right)^{\beta} \frac{(\alpha+\beta)!}{\alpha!\beta!}, \tag{2}$$

$$\Omega_{BN}^m(z) = (pn + a) z^{(m+1)t} \sum_{\alpha=0}^{m} \left(pn\right)^{\alpha} a^{(m-\alpha)} \frac{m!}{\alpha!(m-\alpha)!}, \tag{3}$$

$$\Omega_{BE}^m(z) = \Omega_{BR}^m(z) + \Omega_{BN}^m(z). \tag{4}$$

Для вычисления вероятности достижения цели диалога можно использовать выражение:

$$r_m = \Omega_{BR}^m(z)\Big|_{z=1}. \tag{5}$$

Средняя продолжительность человеко-машинного диалога может быть вычислена по формуле:

$$u_m = \frac{d}{dz}\Omega_{BE}^m(z)\Big|_{z=1}. \tag{6}$$

РЕЗУЛЬТАТЫ РАБОТЫ

На основе применения разработанной модели проведены экспериментальные исследования по оцениванию вероятностно-временных характеристик человеко-машинного диалога. Графики зависимости величин r_m и u_m от значений величины m при $f=0,3$ и $p=0,4$ представлены на рисунках 2 и 3.

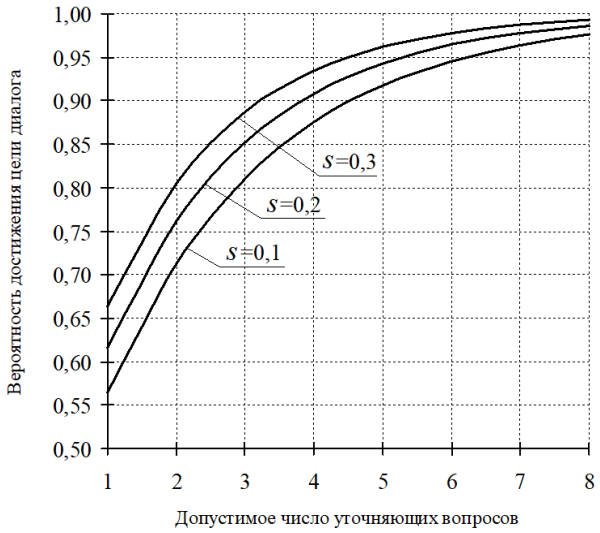


Рис. 2. Графики зависимости величины r_m от значений величины m при $f = 0,3$ и $p = 0,4$

Fig. 2. Graphs of the dependence of the magnitude on the values r_m of the magnitude m at $f = 0,3$ and $p = 0,4$

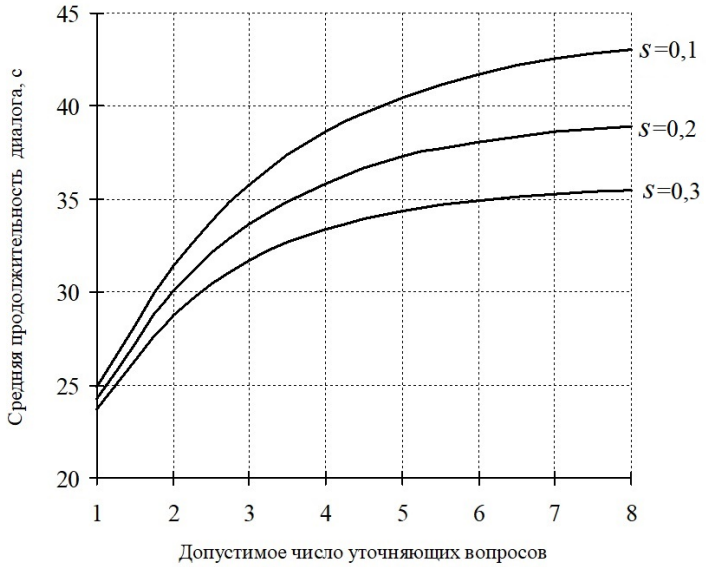


Рис. 3. Графики зависимости величины u_m от значений величины m при $f = 0,3$ и $p = 0,4$

Fig. 3. Graphs of the dependence of the magnitude on the values u_m of the magnitude m at $f = 0,3$ and $p = 0,4$

Допустим заданы требования к характеристикам диалоговой системы, например, вероятность достижения цели человеко-машинного диалога не должна быть ниже 0,9, а средняя продолжительность диалога при этом не должна превышать 40 с. Тогда анализ представленных на рисунках 2 и 3 графиков показывает, что диалоговая система с параметрами $f = 0,3$ и $p = 0,4$ рекомендуется тем пользователям, которые, в случае необходимости, готовы задать хотя бы 4 уточняющих вопроса, а также с вероятностью не ниже 0,2 будут удовлетворены полученными ответами, в которых лишь частично содержится требуемая информация.

Проведены многочисленные эксперименты по исследованию характеристик диалоговой системы «Алиса» и оцениванию показателей, характеризующих особенности её пользователей. В ходе экспериментов проводились диалоги пользователей с этой системой, фиксировались результаты, характеризующие содержание требуемой информации в выдаваемых системой ответах и удовлетворенность ими пользователей. В экспериментальных диалогах приняли участие 134 пользователя, каждый из которых провел с системой по 20 диалогов. Выполнена обработка экспериментальных данных, проведенных с участием всех пользователей, а затем получены результаты вычислений на основе разработанной модели.

Проведен статистический анализ отклонений результатов экспериментов от результатов моделирования. Анализ показал, что отклонения характеристик человеко-машинных диалогов, полученных в результате экспериментов, от характеристик, вычисленных в результате моделирования, не превышают ±3,8 % при доверительной вероятности 0,95. Высокая сходимость результатов экспериментальных исследований и моделирования свидетельствует об адекватности разработанной модели человеко-машинного диалога.

ЗАКЛЮЧЕНИЕ

Модель человеко-машинного диалога учитывает характеристики диалоговой системы, в числе которых следующие величины: вероятность выдачи системой ответа, в котором частично содержится требуемая информация; вероятность того, что выданный системой ответ в полной мере содержит требуемую пользователю информацию; вероятность отсутствия в выданном системой ответе информации, требуемой пользователю. Кроме того, модель учитывает значения параметров, характеризующих особенности пользователя: вероятность удовлетворенности пользователя полученным ответом, в котором частично содержится требуемая информация; вероятность неудовлетворенности пользователя полученным ответом, в котором частично содержится требуемая информация; максимальное количество уточняющих вопросов, задаваемых пользователем в диалоге.

Список литературы

1. Игитян Е.В., К.А. Польщиков, А.Н. Немцев Оценивание вероятностно-временных характеристик человеко-машинного диалога на естественном языке. Экономика. Информатика. – 2023. – Т. 50. – № 1. – С. 162–172.

2. Махди Т.Н., Игитян Е.В., Польщиков К.А., Корсунов Н.И. Оценивание эффективности функционирования диалоговой системы на основе применения нечеткого вывода с нейросетевой настройкой. Экономика. Информатика. – 2022. – Т. 49. – № 2. – С. 356–374.

3. Польщиков К.А., Польщикова О.Н., Игитян Е.В., Балакшин М.С. Алгоритм поддержки принятия решений по выбору средств обработки больших массивов естественно-языковых данных. Научные ведомости Белгородского государственного университета. Серия: Экономика. Информатика. – 2019. – Т. 46. – № 3. – С. 553-562.

4. Польщиков К.А., Лазарев С.А., Константинов И.С., Польщикова О.Н., Свойкина Л.Ф., Игитян Е.В., Балакшин М.С. Модель для оценки эффективности выполнения робототехнической системой коммуникативных функций. СТИН. – 2020. – № 6. – С. 4–7.

5. Агузумцян Р.В., Великанова А.С., Польщиков К.А., Игитян Е.В., Лихошерстов Р.В. О применении интеллектуальных технологий обработки естественного языка и средств виртуальной реальности для поддержки принятия решений при подборе исполнителей проектов. Экономика. Информатика. – 2021. – № 48 (2). – С. 392-404.

References

1. Igityan E.V., Polshchikov K.A., Nemtsev A.N. Estimation of probabilistic-temporal characteristics of human-machine dialogue in natural language. Economics. Information technologies. – 2023. – Vol. 50. – No. 1. – pp. 162-172.

2. Mahdi T.N., Igityan E.V., Polshchikov K.A., Korsunov N.I. Evaluation of the effectiveness of the functioning of a dialog system based on the use of fuzzy inference with neural network setup. Economics. Information technologies. – 2022. – Vol. 49. – No. 2. – pp. 356-374.

3. Polshchikov K.A., Polshchikova O.N., Igityan E.V., Balakshin M.S. Algorithm of decision support for the choice of means of processing large arrays of natural language data. Scientific Bulletin of Belgorod State University. Series: Economics. Computer science. – 2019. – Vol. 46. – No. 3. – pp. 553-562.

4. Polshchikov K.A., Lazarev S.A., Konstantinov I.S., Polshchikova O.N., Svoikina L.F., Igityan E.V., Balakshin M.S. A model for evaluating the effectiveness of performing communicative functions by a robotic system. STIN. – 2020. – No. 6. – pp. 4-7.

5. Aguzumtsyan R.V., Velikanova A.S., Polshchikov K.A., Igityan E.V., Likhosherstov R.V. On the application of intelligent natural language processing technologies and virtual reality tools to support decision-making in the selection of project performers / / Economics. Information technologies. – 2021. – № 48 (2). – Pp. 392-404.

Бузов Павел Андреевич, генеральный директор АО «СОФТКОННЕКТ»
Игитян Елена Владимировна, старший преподаватель кафедры информационных и робототехнических систем, институт инженерных и цифровых технологий
Польщиков Константин Александрович, доцент кафедры информационных и робототехнических систем программного обеспечения вычислительной техники и автоматизированных систем, институт инженерных и цифровых технологий
Лебедева Анна Сергеевна, аспирант 3 курса, направления «Информатика и вычислительная техника», кафедра прикладной информатики и информационных технологий, институт инженерных и цифровых технологий

Buzov Pavel Andreevich, General Director of SOFTCONNECT JSC
Igityan Elena Vladimirovna, Senior Lecturer, Department of Information and Robotic Systems, Institute of Engineering and Digital Technologies
Polshchikov Konstantin Alexandrovich, Associate Professor of the Department of Information and Robotic Systems of Computer Engineering and Automated Systems, Institute of Engineering and Digital Technologies
Lebedeva Anna Sergeevna, 3rd year postgraduate student, Computer Science and Computer Engineering, Department of Applied Computer Science and Information Technology, Institute of Engineering and Digital Technologies

УДК 004.855.5 **DOI: 10.18413/2518-1092-2022-8-3-0-5**

Жихарев А.Г.[1]
Черных В.С.[2]

КЛАССИФИКАЦИЯ РЕЧЕВЫХ ДАННЫХ ПО ЭМОЦИОНАЛЬНОМУ ФОНУ

[1)] Белгородский государственный технологический университет им. В.Г. Шухова, ул. Костюкова, 46, Белгород, 308012, Россия
[2)] Белгородский государственный университет, ул. Победы, 85, Белгород, 308015, Россия

e-mail: zhikharev@bsu.edu.ru

Аннотация
В данной работе рассматривается алгоритм классификации речевых данных по эмоциональному фону, разработанный авторами. В частности, описывается нейронная сеть, созданная с целью распознавания восьми различных эмоций в речи. Для обучения нейронной сети была использована обучающая выборка, полученная из датасета RAVDESS, который содержит 1440 аудиофайлов. Эти аудиофайлы содержат речь 24 актеров (12 женщин и 12 мужчин) с нейтральным североамериканским акцентом.
В работе описывается процесс обучения нейронной сети с использованием библиотеки Keras, включая архитектуру сети, размеры слоев, функции активации и методы оптимизации. Также обсуждаются этапы предварительной обработки и подготовки исходных аудиоданных перед обучением сети.
Полученные результаты исследования показывают, что разработанная нейронная сеть обладает высокой производительностью и способностью распознавать эмоции с точностью 80%.
Ключевые слова: аудиопризнаки; аудио; аудиофайл; аудиоданные; эмоциональный фон; классификация; модель; слой
Для цитирования: Жихарев А.Г., Черных В.С. Классификация речевых данных по эмоциональному фону // Научный результат. Информационные технологии. – Т.8, №3, 2023. – С. 34-44. DOI: 10.18413/2518-1092-2022-8-3-0-5

Zhikharev A.G.[1]
Chernykh V.S.[2]

CLASSIFICATION OF SPEECH DATA BY EMOTIONAL BACKGROUND

[1)] Belgorod State Technological University named after V.G. Shukhov, 46 Kostyukova Str., Belgorod, 308012, Russia
[2)] Belgorod State University, 85 Pobedy Str., Belgorod, 308015, Russia

e-mail: zhikharev@bsu.edu.ru

Abstract
In this paper, the algorithm of classification of speech data by emotional background, developed by the authors, is considered. In particular, it describes a neural network created to recognize eight different emotions in speech. To train the neural network, a training sample obtained from the RAVDESS dataset, which contains 1440 audio files, was used. These audio files contain the speech of 24 actors (12 women and 12 men) with a neutral North American accent.
The paper describes the process of training a neural network using the Keras library, including the network architecture, layer sizes, activation functions and optimization methods. The stages of preprocessing and preparation of the original audio data before training the network are also discussed.
The results of the study show that the developed neural network has high performance and the ability to recognize emotions with an accuracy of 80%.
Keywords: audio attributes; audio; audio file; audio data; emotional background; classification; model; layer

For citation: Zhikharev A.G., Chernykh V.S. Classification of speech data by emotional background // Research result. Information technologies. – T.8, №3, 2023. – P. 34-44. DOI: 10.18413/2518-1092-2022-8-3-0-5

ВВЕДЕНИЕ

В данный момент мы находимся в пике оптимистического развития искусственного интеллекта, и всё чаще слышим о создании новых нейронных сетей для решения самых разнообразных и важных проблем, начиная с постановки медицинских диагнозов и заканчивая усовершенствованием цифровых помощников [1]. В современном мире обработка и анализ больших объемов данных становятся все более важными задачами. Особое внимание уделяется анализу речевых данных, которые могут содержать важную информацию о настроении и эмоциональном состоянии говорящего. Одним из ключевых направлений в этой области является классификация речевых данных по эмоциональному фону.

В данной работе будет рассмотрен модуль для обработки голосовых данных и распознавания эмоций.

НАЧАЛЬНЫЕ УСЛОВИЯ

Выбор обучающих данных является важнейшей составляющей в обучении нейронной сети, поэтому тут необходимо выбрать наикачественнейшие данные, выбор пал на датасет RAVDESS, он обоснован тем, что аудиофайлы записаны на высококачественном оборудовании. Только аудио часть датасета RAVDESS содержит 1440 файлов (.wav) с речью 24 актеров (12 женщин, 12 мужчин) на нейтральном североамериканском акценте. Так же сами аудио файлы имеют удобные для понимания и расшифровки метки, пример на рисунке.

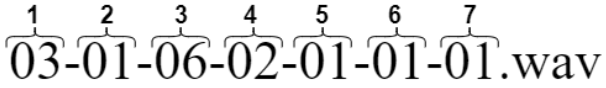


Рис. 1. Пример названия аудиофайла
Fig. 1. Example of an audio file name

1. Режим (01 = видео с аудиоряд, 02 = только видео, 03 = только аудио.
2. Вокальный канал (01 = речь, 02 = песня).
3. Эмоция (01 = нейтрально, 02 = спокойно, 03 = счастье, 04 = грусть, 05 = злость, 06 = испуг, 07 = отвращение, 08 = удивление).
4. Эмоциональная напряженность (01 = нормальная, 02 = сильная). ПРИМЕЧАНИЕ: Для "нейтральной" эмоции не существует сильной интенсивности.
5. Фраза (01 = "Дети разговаривают у двери", 02 = "Собаки сидят у двери").
6. Повторение (01 = 1-е повторение, 02 = 2-е повторение).
7. Актер (с 01 по 24. Актеры с нечетными номерами – мужчины, актеры с четными номерами – женщины).

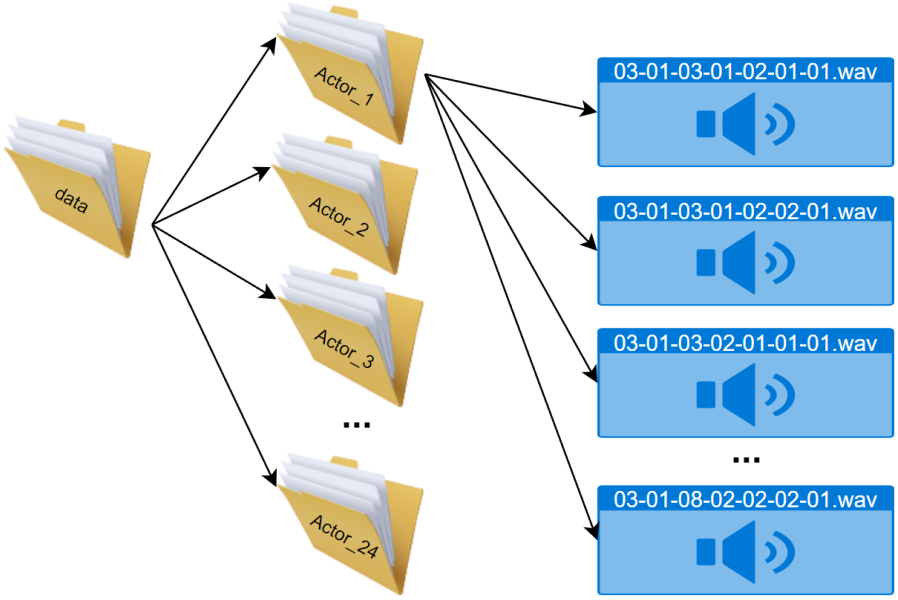


Рис. 2. Структура хранения аудиофайлов
Fig. 2. Structure of audio file storage

АУДИОПРИЗНАКИ

Эмпирическим путём было выявлено, что для получения хороших результатов подходят следующие аудиопризнаки:

1. MFCC (Mel-Frequency Cepstral Coefficients Мел-кепстральные коэффициенты) – это метод извлечения характеристик из аудиосигнала, который используется для представления звуковых сигналов в виде векторов признаков. Он является одним из наиболее популярных методов в анализе и распознавании речи.

2. Chroma (хрома) в контексте аудио-обработки является вектором, который представляет распределение энергии звукового сигнала по музыкальным тональностям (нотам). Chroma-вектор позволяет оценить наличие и относительную интенсивность различных музыкальных тональностей в аудиозаписи [2].

3. Мел-спектрограмма – это представление звукового сигнала, которое позволяет увидеть, какая часть частот присутствует в звуке в разные моменты времени. Она показывает, как энергия звука распределена по различным частотам в зависимости от времени [3].

4. RMS (Root Mean Square Среднеквадратичный корень) – это статистическая мера, используемая для измерения среднеквадратического значения амплитуды звукового сигнала. Она представляет собой квадратный корень из среднего значения квадратов амплитудных значений во временной области сигнала.

ИЗВЛЕЧЕНИЕ ПРИЗНАКОВ

Извлечение аудио признаков происходит следующим образом, берётся аудиофайл для дальнейшего анализа, затем происходит сегментация аудио файла, это сделано для того, чтобы получить одинаковое количество признаков из аудиофайлов разной продолжительности, после этого происходит получение аудио признаков, дальше происходит их нормализация, затем аудио признаки объединяются и в итоге добавляются в общий массив признаков других аудиофайлов.

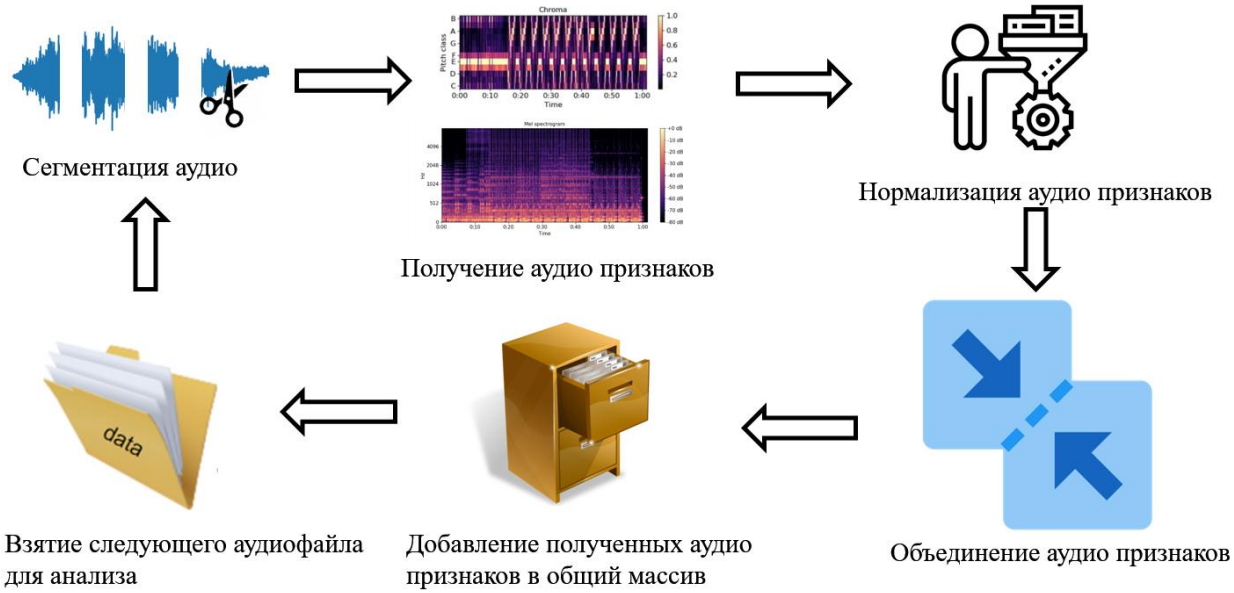


Рис. 3. Общая схема извлечения признаков из аудиоданных
Fig. 3. General scheme of feature extraction from audio data

Для эффективного распознавания эмоций в речевых данных важно извлечь соответствующие признаки из аудиофайлов. Однако, аудиофайлы могут иметь разную длительность, что создает проблемы при получении универсальных и сравнимых признаков для анализа. В целях решения этой проблемы предлагается следующий подход:

1. Округление длительности – В первую очередь, каждая длительность аудиофайла округляется до ближайшей целой секунды.

2. Разбиение на сегменты – Затем каждый аудиофайл разбивается на 20 равных сегментов. Это позволяет получить более детальное представление аудио-сигнала и учитывать его изменения во времени. Каждый сегмент имеет одинаковую длительность, что обеспечивает сравнимость признаков между различными аудиофайлами.

3. Извлечение признаков – После разбиения на сегменты из каждого сегмента извлекаются признаки, извлекаются они библиотекой librosa. Эти признаки будут включать такие характеристики, как MFCC, хромаграммы, мел-спектрограммы и RMS. Извлечение признаков позволяет представить аудиофайлы в виде числовых векторов, содержащих информацию о специфических акустических характеристиках каждого сегмента.

Если аудиофайлы имеют длительность от 3 до 5 секунд, то деление на 20 частей имеет смысл, чтобы получить более мелкие сегменты и извлечь признаки из них.

Каждый аудиофайл разбивается на 20 равных сегментов. При этом каждый сегмент будет иметь длительность, соответствующую 1/20 от общей длительности аудиофайла. Таким образом, каждый сегмент будет иметь продолжительность от 0.15 до 0.25 секунды. Важно отметить, что получаемые массивы данных из RMS строго зависят от длинны аудиозаписи, поэтому массивы имеют фиксированный размер, если данных недостаточно, то они дополняются нулями, иначе обрезаются.

Нормализация данных в аудио признаках выполняется с целью привести их к общему масштабу и диапазону значений. Опытным путём было выявлено, что метод нормализации L2-нормы, которая основывается на вычислении евклидовой нормы векторов признаков, имеет наивысшую эффективность. Этот подход позволяет привести значения признаков к единичной длине, что упрощает их сравнение и улучшает стабильность процесса обучения.

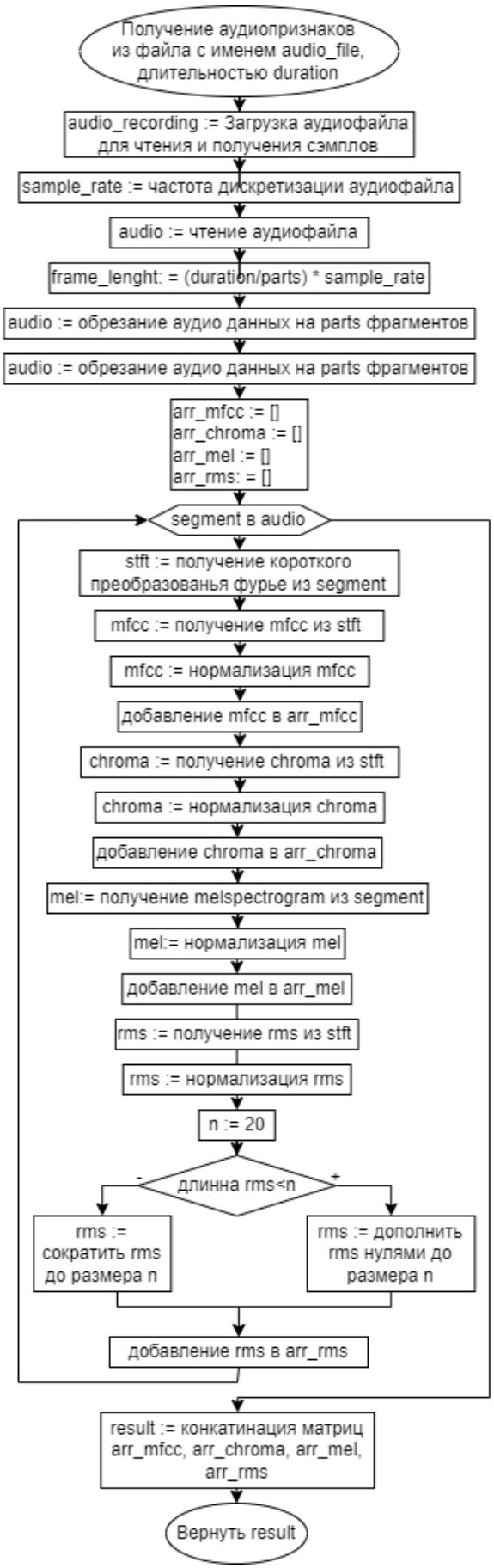


Рис. 4. Получение аудио признаков из файла
Fig. 4. Getting audio features from a file

МОДЕЛЬ НЕЙРОННОЙ СЕТИ

BatchNormalization: Этот слой нормализует входные данные путем стандартизации их по среднему значению и дисперсии. Это помогает улучшить стабильность и скорость обучения модели.

Bidirectional LSTM: LSTM (Long Short-Term Memory сети долгой краткосрочной памяти) – это рекуррентный слой, который способен сохранять информацию о предыдущих состояниях. Bidirectional LSTM обрабатывает последовательности в обоих направлениях, что позволяет модели улавливать и использовать информацию из прошлого и будущего контекста [4]. Он эффективно моделирует зависимости во временных данных, таких как звуковые сигналы речи.

LSTM: Еще один слой LSTM используется для дальнейшего извлечения временных зависимостей из последовательностей данных. Это помогает модели уловить более сложные и долгосрочные зависимости в аудиоданных [5].

Conv1D: Слой одномерной свертки используется для обнаружения локальных шаблонов и особенностей в аудиоданных.

Dropout: Слой Dropout используется для предотвращения переобучения модели путем случайного обнуления некоторых элементов выходных данных во время обучения. Это помогает улучшить обобщающую способность модели и предотвратить переобучение.

Flatten: Слой Flatten преобразует многомерный выход из предыдущего слоя в одномерный вектор. Он подготавливает данные для передачи в полносвязные слои модели.

Dense: Полносвязные слои применяются для классификации. Они содержат нейроны, соединенные со всеми нейронами предыдущего слоя.

Совмещение Conv1D и LSTM позволяет моделировать как глобальные, так и локальные зависимости в данных. LSTM слои помогают захватывать долгосрочные зависимости и общий контекст, в то время как Conv1D слои улавливают локальные паттерны и временные особенности данных. Комбинирование этих двух типов слоев дает возможность модели извлекать более информативные признаки из входных данных и лучше понимать структуру временных рядов.

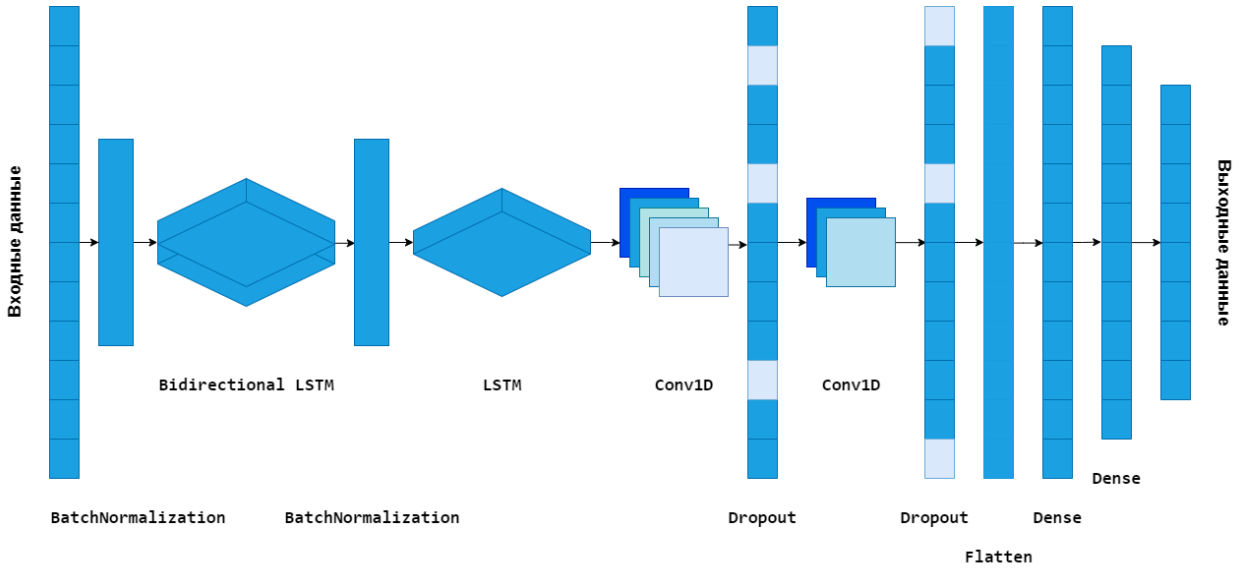


Рис. 5. Структура модели нейронной сети
Fig. 5. Structure of the neural network model

Первый слой BatchNormalization имеет значение momentum=0.91 означает, что модель будет сохранять 91% информации о статистике предыдущего пакета и использовать его для нормализации текущего пакета данных. Это позволяет модели быстрее адаптироваться к изменениям в данных и повышает устойчивость обучения. Более высокое значение momentum может привести к более

стабильной и более медленной адаптации, в то время как более низкое значение может привести к более быстрой, но менее стабильной адаптации.

Второй слой Bidirectional LSTM имеет 240 нейронов и функцию активации 'tanh', все слои имеют данную функцию активации, так как она показала наивысшую эффективность.

Третий слой BatchNormalization имеет значение momentum=0.95.

Четвертый слой LSTM имеет 140 нейронов.

Пятый слой Conv1D имеет 36 нейронов с ядром свёртки 5 и L2-регулиризация со значением 0,0001(kernel_regularizer=regularizers.l2(0.0001)).

Шестой слой Dropout имеет значение 0.2.

Седьмой слой Conv1D имеет 20 нейронов с ядром свёртки 3.

Восьмой слой Dropout имеет значение 0.2.

Десятый слой Dense имеет 140 нейронов.

Одиннадцатый слой Dense имеет 40 нейронов.

Двенадцатый и он же выходной слой имеет 8 нейронов, т.е. равную количество эмоций, а так же функцию активации ‘softmax’.

В данной модели используется функция потерь categorical_crossentropy, так как решается задача многоклассовой классификации.

Оптимизатор adam выбран для эффективного обновления весов модели в процессе обучения, обеспечивая быструю сходимость и хорошие результаты.

РЕЗУЛЬТАТЫ РАБОТЫ

Исходные данные были поделены на обучающие и тестовые, в тестовые данные входят 20% от изначальной.

Рис. 6. Аудиоданные из датасета RAVDESS
Fig. 6. Audio data from the RAVDESS dataset

Рис. 7. Обучающие аудиоданные из датасета RAVDESS
Fig. 7. Training audio data from the RAVDESS dataset

Рис. 8. Тестовые аудиоданные из датасета RAVDESS
Fig. 8. Test audio data from the RAVDESS dataset

Каждый класс имеет различное количество аудио файлов как в обучающем, так и в тестовом наборе данных. Это позволяет модели получить разнообразный набор примеров и обучиться на различных эмоциональных фоновых состояниях.

В результате обучения модели на разнообразном наборе данных, получили хорошие результаты точности и потерь.

Анализируя графики, можно заметить, что уже на 20-й эпохе обучения точность модели на обучающих данных достигла значения 100%, тогда как на тестовых данных она составила 78%. Это свидетельствует о том, что модель успешно обучается и справляется с классификацией эмоций на обучающих и незнакомых данных.

Кроме того, потери на обучающих данных приближаются к нулю, что указывает на эффективность обучения модели на тренировочном наборе данных. В то же время, потери на тестовых данных составляют примерно 0.77, что говорит о том, что модель делает некоторые ошибки в предсказаниях на незнакомых данных, но все равно достигает хороших результатов.

Однако, самые лучше показатели были достигнуты на 67-й эпохе. Точность модели увеличилась до 83%, что является высоким значением для данной задачи. При этом потери также немного увеличились до 0.81, но оставались на приемлемом уровне.

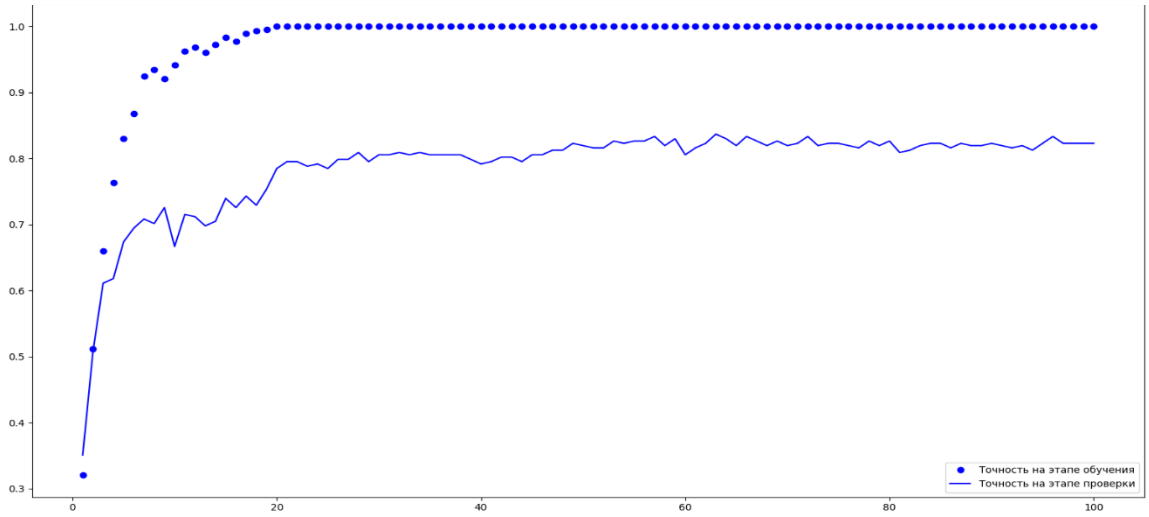


Рис. 9. Точность на этапах обучения и проверки
Fig. 9. Accuracy at the stages of training and verification

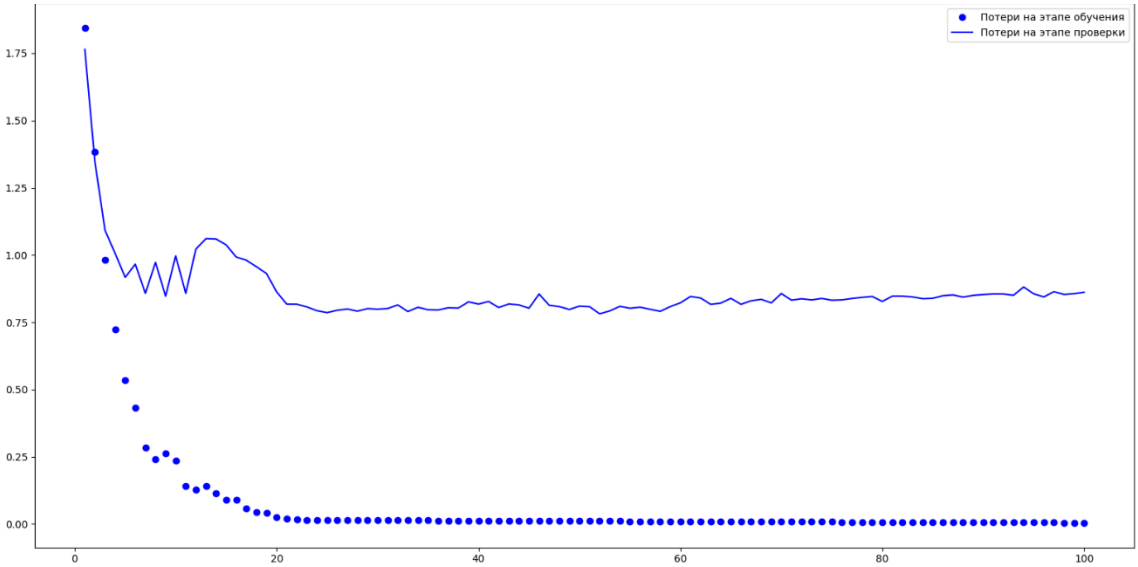


Рис. 10. Потери на этапах обучения и проверки
Fig. 10. Losses at the stages of training and verification

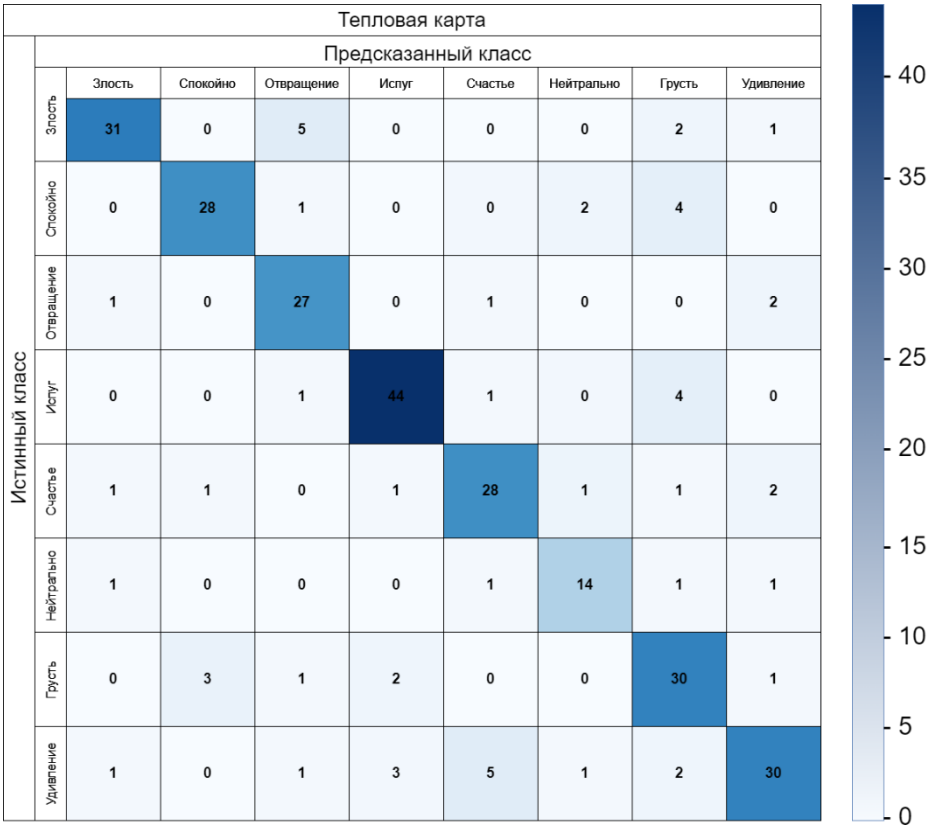


Рис. 11. Тепловая карта результатов обученной модели
Fig. 11. Heat map of the results of the trained model

ЗАКЛЮЧЕНИЕ

С использованием библиотеки Keras и соответствующих алгоритмов была создана модель, которая показала высокую точность и низкие потери. Графики точности и потерь подтверждают успешность обучения модели, достигнув наиболее оптимальных показателей на 67 эпохе с точностью в 83% и потерями в 0.81.

Для дальнейшего улучшения модели классификации эмоций на основе аудиоданных можно увеличить размер обучающего набора данных, чтобы обеспечить большую обобщающую способность модели. Несмотря на достигнутые высокие показатели точности и потерь, данная модель имеет потенциал для дальнейшей оптимизации и улучшения. Дальнейшие исследования и разработки в этой области могут привести к созданию более точных и эффективных моделей классификации эмоций на основе аудиоданных.

Список литературы

1. Шолле Ф. Глубокое обучение на Python. 2-е межд. издание. – СПб.: Питер, 2023. – 576 с. – ISBN 978-5-4461-1909-7.

2. Han K., Lee K., Kim H.G. Music emotion recognition using chroma feature-based probabilistic neural network. Multimedia Tools and Applications. – 2017. – Том №76, Выпуск №3. – С. 3691-3710.

3. Getting to Know the Mel-Spectrogram. [Электронный ресурс] – Электрон, дан., 2019. – URL: https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0

4. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Nature. – 2005. – Том №18, Выпуск №5-6. – С. 602-610.

5. Understanding LSTM Networks. [Электронный ресурс] – Электрон, дан., 2015. – URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/

References

1. Chollet F. Deep Learning with Python. 2nd interd. edition. – SPb.: St. Petersburg: St. Petersburg. – 576 p. – ISBN 978-5-4461-1909-7.
2. Han K., Lee K., Kim H.G. Music emotion recognition using chroma feature-based probabilistic neural network. Multimedia Tools and Applications. – 2017. – V. №76, Issue №3. – P. 3691-3710.
3. Getting to Know the Mel-Spectrogram. [Electronic resource] – Electronic data, 2019. – URL: https://towardsdatascience.com/getting-to-know-the-mel-spectrogram-31bca3e2d9d0.
4. Graves A., Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Nature. – 2005. – V. №18, Issue №5-6. – P. 602-610.
5. Understanding LSTM Networks. [Electronic resource] – Electronic data, 2015. – URL: http://colah.github.io/posts/2015-08-Understanding-LSTMs/.

Жихарев Александр Геннадиевич, доктор технических наук, доцент, доцент кафедры программного обеспечения вычислительной техники и автоматизированных систем
Черных Владимир Сергеевич, студент 1 курса Магистратуры, направления «Прикладная информатика», кафедра прикладной информатики и информационных технологий, институт инженерных и цифровых технологий

Zhikharev Alexander Gennadievich, Doctor of Technical Sciences, Associate Professor, Associate Professor of the Department of Software for Computer Engineering and Automated Systems
Chernykh Vladimir Sergeevich, 1st year Master's student, Applied Informatics, Department of Applied Informatics and Information Technologies, Institute of Engineering and Digital Technologies

UDC 004.62 **DOI: 10.18413/2518-1092-2022-8-3-0-6**

Sokolova E.V.

INTELLECTUAL ANALYSIS OF SCIENTOMETRIC INDICATORS OF RUSSIAN COMPUTER SCIENCE JOURNALS

Federal State Budgetary Educational Institution of Higher Education "Ural State University of Economics"
8 Marta str./Narodnaya Volya, 62/45, Yekaterinburg, 620144

e-mail: sokolova_ev@usue.ru

Abstract
The article discusses issues related to forecasting scientometric indicators and clustering Russian scientific journals in the field of computer science with the aim of evaluating the current state of science and the progress of research in the country, as well as the state of the competitive market for scientific publications. The study is based on the intelligent analysis of results using forecasting models such as ARIMA, ETS, and nnetar, developed in the RStudio environment. The research findings can be valuable for representatives of scientific journal publishers in the field of the State Classifier of Scientific and Technical Information (GRNTI) "Informatics. Information Technologies" when forming development and promotion strategies in a dynamically changing scientific world. The forecasted results indicate that most industry representatives will maintain their current indicators or change insignificantly, but there is potential for the emergence of new journals. It has also been revealed that journals not included in the Higher Attestation Commission (HAC) list often show slower growth, which is directly related to changes in the calculation of the Science Index indicator since the spring of 2023. However, inclusion is not a guarantee of growth at the moment. Significant growth is observed only in journals that are either classified as category 1 by the HAC or classified as category 2 by the HAC and are included in the Russian Science Citation Index (RSCI) Core simultaneously.
Keywords: scientometric indicators; publication activity; scientific journals; information technology; forecasting; intellectual analysis
For citation: Sokolova E.V. Intellectual analysis of scientometric indicators of russian computer science journals // Research result. Information technologies. – T.8, №3, 2023. – P. 45-55. DOI: 10.18413/2518-1092-2022-8-3-0-6

Соколова Е.В.

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ НАУКОМЕТРИЧЕСКИХ ПОКАЗАТЕЛЕЙ РОССИЙСКИХ НАУЧНЫХ ЖУРНАЛОВ ПО ИНФОРМАТИКЕ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Уральский государственный экономический университет»
ул. 8 Марта/Народной Воли, 62/45, г. Екатеринбург, 620144

e-mail: sokolova_ev@usue.ru

Аннотация
В статье рассматриваются вопросы прогнозирования наукометрических показателей и кластеризации российских научных журналов по информатике с целью оценки текущего развития науки и прогресса научных исследований по тематике в стране, а также состояния конкурентного рынка научных изданий. Исследование базируется на интеллектуальном анализе результатов использовании моделей прогнозирования ARIMA, ETS и nnetar, разработанных в среде RStudio. Результаты исследования могут быть полезны для представителей издательств научных журналов рубрики ГРНТИ «Информатика. Информационные технологии» при формировании стратегии развития и продвижения в динамично меняющемся научном мире. Полученные результаты прогноза свидетельствуют о том, что большинство представителей отрасли сохранят свои текущие показатели или изменится незначительно, однако, существует потенциал для вхождения новых журналов.

Также было выяснено, что зачастую меньший рост показывают журналы, не входящие в Ядро РИНЦ, что напрямую связано с изменением расчета показателя Science Index от весны 2023 г., однако, вхождение также не является гарантией роста на текущий момент. Тогда как заметный рост отмечен лишь у журналов, которые или причислены к 1 категории ВАК, или причислены ко 2 категории ВАК и входят в Ядро РИНЦ одновременно.
Ключевые слова: наукометрические показатели; публикационная активность; научные журналы; информационные технологии; прогнозирование; интеллектуальный анализ
Для цитирования: Соколова Е.В. Интеллектуальный анализ наукометрических показателей российских научных журналов по информатике // Научный результат. Информационные технологии. – Т.8, №3, 2023. – С. 45-55. DOI: 10.18413/2518-1092-2022-8-3-0-6

INTRODUCTION

The study and analysis of peer-reviewed scientific journals are essential components of assessing the development of a scientific field within a specific country. Peer-reviewed publications serve as the key source for disseminating information within the scientific community and contribute to the spread of accurate and reliable research findings [18]. They provide a platform for researchers to share their results and ideas while also receiving rigorous evaluation and feedback from colleagues [2]. The peer-review process ensures the quality and reliability of published research and helps maintain the integrity of the scientific community.

In the analysis of the publication activity of scientific journals, various scientometric metrics and indicators are taken into account, measuring the conditional influence, visibility, and influence of journals [20]. The analysis of these indicators provides researchers with a comprehensive understanding of the state of research in a particular field and aids in identifying emerging trends and key knowledge generation centers.

For publishers aiming to promote their journals and gain a competitive advantage in the scientific community, it is also crucial to assess the scientometric indicators of competing journals. Systematic analysis enables publishers to identify the strengths and weaknesses of competing journals in terms of citation, publication frequency, international collaboration, and other significant factors. This, in turn, helps improve existing editorial strategies, identify potential collaboration opportunities, and develop targeted marketing campaigns to attract authors, reviewers, and readers.

Furthermore, the analysis of the effectiveness of journals operating in a specific scientific field within one country allows for comparative analysis and comparison of national and international indicators [21]. Comparing scientometric indicators of journals within a country with those of other countries enables the evaluation of a country's position and contribution to the global research landscape. It also assesses the competitiveness of research institutions and informs resource allocation decisions, fostering increased collaboration and partnership to strengthen scientific production within the country.

For a variety of reasons and the relevance of the field, the conducted research is primarily based on the thematic direction of informatics and information technology. Specifically, it focuses on scientific journals belonging to this subject area. However, the methods used and the model being developed are not tied to any specific direction. The GRNTI category "Informatics. Information Technologies" on the main Russian scientific platform eLIBRARY.RU is represented by 337 journals, which is fewer than 33 other GRNTI categories (excluding multidisciplinary ones) out of 68 (as of May 1, 2023) [16]. Among the journals in the field of Informatics, only 98 scientific journals are currently publishing, accepting publications in Russian, and indexed in the Russian Science Citation Index (RSCI), of which only 55 are included in the list of recommended journals HAC as of May 2023 [4]. In the author's opinion, this is insufficient for such a significant field of science and requires further study.

MATERIALS AND METHODS

Intellectual models can become a preferred approach for data analysis due to their exceptional capabilities in uncovering complex patterns and extracting significant insights from intricate datasets. L. Leydesdorff developed his own model for analyzing time series data as early as 1990, which was applicable

to data on national productivity across countries in the European Economic Community (EEC) and the United States [9]. A. Fernández-Cano and M. Torral evaluated the application of various analytical methods for classifying time series data of scientific growth, ranging from visual analysis of deterministic graphs to curve fitting using exponential smoothing and autoregressive integrated moving average models. Tseng Y. H. and co-authors used multiple trend indices to identify emerging research topics [17].

One of the widely used intellectual models that can be employed in scientometric analysis is the ARIMA (AutoRegressive Integrated Moving Average) model [3; 13].

ARIMA, a time series forecasting model, helps predict future journal indicators by analyzing historical data. It examines patterns and trends in metrics like citations, publication frequency, and impact factors, offering insights into a journal's growth [8; 11]. Vasileiadou E. utilized ARIMA modeling to study the dynamics and types of research collaboration [19].

The nnetar model, based on neural networks for autoregression, applies deep learning concepts and algorithms to model complex dependencies between historical series values and their future values [8]. The neural network adapts to the characteristics of the time series and can capture nonlinear relationships and the dynamics of the series.

The ETS model, or exponential smoothing of time series, is a classic statistical forecasting method widely used in time series analysis [11]. This method is based on the idea of smoothing and assigning different weights to historical values of the series to obtain a forecast based on exponentially weighted averages. ETS takes into account trends, seasonality, and random factors in time series, allowing it to model various features of their behavior.

The selection of these three methods—auto.arima, nnetar, and ETS—is driven by their combined advantages and the ability to complement each other. The automatic ARIMA model (auto.arima) offers convenience and ease of use, as it automatically selects the optimal model parameters based on information criteria. The nnetar model, on the other hand, provides flexibility in modeling nonlinear dependencies, which can be important for predicting journal rankings. The ETS model complements the previous two models by accounting for trends and seasonality, enabling a more accurate capture of various characteristics of time series data. The combined use of these three models allows for the consideration of different aspects and characteristics of journal ranking data, enhancing the accuracy and reliability of forecasts, as well as their comparability.

However, it is important to note that the successful application of these methods requires careful consideration of the data, model assumptions, and appropriate validation methods to ensure the reliability and stability of the estimates.

Scientometric indicators can vary depending on the database in which they are calculated [6]. The scientific electronic library eLIBRARY.RU is the largest scientific database in Russia, and its indexing is a primary requirement for the existence of a scientific journal. eLIBRARY.RU accumulates over 56,000 scientific journals, of which more than 5,600 journals are indexed in the RSCI. Users of the platform can analyze the publication activity of each journal, including changes in various indicators over the years. However, the mentioned indicators are mostly interrelated and correlated based on calculation methods. Based on the existing policy for regulating scientific activities, the primary indicator can be considered the journal's ranking in the Science Index. The integral indicator of a journal in the Science Index system takes into account most of the mentioned indicators and indexes, and therefore, despite the theoretical shortcomings of its calculation method, it will be prioritized in assessment and forecasting [12].

The forecasting involves predicting the journal ranking indicator in the Science Index for each individual journal based on time series. The input data include a column with dates and subsequent columns representing ranking indicators for only 30 journals that have consistently published from 2012 to 2021 (Table 1). In this analysis, indicators were forecasted for not all journals but only for established publications that have been consistently published for the last 9-10 years. This limitation is due to the fact that these journals are the main leaders in the market and exhibit stable trends that can already be used for forecasting. Analyzing the indicators of such journals will help understand and forecast future trends and changes in the industry.

Таблица 1

Структура и пример части данных по изменению показателя журналов в рейтинге Science Index для прогнозирования

Table 1

Structure and example of the data part on the change in the index of journals in the Science Index ranking for prediction

DateJ	J75402	J7310	J27958	J28165	.	J74841
01.01.2012	2,087	2,346	4,045	1,518	.	1,896
01.01.2013	2,397	2,787	4,352	1,942	.	2,342
01.01.2014	2,66	2,743	5,122	2,41		2,427
01.01.2015	2,986	2,964	4,783	2,924		2,604
01.01.2016	2,804	2,956	5,27	2,883		3,036
01.01.2017	2,92	3,183	5,307	3,149		2,901
01.01.2018	3,537	3,454	5,318	3,171		3,245
01.01.2019	4,006	3,557	6,022	3,176		3,251
01.01.2020	4,027	3,882	5,649	3,063		3,713
01.01.2021	3,694	4,239	6,365	3,112		3,441

All three models use the journal's rating history as the primary predictor for forecasting future values, specifically the indicator for the year 2022. They take into account the temporal dependency and changes in the rating over time. The output of each model consists of forecasted rating values for the year 2022 for each individual journal based on historical data.

It is important to note that for the purpose of clustering and analysis, predictors were selected based on their influence on an author's decision when choosing a journal to publish their research results. In other words, they were chosen from the author's perspective and interests. Authors typically choose journals based on factors such as indexing, a limited number of indicators (if there are requirements from their organization or personal preferences), and the perceived "complexity" of publishing in a journal, which is subjectively assessed by each author.

Based on this, and considering the high need for journals to attract new authors, the predictors considered during clustering were chosen to reflect relevant factors that distinguish a journal from the author's perspective. Thus, the clustering of the 59 scientific journals in the sample was based on a combination of several indicators (Table 2):

1. "SI" (journal ranking indicator in the Science Index).
2. "Articles" (the number of articles published by the journal in the last year).
3. "HAC" (the journal's inclusion in the list of journals recommended by the Higher Attestation Commission (HAC), with category information, where the highest category is 3) [19].
4. "CORE" (the journal's inclusion in the Russian Science Citation Index (RSCI) core).
5. "PERIOD" (average number of days for an article to be published after being received by the editorial office).
6. "REJECTION" (the proportion of publications rejected by the journal's editorial board).

These predictors were chosen to provide a comprehensive view of the factors that matter most to authors when selecting a journal for publication and to enable effective clustering and analysis based on these criteria.

Таблица 2

Структура и пример части данных по журналам для кластеризации

Table 2

Structure and example of log data part for clustering

Journal	SI	Articles	HAC	CORE	PERIOD	REJECTION
J10273	3,376	59	2	0	90	0,2
J10570	6,938	38	3	1	180	0,5
J11926	3,94	30	3	0	30	0
…						
J9678	6,42	47	0	1	90	0,05

Attracting new authors is a crucial component of the vitality and success of a scientific journal. This is driven by the desire to draw in promising new authors, as their research and publications can significantly enhance the prestige and reputation of the journal. Evaluating competitors based on their attractiveness to potential authors helps journals understand their position in the scientific publishing market and identify unique advantages they can offer. It also enables the development of strategies for attracting and retaining talented authors, contributing to the growth and prosperity of the journal.

The selected forecasting and clustering models are implemented in the RStudio environment using the R programming language [14-15]. R provides the necessary tools and libraries for data analysis at all stages, from data processing to the visualization of results [10]. When working with models for analyzing scientometric indicators of scientific journals, several libraries were used, including "Tidyverse," "nnet," "forecast," "cluster," "randomForest," and others.

Forecasting the Science Index journal rating includes three similar blocks: ARIMA, nnetar, and ETS.

For example, ARIMA: The data is transformed into time series and analyzed using auto.arima() from the "forecast" package. A model is built for each object, and forecasts are made one step ahead. The results are stored in the variable predicted_values_autoarima.

Forecasts are generated using the print() function. A combined dataframe with the results of all three methods is created using cbind(). Then, a column with the object number is added, and the "model" and "forecast" columns are removed from combined_results_selected. The data is transformed into a long format using pivot_longer() from "tidyr."

Next, a comparison chart of forecasts is created using ggplot(). Data is sourced from combined_results_long, where the "Object" and "Forecast" columns define the X and Y axes. The chart uses the "bar" geometry to display forecasted values by methods.

Additionally, a "boxplot" chart is created to compare the accuracy of forecasting methods. The geom_boxplot() function is used to display the spread of values by methods.

As a result, the model provides users with forecasted Science Index journal rating values for 2022 for each method (ARIMA, nnetar, ETS) for each data object (journal).

The model also performs a cluster analysis of journal data using the random forest method after loading the necessary packages and reading the data. Rows with missing values (NA) are removed using na.omit(), while object identifiers are stored in the variable object_ids.

Then, the column of object identifiers is removed from the data. A min_max_normalize function is created to normalize the data using the min-max method with weights considered.

In the next block, data clustering is performed using the Random Forest algorithm. A Random Forest model (rf_model) is created, and clustering is carried out on normalized data, excluding the last column. Then, a proximity matrix (prox_matrix) is extracted from the Random Forest model, reflecting the similarity between objects.

Next, the optimal number of clusters is determined using the elbow method [1]. For each value of k from 2 to 10, the sum of squared distances within the cluster (WSS) is calculated. Then, the derivative and second derivative of WSS are computed. The optimal k is determined as the first value of k at which the second derivative of WSS becomes negative.

Data clustering is performed using the optimal number of clusters (optimal_k) with the help of the cutree function, and the result is stored in the variable cluster_assignments. The clusters are added to the normalized data as a new column called "Cluster."

Then, a data frame wss_df is created containing WSS values for k from 2 to 10. A wss_plot is generated using ggplot2, showing the relationship between WSS and the number of clusters (k) and highlighting the optimal k with a red dot.

Finally, box-and-whisker plots are created for the values of the silhouette width in each cluster, displaying the distribution of values for each cluster.

As a result of executing this model, the user receives cluster labels, the optimal number of clusters, and corresponding plots.

RESULTS

The comparison of forecasting results indicates the relative similarity among all the methods used. Often, the difference between values is around 0.2-0.3 units, occasionally reaching 1 unit for journals with initially high values in the Science Index journal ranking and a sharp positive change in recent years (see Figures 1-2).

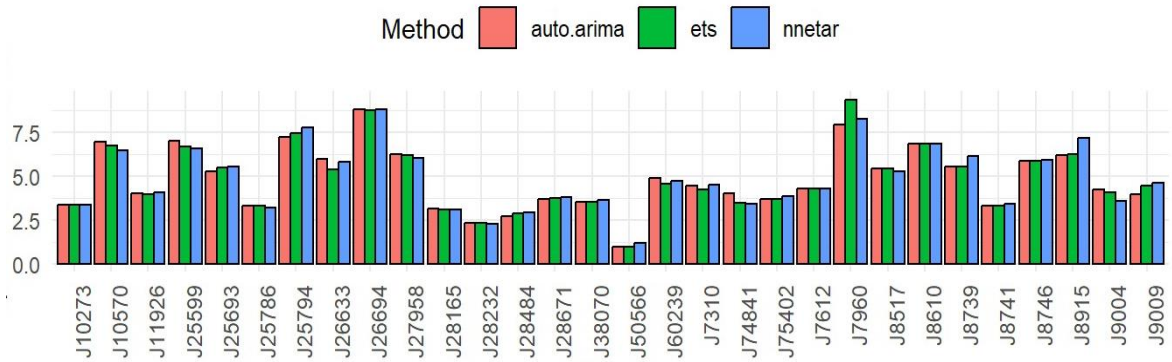


Рис. 1. Сравнение результатов прогнозирования интеллектуальной модели (Прогнозируемое значение/Объект)

Fig. 1. Comparison of intelligent model prediction results (Predicted value/Object)

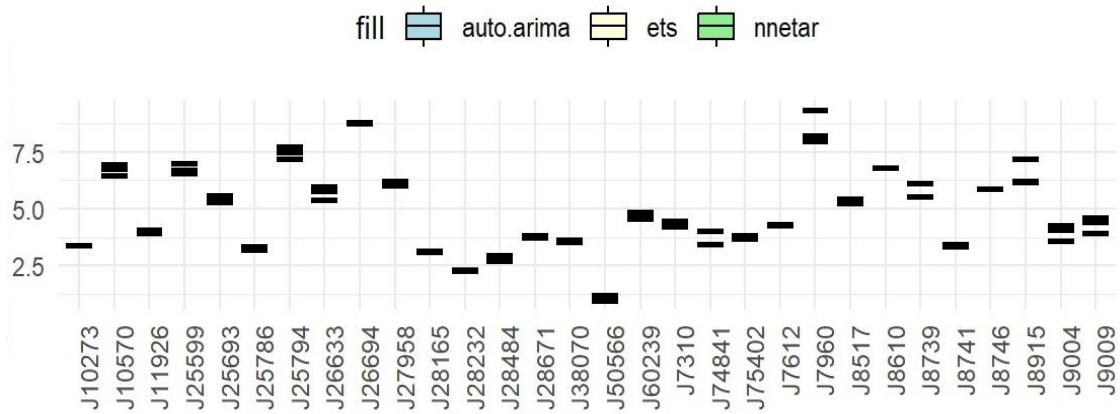


Fig. 2. Comparison of forecast accuracy of the intelligent model (Predicted value/Object)

Рис. 2. Сравнение точности прогнозов интеллектуальной модели (Прогнозируемое значение/Объект)

Furthermore, based on the obtained Mean Squared Error (MSE) values, it can be concluded that the auto.arima model demonstrates the best performance among the forecasting models. Comparing the MSE values, it can be noted that the MSE for auto.arima is 7.55, whereas for the nnetar model, this value is 7.62, and for the ETS model, it is 8.17 [7].

Based on these results, it can be assumed that the auto.arima model provides more accurate forecasts with values closer to the actual data (which are currently unavailable). This may be attributed to the advantages of the automatic ARIMA model selection method, which considers various combinations of model parameters and selects the optimal model most suitable for the time series. However, it is essential to note that without actual data, it is not possible to fully assess the quality of the models and claim that the selected model is optimal or correct.

To determine the optimal number of clusters, the Elbow method (WSS) was used, which is based on the sum of squares of distances within clusters. The goal was to choose a number of clusters after which an increase in the number of clusters no longer significantly reduces the sum of squares of distances.

Based on the obtained values, it can be observed that the reduction in WSS slows down after choosing 4 or 5 clusters (see Figure 3). After this point, the decrease in WSS becomes less substantial. Therefore, in this case, the choice was made for 4 or 5 clusters as the optimal number, considering the data structure and volume. Ultimately, clustering was performed with 4 clusters.

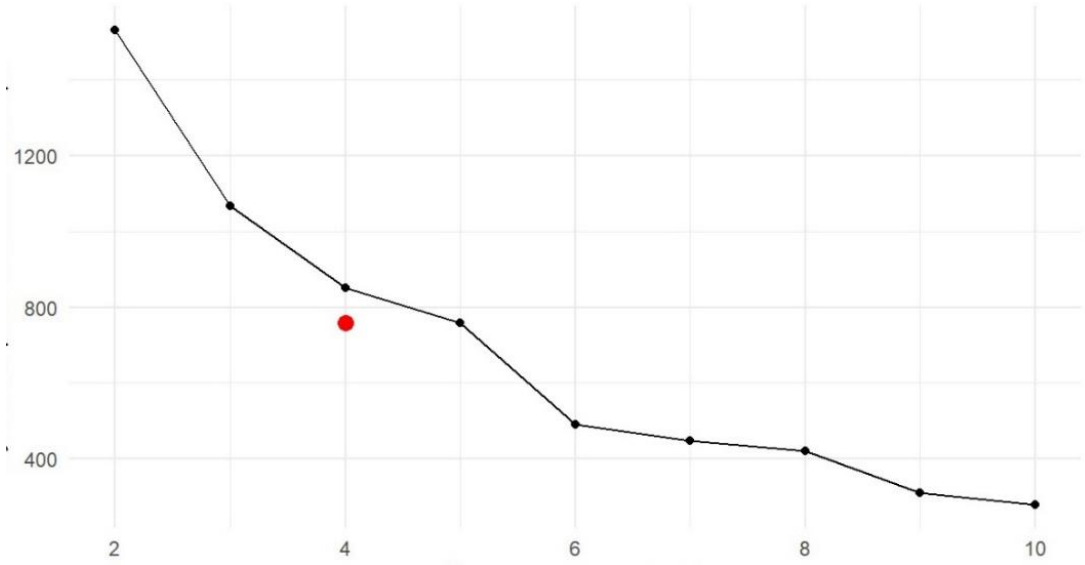


Fig. 3. Variation of the sum of squares of distances within clusters with respect to the number of clusters (Sum of squared distances within the cluster/Number of clusters)

Рис. 3. Изменение суммы квадратов расстояния внутри кластеров относительно числа кластеров (Сумма квадратов расстояний внутри кластера/Количество кластеров)

The evaluation of the quality of the intellectual models for analyzing journal bibliometrics indicates that both the forecasting models and the clustering model are of high quality and can be successfully applied in future practical tasks.

The results obtained from the forecast based on the analysis of a journal's Science Index rating suggest that the majority of journals among the well-established ones in the field, provided they maintain their existing editorial policies, will either maintain their current indicators or experience minor changes. However, six journals have been identified with predicted growth and another six journals with expected significant decreases in their indicators.

It is worth noting that for these six journals with predicted growth, the forecast indicates an increase in their indicators at the level of 0.35, which is less significant compared to existing results but substantial when compared to the industry average. On the other hand, for six journals, a significant decrease in their indicator is forecasted in the range from -0.2 to 1 (see Figure 4). This may be attributed to changes in the journal's research focus, competition with other publications, or issues affecting the quality of publications. Such declines can have a substantial impact on the status and influence of these journals within the scientific community.

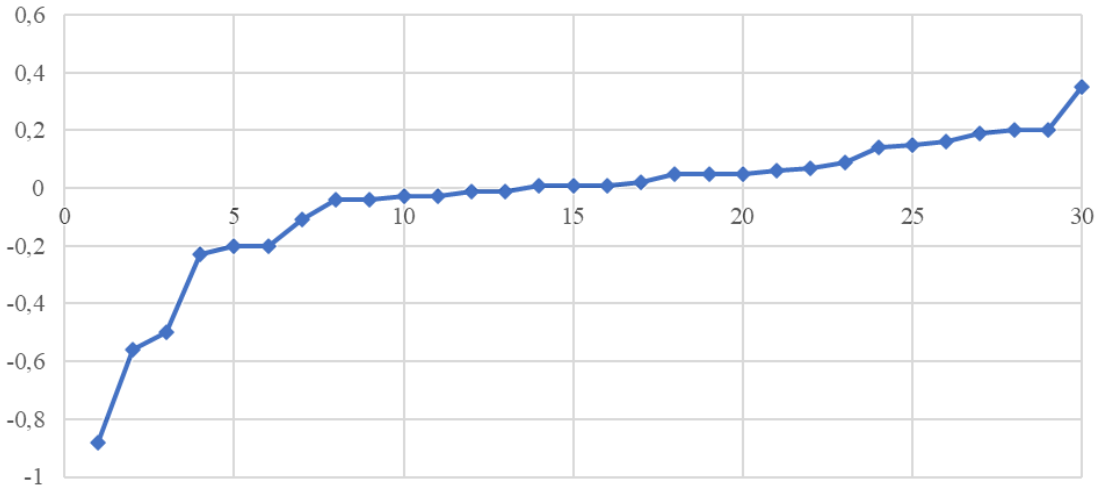


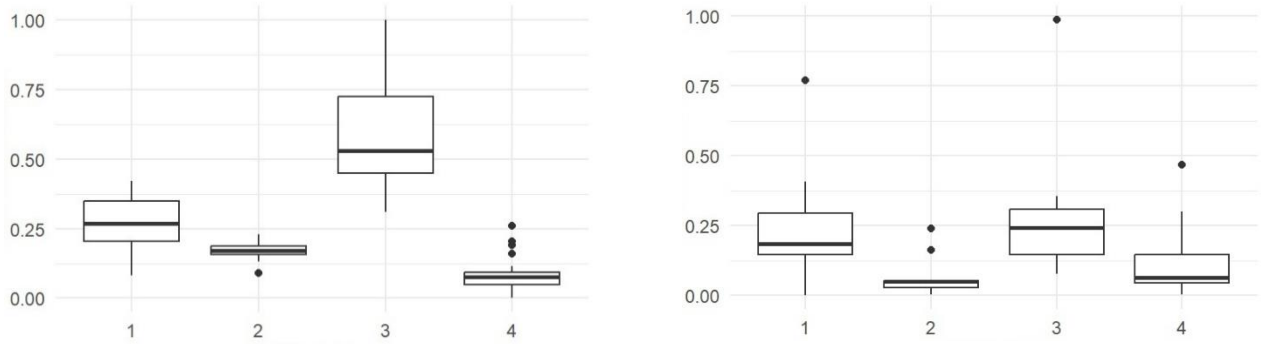
Fig. 4. Projected change in the journal's index in the Science Index ranking for 2022 relative to the current year 2021

Рис. 4. Прогнозируемое изменение показателя журнала в рейтинге Science Index на 2022 г. относительно текущего 2021 г.

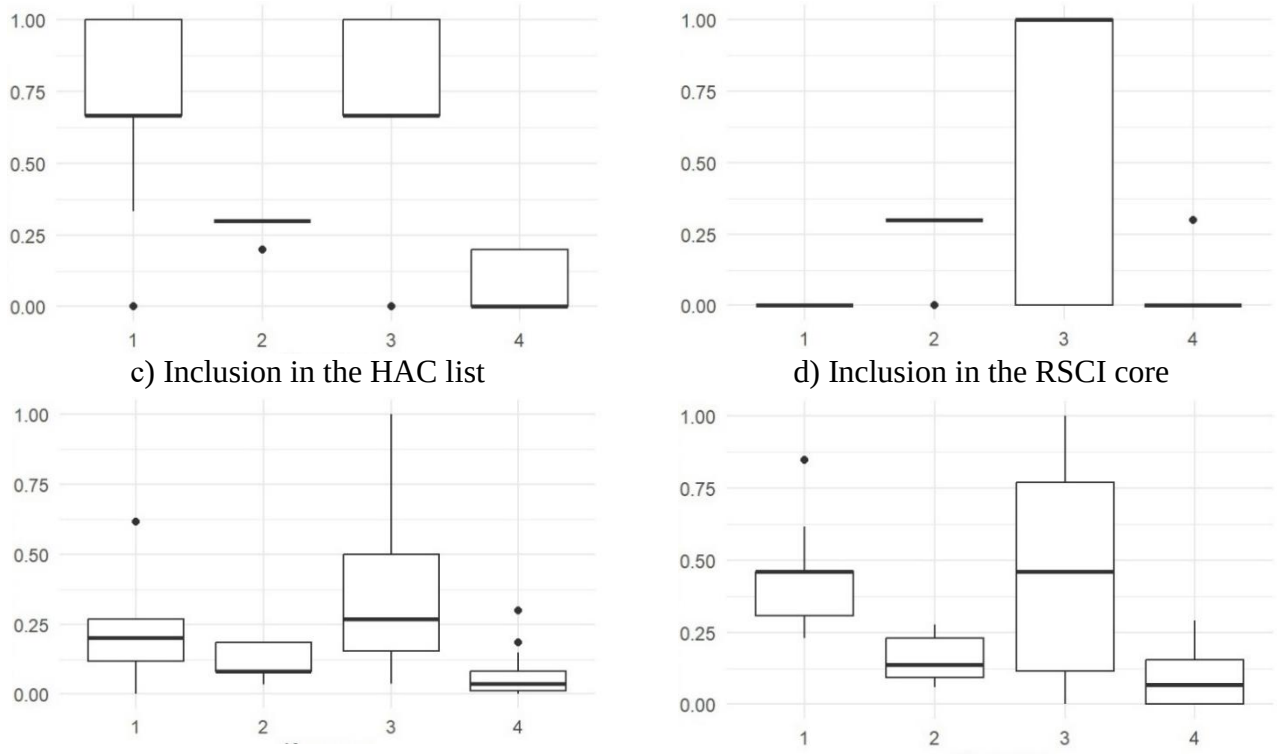
It was also found that journals not included in the RSCI tend to show slower growth. This is directly related to the change in the calculation of the Science Index rating from the spring of 2023. However, inclusion in RSCI is not necessarily a guarantee of growth at the moment. Noticeable growth has been observed primarily in journals that are either classified as Category 1 in the HAC or classified as Category 2 in HAC and are also included in RSCI simultaneously. It can be assumed that such dependencies will become less relevant in the near future due to the recent change in the calculation of the Science Index rating, to which the scientific community has not yet fully adapted.

During clustering, the journals were divided into four groups or clusters, with numbering unrelated to their means. The largest cluster, Cluster 4, included 25 journals with the lowest average Science Index rating within the cluster. These journals rarely even belong to HAC Category 3 and RSCI. Moving down in terms of the number of journals included, Cluster 1 contained 14 journals. It also included journals with unremarkable Science Index ratings, which are more likely to belong to HAC Categories 1-2 but are not part of the RSCI journal list. Both Cluster 4 and Cluster 1 journals share a relatively slow publication pace, but Cluster 1 journals have a higher proportion of rejected articles.

Clusters 2 (9 representatives) and 3 (11 representatives) had more similarities. However, all the journals that combine high Science Index ratings, RSCI inclusion, and HAC Category 1 were placed in Cluster 3, setting it apart from the others. Journals in these categories often reject a larger proportion of submitted articles compared to Categories 1 and 4 and require a significantly longer time for publication, which may be due to a more rigorous peer-review system (see Figure 5).



a) Sciecne Index b) Number of articles



c) Inclusion in the HAC list

d) Inclusion in the RSCI core

e) Publication period (days)

f) Percentage of rejected articles

Fig. 5. Structure of values of journals included in clusters by indicators

Рис. 5. Структура значений журналов, входящих в кластеры, по показателям

In the context of forecasting, it is anticipated that the indicators of journals from Clusters 4 and 2 are likely to decrease, while Cluster 3 may remain relatively stable. On the other hand, Cluster 1 and some Cluster 3 journals are expected to experience growth.

CONCLUSION

The conducted forecasting and clustering of scientometric indicators for Russian journals specializing in the field of informatics have shown relative stability within the industry. This segment of scientific publications is not as numerous compared to some other fields, which affects the assessment but also implies that there is room for new journals in this area. Overall, it is recommended to strive for journal inclusion in the Higher Attestation Commission (HAC) and especially in the Core of the RSCI journal list.

Increase in the number of publications doesn't always lead to improved scientometric indicators, and vice versa. Therefore, it's necessary to find a balance between the quantity of publications and the quality of their content. Maintaining a reasonable time from submission to publication that is convenient for authors is important.

There are journals with high Science Index ratings and a short time to publication, but they also have a significant proportion of rejected works. Achieving such status requires rigorous selection and a commitment to maintaining high publication quality. It's important to strike a balance between a quick review process and rigorous academic selection to attract high-quality and impactful research that contributes to journal metrics.

Author plans to continue working on this research, expanding the scope to consider additional factors and indicators influencing journal development and reflecting its position in the competitive publishing market. This will provide a more comprehensive understanding of the dynamics of scientific journals in the field of informatics and offer valuable insights for journal editors, publishers, and the broader scientific community.

References

1. Aggarwal D., Sharma D. Application of clustering for student result analysis. International Journal of Recent Technology and Engineering. 7(6) (2019): 50-53.
2. Armstrong J.S. Research on scientific journals: Implications for editors and authors. Journal of forecasting. 1(1) (1982): 83-104.
3. Aydynov Z.P., Nurkasheva N.S., Karabasov R.A. Fundamentals of time series forecasting based on the ARIMA method. Statistics, accounting and audit. 4(75) (2019): 184-191.
4. Distribution of journals included in the list of peer-reviewed scientific publications in which the main results of dissertations for the degree of candidate of sciences, the degree of doctor of sciences should be published by category. Higher Attestation Commission (HAC) under the Ministry of Science and Higher Education of the Russian Federation. URL: https://HAC.minobrnauki.gov.ru/uploader/loader?type=19&name=92263438002&f=14239 (date of application: 15.05.2023).
5. Ferreira M., Santos A., Lucio P. Short-term forecast of wind speed through mathematical models. Energy Reports. 5 (2019): 1172-1184.
6. Ivancheva L. Scientometrics today: A methodological overview. Collnet Journal of Scientometrics and Information Management. 2(2) (2008):47-56.
7. Kalugina M.I., Begicheva S.V. Modern possibilities of visualization of research results in the R environment. BI-technologies and corporate information systems in business process optimization: Materials of the IV International scientific and practical conference. Yekaterinburg: Ural State University of Economics (2016): 51-55.
8. Kryukov Yu.A., Chernyagin, D.V. Arima – a model for predicting traffic values. Information technologies and computing systems. 2 (2011): 41-49.
9. Leydesdorff L. The prediction of science indicators using information theory. Scientometrics. 19(3-4) (2019): 297–324. DOI: 10.1007/bf02095353.
10. Liu H. Forecast of the trend in incidence of acute hemorrhagic conjunctivitis in China from 2011–2019 using the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Exponential Smoothing (ETS) models. Journal of infection and public health. 13(2) (2020): 287-294.
11. Min A. A study of likelihood functions of ARIMA (autoregressive integrated moving average) processes (1976).
12. Moskaleva O. Russian index of science citation: Overview and review. Scientometrics. 116 (2018): 449-462.
13. Najarzadeh R. Multi-objective TS fuzzy control of Covid-19 spread model: An LMI approach. Biomedical Signal Processing and Control. 79 (2023): 104107.
14. RDocumentation. URL: https://www.rdocumentation.org/ (date of application: 15.05.2023).
15. R-Studio. URL: https://www.r-studio.com/ru/ (date of application: 15.05.2023).
16. Scientific Electronic Library «eLibrary.Ru». URL: https://elibrary.ru/ (date of application: 15.05.2023).
17. Tseng Y.H., Lin Y.I., Hung W.C., Lee C.H. A comparison of methods for detecting hot topics. Scientometrics. 81(1) (2009): 73–90. DOI: 10.1007/s11192-009-1885-x.
18. Van Fleet D.D., McWilliams A., Siegel D.S. A theoretical and empirical analysis of journal rankings: The case of formal lists. Journal of management. 26(5) (2000): 839-861.
19. Vasileiadou E. Stabilisation operationalised: Using time series analysis to understand the dynamics of research collaboration. Journal of Informetrics. 3(1) (2009): 36–48. DOI: 10.1016/j.joi.2008.11.002.
20. Wallin J.A. Bibliometric methods: pitfalls and possibilities. Basic & clinical pharmacology & toxicology. 97(5) (2005): 261-275.
21. Yazdani K. Scientometrics: Review of concepts, applications, and indicators. Iranian Journal of Epidemiology. 10(4) (2005): 78-88.

Список литературы

1. Aggarwal D., Sharma D. Application of clustering for student result analysis. International Journal of Recent Technology and Engineering. 7(6) (2019): 50-53.
2. Armstrong J.S. Research on scientific journals: Implications for editors and authors. Journal of forecasting. 1(1) (1982): 83-104.
3. Айдынов З.П., Нуркашева Н.С., Карабасов Р.А. Основы прогнозирования временных рядов на основе метода ARIMA // Статистика, учет и аудит. 2019. № 4(75). С. 184-191.

4. Распределение журналов, входящих в перечень рецензируемых научных изданий, в которых должны быть опубликованы основные результаты диссертаций на соискание ученой степени кандидата наук, ученой степени доктора наук по категориям // Высшая аттестационная комиссия (ВАК) при Министерстве науки и высшего образования Российской Федерации. URL: https://vak.minobrnauki.gov.ru/uploader/loader?type=19&name=92263438002&f=14239 (дата обращения: 15.05.2023).

5. Ferreira M., Santos A., Lucio P. Short-term forecast of wind speed through mathematical models. Energy Reports. 5 (2019): 1172-1184.

6. Ivancheva L. Scientometrics today: A methodological overview. Collnet Journal of Scientometrics and Information Management. 2(2) (2008):47-56.

7. Калугина М.И., Бегичева С.В. Современные возможности визуализации результатов исследований в среде R // BI-технологии и корпоративные информационные системы в оптимизации бизнес-процессов: Материалы IV Международной научно-практической конференции. Екатеринбург: Уральский государственный экономический университет, 2016. С. 51-55.

8. Крюков Ю.А., Чернягин Д.В. Arima – модель прогнозирования значений трафика // Информационные технологии и вычислительные системы. 2011. № 2. С. 41-49.

9. Leydesdorff L. The prediction of science indicators using information theory. Scientometrics. 19(3-4) (2019): 297–324. DOI: 10.1007/bf02095353.

10. Liu H. Forecast of the trend in incidence of acute hemorrhagic conjunctivitis in China from 2011–2019 using the Seasonal Autoregressive Integrated Moving Average (SARIMA) and Exponential Smoothing (ETS) models. Journal of infection and public health. 13(2) (2020): 287-294.

11. Min A. A study of likelihood functions of ARIMA (autoregressive integrated moving average) processes (1976).

12. Moskaleva O. Russian index of science citation: Overview and review. Scientometrics. 116 (2018): 449-462.

13. Najarzadeh R. Multi-objective TS fuzzy control of Covid-19 spread model: An LMI approach. Biomedical Signal Processing and Control. 79 (2023): 104107.

14. RDocumentation. URL: https://www.rdocumentation.org/ (date of application: 15.05.2023).

15. R-Studio. URL: https://www.r-studio.com/ru/ (date of application: 15.05.2023).

16. Scientific Electronic Library «eLibrary.Ru». URL: https://elibrary.ru/ (date of application: 15.05.2023).

17. Tseng Y.H., Lin Y.I., Hung W.C., Lee C.H. A comparison of methods for detecting hot topics. Scientometrics. 81(1) (2009): 73–90. DOI: 10.1007/s11192-009-1885-x.

18. Van Fleet D.D., McWilliams A., Siegel D.S. A theoretical and empirical analysis of journal rankings: The case of formal lists. Journal of management. 26(5) (2000): 839-861.

19. Vasileiadou E. Stabilisation operationalised: Using time series analysis to understand the dynamics of research collaboration. Journal of Informetrics. 3(1) (2009): 36-48. DOI: 10.1016/j.joi.2008.11.002.

20. Wallin J.A. Bibliometric methods: pitfalls and possibilities. Basic & clinical pharmacology & toxicology. 97(5) (2005): 261-275.

21. Yazdani K. Scientometrics: Review of concepts, applications, and indicators. Iranian Journal of Epidemiology. 10(4) (2005): 78-88.

Sokolova Elizaveta Vitalievna, Specialist of the Department of Scientometrics, R&D and Ratings, Ural State University of Economics

Соколова Елизавета Витальевна, специалист Управления наукометрии, научно-исследовательской работы и рейтингов, Уральский государственный экономический университет

УДК 004.65 **DOI: 10.18413/2518-1092-2022-8-3-0-7**

Гармашев М.А.
Резников Н.Г.

ТЕХНОЛОГИЯ NOSQL: ТЕХНИЧЕСКИЕ ОСОБЕННОСТИ, ПЕРСПЕКТИВЫ РАЗВИТИЯ И СРАВНИТЕЛЬНЫЕ ПРЕИМУЩЕСТВА ПЕРЕД SQL

Белгородский государственный национальный исследовательский университет,
ул. победы, 85, Белгород, 308015, Россия

e-mail: garmashev@bsu.edu.ru

Аннотация
В данной обзорной статье рассматриваются концепции, технические особенности и перспективы развития технологии NoSQL, подчеркиваются ее преимущества по сравнению с обычными базами данных SQL. Базы данных NoSQL привлекли внимание благодаря своей способности эффективно обрабатывать огромные объемы неструктурированных и полуструктурированных данных. Авторы проводят краткий анализ ключевых характеристик баз данных NoSQL через частичное сравнение с классической SQL. В статье рассматриваются преимущества технологии NoSQL по сравнению с традиционными базами данных SQL, такие как повышенная производительность, упрощенное моделирование данных, повышенная масштабируемость и поддержка распределенных сред. В заключение перечислены будущие перспективы технологии NoSQL и рост используемости в современных приложениях, ориентированных на большие данные.
Ключевые слова: информация; аналитика; интернет; хранилище; ключи; значения; база данных; графовая база данных; распределенная база данных; горизонтальное масштабирование
Для цитирования: Гармашев М.А., Резников Н.Г. Технология NoSQL: технические особенности, перспективы развития и сравнительные преимущества перед SQL // Научный результат. Информационные технологии. – Т.8, №3, 2023. – С. 56-62. DOI: 10.18413/2518-1092-2022-8-3-0-7

Garmashev M.A.
Reznikov N.G.

NOSQL TECHNOLOGY: TECHNICAL FEATURES, DEVELOPMENT PROSPECTS AND COMPARATIVE ADVANTAGES OVER SQL

Belgorod State National Research University, 85 Pobedy St., Belgorod, 308015, Russia

e-mail: garmashev@bsu.edu.ru

Abstract
This review article examines the concepts, technical features, and future developments of NoSQL technology, highlighting its advantages over conventional SQL databases. NoSQL databases have gained attention due to their ability to efficiently process huge volumes of unstructured and semi-structured data. The authors provide a brief analysis of the key characteristics of NoSQL databases through a partial comparison with classical SQL. This article discusses the benefits of NoSQL technology over traditional SQL databases, such as improved performance, simplified data modeling, increased scalability, and support for distributed environments. Finally, the future prospects of NoSQL technology and its increasing use in modern big data applications are listed.
Keywords: information; analytics; Internet; storage; keys; values; database; graph database; distributed database; horizontal scaling
For citation: Garmashev M.A., Reznikov N.G. NoSQL technology: technical features, development prospects and comparative advantages over SQL // Research result. Information technologies. – T.8, №3, 2023. – P. 56-62. DOI: 10.18413/2518-1092-2022-8-3-0-7

ВВЕДЕНИЕ

В последние годы наблюдается значительный рост использования технологии NoSQL благодаря ее способности обрабатывать большие объемы неструктурированных данных и обеспечивать высокую масштабируемость и производительность. Программисты заинтересованы в использовании технологии NoSQL, поскольку она обеспечивает гибкое моделирование данных, более быстрые циклы разработки и горизонтальную масштабируемость. Базы данных NoSQL также поддерживают различные типы данных и хорошо подходят для таких случаев использования, как аналитика в реальном времени, системы управления контентом и приложения с быстро меняющимися требованиями к данным. Базы данных NoSQL используются различными организациями в России, включая технологические компании, стартапы, государственные учреждения, исследовательские институты и предприятия в различных секторах, таких как финансы, телекоммуникации, электронная коммерция и производство. В число известных компаний, использующих базы данных NoSQL в России, входят Яндекс, Mail.Ru Group, Сбербанк, Газпром, Ростелеком и ВКонтакте. NoSQL (что означает «не только SQL») – это категория систем баз данных, которые отличаются от традиционных реляционных баз данных. Базы данных NoSQL не были изобретены одним человеком, а возникли как ответ на требования современных веб-приложений и больших данных. Различные компании и разработчики внесли свой вклад в разработку баз данных NoSQL, включая Apache Cassandra, MongoDB и CouchDB.

КРАТКИЙ АНАЛИЗ ПЕРСПЕКТИВЫ ТЕХНОЛОГИИ NOSQL

1. Гибкость схемы:

Базы данных NoSQL предлагают гибкую модель данных, позволяющую разработчикам хранить и извлекать данные, не придерживаясь предопределенной схемы. Такая гибкость учитывает меняющиеся требования к данным и облегчает быструю разработку и итерацию, что делает NoSQL идеальным для гибкой разработки и приложений больших данных.

2. Масштабируемость и высокая производительность:

Базы данных NoSQL предназначены для горизонтального масштабирования, то есть они могут обрабатывать большие объемы данных, распределяя их по нескольким серверам или кластерам. Это обеспечивает высокую пропускную способность записи и чтения, что делает базы данных NoSQL хорошо подходящими для обработки огромных рабочих нагрузок и обслуживания приложений реального времени.

3. Репликация и доступность данных:

Большинство баз данных NoSQL предлагают автоматическую репликацию данных, гарантируя, что данные реплицируются на несколько узлов для обеспечения отказоустойчивости и высокой доступности. Эта функция предотвращает единые точки отказа и обеспечивает встроенную избыточность, гарантируя долговечность данных и бесперебойное обслуживание.

4. Запрос на гибкость:

Базы данных NoSQL используют различные модели запросов, такие как базы данных «ключ-значение», «документ», «семейство столбцов» и «граф». Каждая модель предлагает уникальные возможности запроса, подходящие для конкретных шаблонов доступа к данным и сценариев использования. Базы данных NoSQL часто обеспечивают эффективный поиск по ключам, широкие возможности выполнения запросов и поддержку сложных структур данных [1].

5. Распределенная и отказоустойчивая архитектура:

Базы данных NoSQL созданы с учетом распределенной архитектуры. Они используют такие методы, как сегментирование и секционирование, для распределения данных по нескольким серверам, обеспечивая горизонтальное масштабирование. Кроме того, базы данных NoSQL предназначены для корректной обработки сбоев серверов, обеспечивая высокую доступность и отказоустойчивость.

6. Поддержка неструктурированных и полуструктурированных данных:

Базы данных NoSQL превосходно справляются с обработкой неструктурированных и полуструктурированных данных, таких как документы JSON, XML или двоичный контент. Они обеспечивают гибкость хранения и индексирования этих разнообразных форматов данных, позволяя эффективно находить и анализировать сложные структуры данных.

7. Типы баз данных NoSQL

- Документоориентированные базы данных (например, MongoDB)
- Столбчатые базы данных (например, Apache Cassandra)
- Базы данных «ключ-значение» (например, Redis)
- Графовые базы данных (например, Neo4j)

СРАВНЕНИЕ БАЗЫ ДАННЫХ NOSQL И БАЗЫ ДАННЫХ SQL

1. Модель данных. Базы данных SQL используют структурированную и предопределенную схему с таблицами, строками и столбцами, тогда как базы данных NoSQL используют гибкую и динамическую модель данных, часто основанную на парах ключ-значение, документах, графиках или хранилищах с широкими столбцами.

2. Масштабируемость. Базы данных NoSQL предназначены для горизонтального масштабирования, что означает, что они могут обрабатывать большие объемы данных и высокую нагрузку трафика на нескольких серверах. Базы данных SQL обычно масштабируются вертикально, требуя более мощного оборудования по мере роста данных.

3. Гибкость схемы. Базы данных NoSQL обеспечивают гибкость схемы, позволяя легко добавлять, изменять и удалять поля в модели данных, не затрагивая существующие данные. Базам данных SQL требуется предопределенная схема, а изменения могут потребовать модификации схемы и миграции данных.

4. Язык запросов. Базы данных SQL используют язык запросов SQL, стандартизированный язык для управления данными и манипулирования ими. Базы данных NoSQL часто имеют свои собственные языки запросов, которые могут различаться в зависимости от типа базы данных. Некоторые базы данных NoSQL также поддерживают запросы, подобные SQL.

5. Соответствие ACID. Базы данных SQL обычно соответствуют требованиям ACID (атомарность, согласованность, изоляция, долговечность), что обеспечивает согласованность и надежность данных. Базы данных NoSQL по умолчанию обеспечивают итоговую согласованность, отдавая приоритет высокой доступности и масштабируемости, а не строгой согласованности.

6. Варианты использования. Базы данных SQL хорошо подходят для приложений со структурированными и сложными связями, требующих строгой целостности данных, таких как транзакционные системы и приложения, использующие финансовые данные. Базы данных NoSQL часто используются в сценариях с большими объемами неструктурированных или полуструктурированных данных, обработкой данных в реальном времени и высокими требованиями к масштабируемости [2]. На рисунке представлен краткий анализ сравнения разных технологий.

Рис. Краткий анализ сравнения разных технологий
Fig. Brief comparison analysis of different technologies

Важно отметить, что это общие характеристики, и существуют различные типы баз данных NoSQL, каждая из которых имеет свои сильные и слабые стороны. Выбор между SQL и NoSQL зависит от конкретных требований и характера разрабатываемого приложения или системы.

НЕКОТОРЫЕ РАСПРОСТРАНЕННЫЕ ОШИБКИ ПРИ ПЕРЕХОДЕ ОТ БАЗ ДАННЫХ SQL К БАЗАМ ДАННЫХ NOSQL

Следует знать, что успешный переход от SQL к NoSQL требует понимания различий между этими типами баз данных и соответствующей адаптации подхода программиста.

1. Игнорирование моделирования данных. Базы данных NoSQL требуют другого подхода к моделированию данных по сравнению с SQL. Неспособность разработать соответствующую модель данных может привести к проблемам с производительностью и неэффективным запросам.
2. Чрезмерное использование денормализации. Хотя в некоторых случаях денормализация может улучшить производительность запросов, чрезмерное ее использование может привести к избыточности данных и увеличению сложности. Важно найти баланс и тщательно взвесить компромиссы.
3. Пренебрежение индексацией. Базы данных NoSQL часто предоставляют гибкие варианты схемы, но эта гибкость может привести к проблемам с производительностью, если индексы не используются эффективно. Правильная индексация имеет решающее значение для оптимизации производительности запросов.
4. Не учитывается масштабируемость. Базы данных NoSQL известны своей масштабируемостью, но важно планировать будущий рост. Неспособность разработать масштабируемую архитектуру может привести к трудностям при увеличении объема данных или рабочей нагрузки.
5. Недооценка проблем согласованности данных. Базы данных NoSQL, особенно ориентированные на высокую масштабируемость, могут обеспечивать более слабые гарантии согласованности по сравнению с традиционными базами данных SQL. Очень важно понять модель согласованности выбранной базы данных NoSQL и соответствующим образом спроектировать приложение.

6. Недостаточное знакомство с языками запросов и API. Базы данных NoSQL часто имеют свои собственные языки запросов и API, которые могут отличаться от SQL. Неспособность изучить и понять эти новые языки и API может привести к неэффективным или неверным запросам.

7. Предполагается, что NoSQL является универсальным решением: базы данных NoSQL превосходны в определенных случаях использования, но могут подходить не для всех сценариев. Важно тщательно оценить требования конкретного приложения и выбрать технологию базы данных, которая лучше всего соответствует техническому заданию [3].

NoSQL может различаться по сложности в зависимости от конкретной технологии баз данных и предыдущего опыта работы с базами данных. Некоторые базы данных NoSQL довольно просты в освоении, в то время как другие могут быть требовательны к обучению. Однако в целом базы данных NoSQL спроектированы так, чтобы быть гибкими и масштабируемыми, что может облегчить работу с ними в определенных случаях использования по сравнению с традиционными реляционными базами данных.

НЕДОСТАТКИ ТЕХНОЛОГИИ NOSQL

Важно отметить, что обозреваемые недостатки относятся не ко всем базам данных NoSQL, поскольку существуют разные типы и реализации с разными функциями и компромиссами.

1. Отсутствие стандартизированного языка запросов. В базах данных NoSQL часто отсутствует стандартизированный язык запросов, такой как SQL, что усложняет выполнение сложных запросов и анализ данных.

2. Ограниченная поддержка транзакций. Базы данных NoSQL могут не обеспечивать полные гарантии транзакций ACID (атомарность, согласованность, изоляция, долговечность), что делает их менее подходящими для приложений, требующих строгой согласованности данных.

3. Дублирование и денормализация данных. Базы данных NoSQL часто денормализуют данные и дублируют их в нескольких документах или коллекциях для оптимизации производительности чтения. Это может привести к увеличению использования дискового пространства и потенциальной несогласованности данных, если обновления выполняются неправильно.

4. Ограниченные инструменты и экосистема. По сравнению с реляционными базами данных базы данных NoSQL часто имеют меньший набор инструментов и экосистемную поддержку, что может усложнить такие задачи, как мониторинг, резервное копирование и восстановление, а также управление схемой.

5. Недостаточная зрелость и стабильность. Некоторые базы данных NoSQL являются относительно новыми и могут не обладать зрелостью и стабильностью более устоявшихся реляционных баз данных. Это может привести к потенциальным проблемам с ошибками, производительностью и целостностью данных [5].

6. Более крутая кривая обучения. Базы данных NoSQL часто требуют от разработчиков изучения новых моделей данных и языков запросов, что может быть более сложным и трудоемким по сравнению с традиционными базами данных на основе SQL [5].

ЗАКЛЮЧЕНИЕ

В данной обзорной статье представлен всесторонний обзор технологии NoSQL, ее технических особенностей и перспектив развития. В нем подчеркиваются преимущества баз данных NoSQL по сравнению с традиционными системами SQL, подчеркиваются их масштабируемость, гибкость и преимущества в производительности. Рост больших данных и потребность в аналитике в реальном времени создают возможности для баз данных NoSQL для эффективной обработки крупномасштабных и быстро меняющихся наборов данных. Институты и исследовательские организации России активно участвуют в изучении и развитии технологий управления данными, в том числе NoSQL. Потенциальное сотрудничество между научными кругами и промышленными

партнерами облегчит обмен знаниями и может привести к прогрессу в разработке NoSQL. Поскольку объем и сложность данных продолжают расти, технология NoSQL, вероятно, будет играть решающую роль в современном управлении данными и аналитике. Перспективы развития NoSQL в России могут быть многообещающими благодаря растущему спросу на масштабируемые решения для управления данными, процветающей экосистеме стартапов, академическим исследованиям и государственной поддержке.

Список литературы

1. Попов В.Б. Технологии "nosql" в алгоритмах анализа больших данных и искусственном интеллекте / В. Б. Попов, И. В. Гавриков // Проблемы информационной безопасности: V Всероссийская с международным участием научно-практическая конференция, Симферополь – Гурзуф, 14-16 февраля 2019 года / Крымский федеральный университет имени В.И. Вернадского. – Симферополь – Гурзуф: ИП Зуева Т.В., 2019. – С. 158-160. – EDN YXYRAD.

2. Абелян А.А., Шерстобитов Г.В., Уварова А.А., Плужник Е.Н. Нереляционные базы данных. Обзор технологии nosql. Студенческая наука для развития информационного общества: сборник материалов VII Всероссийской научно-технической конференции, Ставрополь, 26–28 декабря 2017 года. Том Часть 2. – Ставрополь: Северо-Кавказский федеральный университет, 2018. – С. 15-19. – EDN YOTOKG.

3. Матыцын А.В., Есина А.Р. Перспективы интеграции технологии nosql в производственный процесс предприятия. Экономическое развитие в XXI веке: тенденции, вызовы и перспективы: Сборник научных трудов IX Международной научно-практической конференции «Горизонты России» В двух частях.

4. Буслаева У.Ю. Тенденции развития noSQL-технологий и систем управления базами данных. Тридцатая годичная сессия ученого совета Сыктывкарского государственного университета имени Питирима Сорокина: Сборник статей Национальной конференции. В 2-х частях, Сыктывкар, 01–28 февраля 2023 года / Отв. редакторы О.А. Сотникова, Н.Н. Новикова. Том Часть 2. – Сыктывкар: Сыктывкарский государственный университет им. Питирима Сорокина, 2023. – С. 276-278. – EDN VBWTJQ.

5. Быкова К.И., Кузьмичева Е.А., Павлов П.А. Преимущества и недостатки больших баз данных, проблемы их сжатия и разработки. Информационные технологии в образовательном процессе вуза и школы: Материалы XVI Всероссийской научно-практической конференции, Воронеж, 30 марта 2022 года / Редколлегия: Р.М. Чудинский (науч. ред.), В.В. Малев, А.А. Малева (отв. ред.). – Воронеж: Воронежский государственный педагогический университет, 2022. – С. 47-53. – EDN FFGFLD.

References

1. Popov V.B., Gavrikov I.V. "Nosql" technologies in big data analysis algorithms and artificial intelligence. Problems of information security: V All-Russian scientific and practical conference with international participation, Simferopol – Gurzuf, February 14-16, 2019 / Crimean Federal University named after V.I. Vernadsky. – Simferopol – Gurzuf: IP Zueva T.V., 2019. – P. 158-160. – EDN YXYRAD.

2. Abelyan A.A., Sherstobitov G.V., Uvarova A.A., Pluzhnik E.N. Non-relational databases. Review of nosql technology. Student science for the development of the information society: collection of materials of the VII All-Russian Scientific and Technical Conference, Stavropol, 26–28 December 2017. Volume Part 2. – Stavropol: North Caucasus Federal University, 2018. – pp. 15-19. – EDN YOTOKG.

3. Matytsyn A.V., Esina A.R. Prospects for integrating nosql technology into the production process of an enterprise. Economic development in the 21st century: trends, challenges and prospects: Collection of scientific papers of the IX International Scientific and Practical conference “Russian Horizons” In two parts

4. Buslaeva U.Yu. Trends in the development of noSQL technologies and database management systems. Thirtieth annual session of the Academic Council of Syktyvkar State University named after Pitirim Sorokin: Collection of articles of the National Conference. In 2 parts, Syktyvkar, February 01–28, 2023 / Rep. editors O.A. Sotnikova, N.N. Novikova. Volume Part 2. – Syktyvkar: Syktyvkar State University named after. Pitirim Sorokina, 2023. – pp. 276-278. – EDN VBWTJQ.

5. Bykova K.I., Kuzmicheva E.A., Pavlov P.A. Advantages and disadvantages of large databases, problems of their compression and development. Information technologies in the educational process of universities and schools: Materials XVI All-Russian Scientific and Practical Conference, Voronezh, March 30, 2022 / Editorial Board: R.M. Chudinsky (scientific editor), V.V. Malev, A.A. Maleva (responsible editor). – Voronezh: Voronezh State Pedagogical University, 2022. – P. 47-53. – EDN FFGFLD.

Гармашев Михаил Александрович, старший преподаватель кафедры гражданского права и процесса юридического института
Резников Никита Григорьевич, ассистент кафедры информационных и робототехнических систем института инженерных и цифровых технологий

Garmashev Mikhail Aleksandrovich, senior lecturer of the Department of Civil Law and Process of the Law Institute
Reznikov Nikita Grigorievich, assistant of the department of information and robotic systems of the Institute of Engineering and Digital Technologies