

ВЕСТНИК НОВОСИБИРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Научный журнал
Основан в ноябре 1999 года

Серия: Информационные технологии

2022. Том 20, № 3

СОДЕРЖАНИЕ

<i>Бутенко Ю. И.</i> Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов	5
<i>Гаркуша В. В., Журавлев С. С., Шакиров С. Р., Яковлев В. В.</i> Отладочный комплекс с имитацией кондуктивных помех для тестирования контроллеров АСУ ТП	14
<i>Клименко О. А., Рычкова Е. В., Шабальников И. В.</i> Развитие информационной среды СО РАН в условиях санкционных ограничений и информационной войны против РФ	29
<i>Дементьева К. И., Ракитский А. А.</i> Метод для восстановления аудиосигнала с помощью свёрточных нейронных сетей.....	38
<i>Сидорова Е. А., Долгова А. В., Железняк С. П.</i> Автоматизированная система контроля формирования индикаторов достижения компетенций при изучении основ алгебры логики	51
<i>Тихобаева О. Ю., Бручес Е. П., Батура Т. В.</i> Извлечение семантических отношений из текстов научных статей.....	65
Информация для авторов	77

VESTNIK

NOVOSIBIRSK STATE UNIVERSITY

Scientific Journal
Since 1999, November
In Russian

Series: Information Technologies

2022. Volume 20, № 3

CONTENTS

<i>Butenko I. I.</i> Model of the Text of a Scientific and Technical Article for Markup in the Corpus of Scientific and Technical Texts	5
<i>Garkusha V. V., Zhuravlev S. S., Shakirov S. R., Yakovlev V. V.</i> Debugging Complex with Simulation of Conductive Interference for Testing Automated Control System Controllers	14
<i>Klimenko O. A., Rychkova E. V., Shabalnikov I. V.</i> Development of the Information Environment of SB RAS under Sanctions and Information Warfare against the Russian Federation	29
<i>Dementyeva K. I., Rakitskiy A. A.</i> Method for Restoring Audio Signal Using Convolutional Neural Networks	38
<i>Sidorova E. A., Dolgova A. V., Zheleznyak S. P.</i> Automated System of Competencies Indicators' Formation Control during the Basic Boolean Algebra Studying	51
<i>Tikhobaeva O. Y., Bruches E. P., Batura T. V.</i> Extracting Semantic Relations from the Texts of Scientific Articles	65
Instructions for Contributors.....	77

Editor in Chief M. M. Lavrentiev

Vice-Editor A. V. Avdeev

Executive Secretary D. P. Iksanova

Editorial Board of the Series

- I. V. Bychkov*, professor, academician (Irkutsk), *B. M. Glinsky*, professor (Novosibirsk)
A. N. Gorban, professor (Lester, GB), *E. P. Gordov*, professor (Tomsk)
B. S. Dobronets, professor (Krasnoyarsk), *A. M. Elizarov*, professor (Kazan)
G. N. Erokhin, professor (Kaliningrad), *A. I. Kamysnikov*, professor (Khanty-Mansiysk)
G. P. Karev, professor (Maryland, USA), *N. A. Kolchanov*, professor, academician (Novosibirsk)
M. M. Lavrentjev, professor (Novosibirsk), *V. E. Malyshev*, professor (Novosibirsk)
N. N. Mirenikov, professor (Aizu, Japan), *N. M. Oskorbin*, professor (Barnaul)
D. E. Palchunov, professor (Novosibirsk), *T. Pizansky*, professor (Ljubljana, Slovenia)
V. P. Potapov, professor (Kemerovo), *O. I. Potaturkin*, professor (Novosibirsk)
V. A. Serebryakov, professor (Moscow), *A. V. Starchenko*, professor (Tomsk)
S. I. Smagin, professor, corresponding member of RAS (Khabarovsk)
D. A. Tusupov, professor (Astana, Kazakhstan)
V. V. Shajdurov, professor, corresponding member of RAS (Krasnoyarsk)
Yu. I. Shokin, professor, academician (Novosibirsk)

*The journal is published quarterly in Russian since 1999
by Novosibirsk State University Press*

*The address for correspondence
Faculty of Information Technologies, Novosibirsk State University
1 Pirogov Street, Novosibirsk, 630090, Russia
Tel. +7 (383) 363 42 46*

E-mail address: inftech@vestnik.nsu.ru

On-line version: <http://elibrary.ru>

Научная статья

УДК 004.89

DOI 10.25205/1818-7900-2022-20-3-5-13

Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов

Юлия Ивановна Бутенко

Московский государственный технический университет им. Н. Э. Баумана

Москва, Россия

iubutenko@bmstu.ru

Аннотация

В статье предложена модель текста научно-технической статьи для автоматизации разметки в корпусе научно-технических текстов. Обосновано, что при создании корпуса научно-технических текстов необходимо учитывать структурные особенности текстов научно-технических статей. Показана необходимость добавления структурной разметки в корпус научно-технических текстов. Отмечено, что тексты научно-технических статей имеют одинаковую для всех текстов этого класса структуру изложения материала, а также содержат ограниченный набор структурных элементов. Проанализированы особенности композиционной организации текстов научно-технических статей. Описано примерное содержание каждого из элементов структуры статьи. Представлена композиционная структура текстов научно-технических статей в нотациях Бекуса-Наура. Предложена модель текста научно-технической статьи в виде графа, вершинами и ребрами которого являются полноценные структурные элементы научно-технической статьи. Обосновано, что представление текста научно-технической статьи в виде графа дает возможность в процессе компьютерного анализа текста определить тип структурного элемента, степень вложенности, за счет подачи научно-технической статьи в виде конечного множества ее составных частей. Обосновано, что наличие структурной разметки в корпусе научно-технических текстов значительно расширит его исследовательский потенциал и послужит базой для задач автоматической обработки научно-технических текстов.

Ключевые слова

научная статья, структура текста, структурный элемент, корпус научно-технических текстов, модель текста

Для цитирования

Бутенко Ю. И. Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 5–13. DOI 10.25205/1818-7900-2022-20-3-5-13

Model of the Text of a Scientific and Technical Article for Markup in the Corpus of Scientific and Technical Texts

Yulia I. Butenko

Bauman Moscow State Technical University

Moscow, Russian Federation

iubutenko@bmstu.ru

Abstract

The paper proposes a model of the text of a scientific and technical article for the automation of markup in the corpus of scientific and technical texts. It is proved that when creating a corpus of scientific and technical texts, it is necessary to take into account the structural features of texts of scientific and technical articles. The necessity of adding structural markup to the corpus of scientific and technical texts has been shown. It is noted that the texts of scientific and technical articles have the same narration structure for all texts in this class, and also contain a limited set of structural elements. The features of compositional organization of the texts of scientific and technical articles are analyzed. The approximate content of each of the elements of article structure is described. Compositional structure of the texts of scientific and

© Бутенко Ю. И., 2022

technical articles in Bekus-Naur notation is presented. A model of the text of a scientific and technical article in the form of a graph, the vertices and edges of which are the full-fledged structural elements of a scientific and technical article, is proposed. It is proved that the representation of a text of scientific and technical article in the form of a graph makes it possible to determine the type of structural element and the degree of nesting in the process of computer analysis of the text by presenting the scientific and technical article as a finite set of its constituent parts. It is proved that the presence of structural markup in the corpus of scientific and technical texts significantly expands its research potential and serves as the basis for the tasks of automatic processing of scientific and technical texts.

Keywords

scientific article, text structure, structural element, corpus of scientific and technical texts, text model

For citation

Butenko Yu. I. Model of the Text of a Scientific and Technical Article for Markup in the Corpus of Scientific and Technical Texts. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 5–13. DOI 10.25205/1818-7900-2022-20-3-5-13

Введение

Отличительной особенностью современного мира является накопление огромного фонда различной информации, большая часть которой хранится в электронном виде. Одними из наиболее представительных информационных ресурсов текстов можно назвать электронные корпуса текстов [1–2]. Для описания подязыка определенной предметной области необходимо использование специальных корпусов узкоспециальных текстов – корпусов научно-технических текстов, так как общие корпуса не подходят для изучения определенных предметных областей в силу их большого объема, разнообразного материала, а также отсутствия специальной терминологии [3].

Тексты, включаемые в корпус, проходят тщательный отбор и группируются в подкорпуса [4]. К источникам текстов, обеспечивающих репрезентативность корпуса научно-технических текстов, прежде всего, следует отнести научно-технические статьи, опубликованные в специализированных журналах.

Электронный корпус представляет собой коллекции текстов и их разметку, зависящую от типа исследования или задачи, для решения которой они созданы [3]. Разметка позволяет сделать корпус гораздо удобнее в использовании и является главной отличительной особенностью корпуса по сравнению с любыми другими коллекциями текстов. Таким образом, создание корпуса научно-технических текстов предполагает наличие лингвистической разметки, которая описывает сугубо лингвистические характеристики языковой выборки корпуса и представляет собой сложный процесс, требующий длительной и кропотливой работы над каждой лексической единицей, представленной в корпусе. Лингвистическая разметка обычно включает в себя разметку морфологическую, синтаксическую, семантическую [5]. Одним из ключевых аспектов проектирования корпусов является также метаразметка текстов – процесс приписывания тексту различных характеристик, описывающих обстоятельства его создания, автора, соотносённости с определенным жанром и стилем изложения [6]. Основное назначение метаразметки – дать возможность пользователям корпуса настроить внешние параметры поиска текстов: например, осуществлять поиск по текстам, созданным авторами определенного года рождения, страны происхождения, гендерной принадлежности. Метаразметка содержит основную информацию о каждом тексте, включенном в корпус.

Стоит отметить, что научно-технические тексты обладают рядом специфических особенностей, которые требуют использования дополнительных видов разметки. К таким особенностям следует отнести композиционную структуру научно-технических текстов, которая может оказывать существенное влияние на результаты их автоматической обработки. Так, например, поиск информации в текстах стандартов необходимо проводить в два этапа. На первом этапе отбирать стандарты или их разделы, а затем искать в отобранном массиве необходимую информацию. Данный факт объясняется особенностями композиционной структуры текстов

стандартов, с одной стороны, и обобщенно-отвлеченным характером лексики, используемой при изложении требований стандартов, с другой стороны [7]. Наличие структурной разметки научно-технических текстов позволит отбирать для исследования только определенные структурные компоненты научно-технического текста, например, аннотации или введения.

Целью статьи является построение модели текста научно-технической статьи для структурной разметки в корпусе научно-технических текстов.

1. Научно-техническая статья как структурированный текст

Научно-техническая статья – это первичный письменный жанр научного дискурса, задачей которого является постановка и решение одной научной проблемы, имеет средний объем, конвенциональную структуру, системы ссылок и выходные данные [8]. Научно-техническим статьям присущи все стиливые особенности научного стиля: точность, логичность изложения материала, эмоциональная нейтральность, наличие специальной терминологии. Результаты анализа текстов статей представлены в таблице.

Результаты анализа текстов научно-технических статей
Results of the Analysis of Texts of Scientific and Technical Articles

Сфера функционирования и типовая ситуация общения	Наука и техника. Обмен информацией о решении научно-технических проблем и задач
Участники речи	Профессиональное сообщество ученых и инженеров. Специалисты, владеющие обширными знаниями о специализированной предметной области статьи
Функция речи	Профессиональная коммуникация специалистов разных специализированных предметных областей.
Тип содержания и предмет речи (тема)	Содержание – конкретное. Тема – результаты определенного исследования или обзор текущего состояния предмета исследования.
Коммуникативная цель	Обмен ясно, кратко и достоверно изложенной информацией о результатах исследования в некоторой предметной области
Стилеобразующие признаки	Ясность, точность, логичность, объективность и точность. Обезличенность информации. Шаблонность, четкое закрепление места за элементами.
Форма бытия	Письменная, в виде текстов установленной формы в научно-технических или специализированных периодических изданиях определенных предметных отраслей.

К ключевым элементам структуры научнотехнической статьи с точки зрения их функциональных и лексико-грамматических особенностей относят [9–10]:

1. Код УДК.
2. Название.

Отражает содержание статьи, обычно состоит не более одиннадцати слов без учета союзов и предлогов, возможны варианты названий, включающие от двух до пятнадцати слов.

3. Информация об авторах.

Включает имя и фамилию автора, место работы автора и контактные данные. Зачастую количество авторов не ограничено, но некоторые специализированные журналы могут устанавливать ограничения, например, не более пяти авторов. Место работы автора включает название организации с указанием города и страны, реже с полным адресом места работы автора. Контактные данные чаще всего представлены адресом электронной почты. Если авторов несколько, то указывается имя автора ответственного за переписку.

4. Аннотация и ключевые слова.

Стандартная аннотация содержит информацию об объекте, цели и методах исследования, основных результатах и выводах. Ключевые слова используют для быстрого поиска информации в больших коллекциях документов.

5. Введение.

Во введении акцентируют внимание на новизне и актуальности работы, приводят обзор литературы предметной области исследования. Элемент «Введение» зачастую заканчивается формулировкой цели работы.

6. Основная часть.

Способы организации данного раздела варьируют в зависимости от цели исследования, обычно выделяют разделы «Материалы и методы», «Обсуждение», «Результаты».

7. Заключение.

Отражает факты, сделанные на основе проведенной работы.

8. Слова благодарности.

Благодарят тех, кто оказал помощь в проведении работы.

9. Ссылки на литературу.

Представляет собой список литературы, процитированной в тексте статьи.

В результате анализа композиционной структуры текстов научно-технических статей, выявлено, что они имеют ярко-выраженную структуру, содержат определенный набор элементов, за каждым из которых закреплено свое место в тексте документа, взаимоувязанную систему заголовков разделов и подразделов.

2. Формальная модель текстов научно-технических статей

В качестве исходных данных выступают результаты композиционного анализа текстов стандартов, полученные в выше. Для решения задачи необходимо разработать формальные средства композиционной структуры научно-технических статей, использование которых позволит осуществлять структурную разметку в корпусе научно-технических текстов.

В результате будет получена модель формального представления текстов научно-технических статей, которая даст возможность при разметке корпуса научно-технических текстов учитывать их композиционную структуру.

Прежде всего, необходимо установить общую структуру языка стандартов как совокупности элементов. Структура совокупности выявлена в результате проведения тщательной классификации исследуемых единиц, то есть их иерархического распределения по определенному признаку на основные разделы, которые далее распадаются на подразделы, пункты, подпункты, требования, которые в свою очередь разбиваются на отдельные более мелкие базисные единицы – предложения. Все это вместе образует номенклатуру – полный подробный перечень отдельных элементов изучаемой совокупности [11].

Любую систему можно представить как некоторую совокупность взаимосвязанных элементов. Каждая из таких систем S_j является отделенной системой (научный / официально-деловой стиль) и может быть представлена как некоторая часть (подсистема) более общей системы S (суперсистема – русский язык) $S_j \in S$.

Взаимосвязь между системами S_j и S построена по иерархическому принципу, который предусматривает подчиненность подсистемы S_j суперсистеме S , в плане своего структурного расположения, так и в плане функционально-коммуникативной направленности составных частей. Отсюда вытекает, что любую систему (совокупность) S можно разделить на подсистемы разных рангов $S_1 - S_2 - S_3$ (определенный естественный язык – стиль – жанр), проводя процесс членения по определенным признакам до получения составных элементов. При этом каждая операция членения системы порождает отдельные подсистемы, что обеспечивает построение некоторого дерева метасистем S , на котором выделены отдельные подсистемы ($S_1; S_2; S_3$), которые относятся к разным уровням ($S_{12}; S_{121}; S_{1211}$). Членение системы может быть произведено рядом способов, при этом генерируется разное количество частей (подсистем, базисных элементов) [12].

Композиционная структура научно-технической статьи в первом приближении состоит из трех частей. Первая часть включает реферативный раздел, под которым понимается совокупность основных конвенциональных элементов, используемая для формальной идентификации первичного документа с учетом его природы, элементов и порядка внешних признаков, которые отличают его от других. Компонентами реферативного раздела являются: код УДК, название статьи, информация об авторах, место работы авторов, аннотация (реферат), ключевые слова. Код УДК не является обязательным элементом статьи. Реферативный раздел зачастую имеет переводную версию на английском языке. Вторая часть статьи представляет корпус научно-технической статьи, который обычно состоит из пяти основных структурных элементов: введение, материал и методы, результаты, обсуждение результатов, заключение. Стоит отметить, что названия разделов, кроме «введения» и «заключения» могут отличаться, однако их содержание соответствует по смыслу заявленным. В связи с тем, что разные научные журналы формируют собственные требования к рукописям статей, структурные элементы корпуса научно-технической статьи могут быть разбиты на пункты и подпункты, которые в свою очередь разбиваются на абзацы и затем предложения. Третья часть статьи является вспомогательным аппаратом публикации, который включает примечание и ссылки на источники [13–14].

В нотациях Бекуса-Наура композиционную структуру текстов научно-технических статей можно задать следующим образом:

$$St_i ::= \langle X^1, X^2, X^3 \rangle$$

где X^1 – реферативный раздел научно-технической статьи, X^2 – корпус научно-технической статьи, X^3 – информативный раздел научно-технической статьи.

X^1 – реферативный раздел научно-технической статьи, состоящий из следующих элементов:

$$X^1 ::= \langle x_{11}, x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17} \rangle | \langle x_{12}, x_{13}, x_{14}, x_{15}, x_{16}, x_{17} \rangle$$

где x_{11} – код УДК, x_{12} – название статьи, x_{13} – информация об авторах, x_{14} – место работы авторов, x_{15} – контактная информация авторов, x_{16} – аннотация, x_{17} – ключевые слова.

X^2 – корпус научно-технической статьи можно представить в виде набора из следующих элементов:

$$X^2 ::= \langle x_{21}, x_{22}, x_{23}, x_{24}, x_{25} \rangle | \langle x_{21}, x_{22}, x_{23}, x_{25} \rangle | \langle x_{21}, x_{23}, x_{25} \rangle | \langle x_{21}, x_{23}, x_{24}, x_{25} \rangle,$$

где x_{21} – введение, x_{22} – материал и методы, x_{23} – результаты, x_{24} – обсуждение результатов, x_{25} – заключение.

X^3 – информативный раздел научно-технической статьи, для которого справедливо

$$X^3 ::= \langle x_{31}, x_{32} \rangle | \langle x_{32} \rangle,$$

где x_{31} – примечания, x_{32} – ссылки на источники.

На рисунке представлена полученная структурная схема элементов текста научно-технической статьи.

На основе проведенного анализа композиционной структуры текстов научно-технических статей модель текста научно-технической статьи St целесообразно представить в виде:

$$St = \langle E^L, R \rangle$$

где E – структурный элемент, R – отношения между структурными элементами, L – уровень структурного элемента. При этом $L = \{l_1, \dots, l_5\}$, где l_1 – раздел, l_2 – пункт, l_3 – подпункт, l_4 – абзац, l_5 – предложение.



Рис. Структурные элементы текстов научно-технических статей
Fig. Structural elements of texts of scientific and technical articles

Такое представление научно-технической статьи порождает дальнейшую возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего исследуемого текста научно-технической статьи в целом. Таким образом, модель композиционной структуры текста научно-технической статьи – это граф, вершинами и ребрами которого являются только полноценные единицы – разделы, пункты, подпункты, то есть наиболее значимые структурные элементы. Наличие структурной разметки текста научно-технической статьи при создании корпуса научно-технических текстов значительно расширит исследовательский потенциал корпуса, что в свою очередь позволит при разработке систем обработки естественного языка учитывать композиционные особенности научно-технических текстов, в целом, и их отдельных структурных компонентов, в частности.

Заключение

В настоящее время для описания подязыка определенной предметной области необходимо использование специальных корпусов узкоспециальных текстов – корпусов научно-технических текстов. К источникам текстов для корпуса научно-технических текстов отнесены научно-технические статьи. Показано, что электронный корпус представляет собой коллекции

текстов и их разметку: морфологическую, синтаксическую, семантическую, метаразметку. Выявлено, научно-технические тексты обладают рядом специфических особенностей в композиционной структуре.

Научно-техническая статья – это первичный письменный жанр научного дискурса, задачей которого является постановка и решение одной научной проблемы, имеет средний объем, конвенциональную структуру, системы ссылок и выходные данные. Научно-техническим статьям присущи все стилевые особенности научного стиля: точность, логичность изложения материала, эмоциональная нейтральность, наличие специальной терминологии. К ключевым элементам структуры научно-технической статьи с точки зрения их функциональных и лексико-грамматических особенностей относят: название, информацию об авторах, аннотацию и ключевые слова, введение, основную часть, заключение, слова благодарности и ссылки на литературу.

Композиционная структура научно-технической статьи состоит из реферативного раздела, корпуса научно-технической статьи и информативного раздела. Компонентами реферативного раздела являются: код УДК, название статьи, информация об авторах, место работы авторов, аннотация (реферат), ключевые слова. Код УДК не является обязательным элементом статьи. Корпус научно-технической статьи состоит из «Введения», «Материалов и методов», «Результатов», «Обсуждения результатов», «Заключения». Информационный раздел включает примечание и ссылки на источники. Композиционная структура текстов научно-технических статей задана в нотациях Бекуса-Наура. Построена модель текста научно-технической статьи для структурной разметки в корпусе научно-технических текстов, которая порождает дальнейшую возможность анализировать с помощью математических методов как вероятностную, так и логическую структуру всего исследуемого текста научно-технической статьи в целом.

Список литературы

1. **Захаров В. П.** Корпуса русского языка // Труды института русского языка имени В. В. Виноградова. 2015. Т.6. С. 20–65.
2. **Кружков М. Г.** Информационные ресурсы контрастивных лингвистических исследований: электронные корпуса текстов // Системы и средства информатики. 2015. Т. 25, № 2. С. 140–159.
3. **Нагель О. В.** Корпусная лингвистика и ее использование в компьютеризированном языковом обучении // Язык и культура. 2008. № 4. С. 53–59.
4. **Соловьева А. Е.** Англоязычные тексты военной авиации как основа лингвистического корпуса // Балтийский гуманитарный журнал. 2019. № 3 (28). С. 369–372.
5. **Лесников В. С.** Виды разметок текстовых корпусов русского языка // Научно-техническая информация. Серия 2. Информационные процессы и системы. 2019. №9. С. 27–30.
6. **Ванюшкин А. С.** О разметке корпусов текстов ключевыми словами // Новые информационные технологии в автоматизированных системах. 2018. № 21. С. 207–211.
7. **Бутенко Ю. И.** Влияние лингвистических особенностей текстов стандартов на информационный поиск // Филологические науки. Научные доклады высшей школы. 2019. № 6. С. 29–35. DOI: 10.20339/PhS.6-19.029
8. **Попова Т. Г.** Структура испанской научно-технической статьи как первичного жанра научного дискурса // Вестник Российского университета дружбы народов. Серия: Русский и иностранные языки, и методика их преподавания. 2004. № 1. С. 108–115.
9. **Романов Д. А.** Кратко о структуре экспериментальной научной статьи на английском языке // Вестник Казанского технологического университета. 2014. Т. 17, № 6. С. 325–327.
10. **Раицкая Л. К.** Структура научной статьи по политологии и международным отношениям в контексте качества научной информации // Полис. Политические исследования. 2019. № 1. С. 167–181.

11. **Sidnyaev N. I.** Mathematical apparatus for engineering-linguistic models // AIP Conference Proceedings. 2019. Vol. 2195. No. 1. P. 020033. DOI: 10.1063/1.5140133
12. **Бутенко Ю. И.** Модель текста стандарта при информационном поиске в коллекции документов нормативной базы // Вестник компьютерных и информационных технологий. 2020. – Т. 17, № 11. С. 23–32. DOI: 10.14489/vkit. 2020.11
13. **Попова Н. Г.** Введение к научной статье на английском языке: структура и композиция // Высшее образование в России. 2015. № 6. С. 52–58.
14. **Иванов В. П.** Как написать научную статью (структура материала и организация работы) // Вестник Полоцкого государственного университета. Серия В. Промышленность. Прикладные науки. 2016. № 3. С. 195.

References

1. **Zakharov V. P.** Russian corpora. *Proceedings of Vinogradov Institute of the Russian Language*, 2015. Vol. 6, pp. 20–65. (in Russ.)
2. **Nagel O. V.** Corpus linguistics and its use in computerized language learning. *Language and Culture*, 2008. No. 4, pp. 53–59. (in Russ.)
3. **Kruzhkov M. G.** Information resources of contrastive linguistic research: electronic corpus of texts. *Systems and means of informatics*, 2015. Vol. 25, no. 2, pp. 140–159. (in Russ.)
4. **Lesnikov V. S.** Types of markup of text corpus of the Russian language. *Scientific and Technical Information. Series 2. Information processes and systems*, 2019. No. 9, pp. 27–30. (in Russ.)
5. **Butenko Iu. I.** Model of the text of the standard in the information search in the collection of documents of the normative base. *Bulletin of Computer and Information Technologies*, 2020. Vol. 17, no. 11, pp. 23–32. DOI: 10.14489/vkit. 2020.11 (in Russ.)
6. **Butenko Iu. I., Semenova E. L.** Influence of linguistic features of standards texts on information search. *Philological Sciences. Scientific reports of higher school*, 2019. No. 6, pp. 29–35. DOI: 10.20339/PhS.6-19.029 (in Russ.)
7. **Sidnyaev N. I., Butenko J. I., Garazha V. V.** Mathematical apparatus for engineering-linguistic models. AIP Conference Proceedings, 2019. Vol. 2195, no. 1, p. 020033. DOI: 10.1063/1.5140133
8. **Romanov D. A.** Briefly about the structure of the experimental scientific article in English. *Bulletin of Kazan Technological University*, 2014. Vol. 17, no. 6, pp. 325–327. (in Russ.)
9. **Raitskaya L. K.** Structure of a scientific article on political science and international relations in the context of the quality of scientific information. *Polis. Politicheskie issledovaniye*, 2019. No. 1, pp. 167–181. (in Russ.)
10. **Popova T. G.** Structure of the Spanish scientific and technical article as a primary genre of scientific discourse. *Bulletin of the Peoples' Friendship University of Russia. Series: Russian and foreign languages and the methodology of their teaching*, 2004. No. 1, pp. 108–115. (in Russ.)
11. **Popova N. G.** Introduction to the scientific article in English: structure and composition. *Higher Education in Russia*, 2015. No. 6, pp. 52–58. (in Russ.)
12. **Ivanov V. P.** How to write a scientific article (material structure and work organization). *Bulletin of Polotsk State University. Series B. Industry. Applied sciences*, 2016. No. 3, p. 195. (in Russ.)
13. **Vanyushkin A. S., Grashchenko L. A.** On the markup of corpus texts with keywords. *New Information Technologies in Automated Systems*, 2018. No. 21, pp. 207–211. (in Russ.)
14. **Solov'eva A. E.** English-language texts of military aviation as the basis of linguistic corpus. *Baltic humanitarian journal*, 2019. No. 3(28), pp. 369–372. (in Russ.)

Информация об авторе

Бутенко Юлия Ивановна, кандидат технических наук, доцент кафедры «Романо-германские языки», Московский государственный технический университет им. Н. Э. Баумана (Москва, Россия)

Information about the author

Butenko I. Yulia, Candidate of Technical Sciences, Associate Professor of the Department of Romano-Germanic Languages, Bauman Moscow State Technical University (Moscow, Russian Federation)

*Статья поступила в редакцию 27.07.2022;
одобрена после рецензирования 07.11.2022; принята к публикации 07.11.2022
The article was submitted 27.07.2022;
approved after reviewing 07.11.2022; accepted for publication 07.11.2022*

Научная статья

УДК 004.05+519.718.2

DOI 10.25205/1818-7900-2022-20-3-14-28

Отладочный комплекс с имитацией кондуктивных помех для тестирования контроллеров АСУ ТП

**Владимир Владимирович Гаркуша¹, Сергей Сергеевич Журавлев²,
Станислав Рудольфович Шакиров³, Владимир Владимирович Яковлев⁴**

^{1,2,4}Федеральный исследовательский центр информационных и вычислительных технологий
Новосибирск, Россия

³Конструкторско-технологический институт научного приборостроения
Сибирского отделения Российской академии наук
Новосибирск, Россия

¹vgar49@mail.ru

²s-zhur@yandex.ru

³shakirov@tdisie.nsc.ru

⁴vy82@mail.ru

Аннотация

В работе представлен полнофункциональный имитационный отладочный комплекс, предназначенный для тестирования программируемых контроллеров АСУ ТП в лабораторных условиях, но максимально приближенных к реальным условиям эксплуатации на объекте, с учетом большой зашумленности входных сигналов, воздействия на них сетевой и импульсных помех большой интенсивности, а также с учетом вариации параметров линии связи. Для имитации наиболее характерных помех разработана структура формирователя кондуктивных помех и параметров линии связи с описанием всех его элементов.

Ключевые слова

имитационный отладочный комплекс, электромагнитная помеха, формирователь кондуктивных помех, имитатор параметров линии связи, адаптер-формирователь помехи

Для цитирования

Гаркуша В. В., Журавлев С. С., Шакиров С. Р., Яковлев В. В. Отладочный комплекс с имитацией кондуктивных помех для тестирования контроллеров АСУ ТП // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 14–28. DOI 10.25205/1818-7900-2022-20-3-14-28

© Гаркуша В. В., Журавлев С. С., Шакиров С. Р., Яковлев В. В., 2022

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)

Вестник НГУ. Серия: Информационные технологии. 2022. Том 20, № 3

Vestnik NSU. Series: Information Technologies, 2022, vol. 20, no. 3

Debugging Complex with Simulation of Conductive Interference for Testing Automated Control System Controllers

Vladimir V. Garkusha¹, Sergey S. Zhuravlev², Stanislav R. Shakirov³,
Vladimir V. Yakovlev⁴

^{1,2,4}Federal Research Center for Information and Computational Technologies
Novosibirsk, Russian Federation

³Technological Design Institute of Scientific Instrument Engineering of Siberian Branch of RAS
Novosibirsk, Russian Federation

¹vgar49@mail.ru

²s-zhur@yandex.ru

³shakirov@tdisie.nsc.ru

⁴vy82@mail.ru

Abstract

The paper presents a full-featured simulation debugging system designed to test programmable controllers for automated process control systems in the laboratory. This system imitates the operating conditions of the automated process control system, as close as possible to the real operating conditions at the automation object. The presented system can form various levels of interference of input signals, influencing the automatic process control system with high-intensity network and impulse noise. The system also allows one to vary the parameters of the communication line. The structure of the conducted interference generator and the variator of the communication line parameters have been developed. A description of all elements necessary for modeling the main types of interference is given. In this work, an analysis of the existing electromagnetic interference in the signal circuits of process control systems was carried out, and the most typical interference was identified. A block diagram of a conducted interference generator and a variator of communication line parameters has been developed. Both functional blocks are part of the modeling and debugging complex. They allow one to simulate the interference environment for the controller under test by introducing the generated interference into the signal communication lines in a conductive way. The structure of the adapter-former of interference for analog signals is worked out in detail with a description of its main components. Recommendations for choosing the element base are given. The practical significance of the performed work lies in the fact that it may improve the efficiency of the complex of control and laboratory tests of the systems being created. This allows one to achieve a reduction in complexity and in setup time during implementation at the automation facility.

Keywords

simulation and debugging complex, electromagnetic interference, conducted interference shaper, communication line parameters simulator, interference shaper adapter

For citation

Garkusha V. V., Zhuravlev S. S., Shakirov S. R., Yakovlev V. V. Debugging Complex with Simulation of Conductive Interference for Testing Automated Control System Controllers. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 14–28. DOI 10.25205/1818-7900-2022-20-3-14-28

Введение

В ФИЦ ИВТ (ранее КТИ ВТ СО РАН и ИВТ СО РАН) много лет проводятся работы по созданию промышленных программируемых контроллеров и АСУ ТП различного назначения на их основе. Разработаны и введены в промышленную эксплуатацию несколько автоматизированных систем, в том числе: АСУ ТП турбокомпрессорной станции и АСУ ТП тепловой станции № 1 СО РАН [1; 2], несколько АСУ ТП для угольных шахт Кузбасса (контроля и управления ленточными конвейерами; наблюдения, оповещения и поиска персонала при авариях; водоотлива; вентиляции и др.) [3; 4]. Работы по внедрению АСУ ТП всегда были сопряжены с большим объемом отладочных работ как при производстве контроллеров и систем, так и с еще большим объемом и трудоемкостью при выполнении пуско-наладочных работ на объектах автоматизации, поэтому в ФИЦ ИВТ всегда большое внимание уделялось средствам отладки систем при их изготовлении, т. е. созданию специализированных отладочных стендов [5].

Как показал опыт внедрения АСУ ТП шахтного и рудничного исполнения, пуско-наладочные работы на месте их монтажа и будущей эксплуатации существенно затруднены как спо-

собами и местами размещения оборудования и кабельных трасс, так и рабочими условиями эксплуатации. При этом, даже когда каждое устройство или рабочая программа системы в отдельности прошли все стадии технического контроля и корректно выполняли свои функции на стадии наладки при производстве, то собранная из них система на месте эксплуатации может иметь в себе новые уязвимости. Поэтому так важно осуществлять многофакторную верификацию и отладку программно-аппаратных средств АСУ ТП на этапе ее производства или модернизации.

Для реализации этих целей в ФИЦ ИВТ ведутся работы по созданию имитационного отладочного комплекса с функцией автоматической верификации алгоритмов управления АСУ ТП специального назначения, а также ее компонент, с использованием моделей технологических процессов угольной шахты.

Существующий в настоящее время отладочный комплекс позволяет проводить тестовые испытания контроллеров АСУ ТП с имитационными сигналами и реальными сигналами без помех, при которых выявляется большинство программных ошибок [6]. Для этих контроллеров, в качестве имитаторов входных реальных сигналов и исполнительных устройств для выходных управляющих сигналов, в отладочном комплексе используются промышленные модули серийного производства и устройства собственной разработки.

Очевидно, что предварительные испытания АСУ ТП, проводимые в лабораторных условиях, и ее эксплуатация в реальных условиях отличаются по степени воздействия внешних факторов. О двух из них и пойдет речь в данной работе:

- во-первых, это электромагнитная обстановка, влияющая на виды и уровни помех, индуцируемых в сигнальных цепях сбора данных и управления;
- во-вторых, это характеристики кабельных линий связи, которые в лабораторных условиях практически идеальны, хотя в реальных условиях эксплуатации такие линии могут достигать длины от сотен метров до нескольких километров, при этом параметры линий могут варьироваться в зависимости от качества используемой кабельной продукции и способа прокладки сигнальных линий.

Чтобы учесть эти факторы при проведении предварительных и тестовых испытаний системы необходимо в существующий имитационный отладочный комплекс добавить дополнительный блок – блок формирования помех и параметров линий связи, управляемый от АРМ тестировщика, что даст возможность тестирования системы на отладочном комплексе, как с целью проверки ее работоспособности с идеальными первичными имитационными сигналами, так и с целью исследования надежности функционирования ее оборудования и алгоритмов, при воздействии вышеуказанных факторов, в условиях разной помеховой обстановки.

Таким образом, целью настоящей работы является разработка структурной схемы полнофункционального имитационного отладочного комплекса. Данный комплекс предназначен для тестирования автоматизированных систем и промышленных программируемых контроллеров в лабораторных условиях, но максимально приближенных к реальным условиям эксплуатации их на объекте. К таким условиям относятся – большая зашумленность сигналов от датчиков, воздействиями сетевой помехи (50 Гц) и импульсных помех большой интенсивности, а также с вариациями параметров линии связи.

Практически все производители оборудования для облегчения процесса его наладки и тестирования, так или иначе всегда сталкиваются с проблемой создания отладочных комплексов разной степени сложности, функциональности и назначения в зависимости от глубины и объема проводимых тестовых испытаний. Речь не идет о специализированных стендах, предназначенных для проведения в обязательном порядке приемо-сдаточных испытаний выпускаемого оборудования в соответствии с его техническими условиями, в которых приводится и схема стенда, и используемые в нем, поверенные измерительные приборы и дополнительно привлеченное оборудование. Это отдельная задача и решается она индивидуально для каждого конкретного изделия в зависимости от его назначения, при этом могут проводиться испытания

на электробезопасность, на соответствие заявленным параметрам, могут быть затронуты вопросы метрологии измерительных каналов и т. д.

Тема создания отладочных стендов не нова, существует достаточно много публикаций по данному вопросу, в которых приводятся описания различных стендов, в том числе виртуальных, испытательного и измерительного оборудования, средств генерации электромагнитных помех (ЭМП) и их измерений, тестового программного обеспечения, и все это разрабатывалось и создавалось как для частного применения [7–10], так и для создания инструментальных средств разработки комплексов (Стенды NI, комплексы на базе xPC Target MathWorks и другие).

В данной работе рассматривается более широкая задача – создание имитационного отладочного комплекса, на котором необходимо выявлять с помощью разноплановых тестов большинство аппаратных, алгоритмических и программных ошибок, допущенных при проектировании и изготовлении программируемых контроллеров и АСУ ТП, перед внедрением их на объекте автоматизации.

Классификация помех

Любой функциональный узел АСУ ТП, будь то управляющий программируемый контроллер с блоками сбора данных и управления, датчики и исполнительные механизмы, линии передачи сигналов и электропитания или программное обеспечение, невозможно полностью оградить от воздействия внешних факторов, каким-либо образом влияющих на качество их функционирования. Поэтому одной из важнейших задач при проектировании систем автоматизации является обеспечение их работоспособности с учетом всех факторов внешнего воздействия, характерных для места размещения и эксплуатации технических средств. Кроме учета таких факторов как: влияние окружающей среды (температура, влажность, давление, осадки, агрессивные среды); механические удары и вибрации; воздействие человека – особое внимание необходимо уделять ЭМП. Влиянию этих помех более всего подвержены электронные блоки устройств АСУ ТП и линии связи, а также алгоритмы замкнутого контура управления.

Вопросам возникновения ЭМП на объектах автоматизации, воздействия их на оборудование, как и способ защиты последнего от воздействия помех, обеспечения помехоустойчивости управляющих систем и электромагнитной совместимости используемой техники, в литературе уделено много внимания [11–13].

В общем случае, ЭМП считается любое электромагнитное явление, которое сможет ухудшить качество функционирования технического средства.

Определение электромагнитной обстановки, то есть выделение ограниченного набора видов ЭМП и их характеристик, наиболее значимо влияющих на устройство в условиях его эксплуатации – является задачей экспертной, а значит субъективной. В национальном стандарте ГОСТ Р 51317.2.5-2000 представлена достаточно подробная классификация электромагнитных явлений и процессов и связанных с ними ЭМП, при этом, предложено для конкретного технического средства и конкретной электромагнитной обстановки, а также места эксплуатации выбрать свой ограниченный перечень ЭМП, достаточно полно (по мнению эксперта) описывающий условия эксплуатации этого технического средства.

Согласно указанного ГОСТ ЭМП подразделяются на излучаемые и кондуктивные. Излучаемые ЭМП возникают и распространяются в непроводящей среде, в то время как кондуктивные возникают и передаются в токопроводящих средах. Эти помехи тесно связаны между собой и на протяжении всего пути распространения могут переходить от одного вида к другому, пока не достигнут своего реципиента. ЭМП может проникать в техническое средство через разные порты [12]: корпус; линии питания (переменного или постоянного тока); линии ввода-вывода; заземление.

Для иллюстрации шумов, приходящих на вход АСУ ТП, на рис. 1 приведен реальный исходный сигнал от термометра сопротивления ТСМ50 (температура стали статора мощного

двигателя), снятый авторами при помощи осциллографа, из которого видно, что на основную наводку от промышленной сети 50 Гц дополнительно накладывается еще широкополосный шум до 10 МГц (похожий на «белый шум») и хаотические импульсные помехи большой амплитудой.

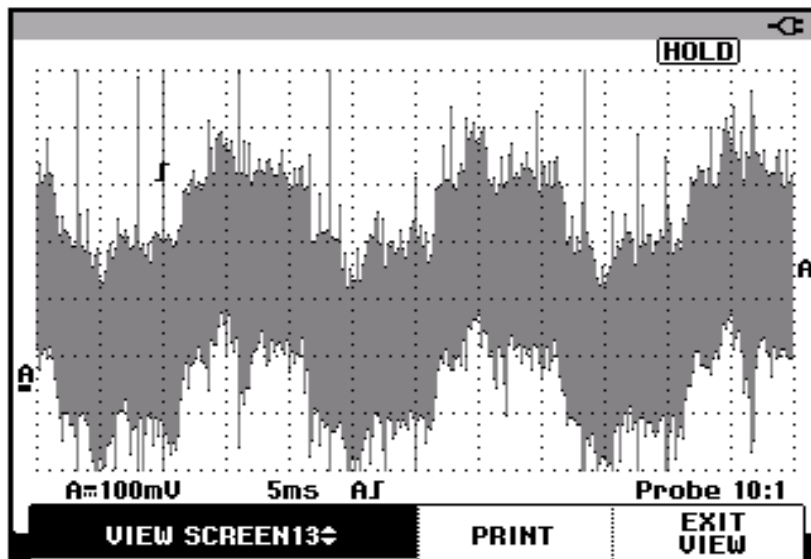


Рис. 1. Исходный сигнал температуры стали статора двигателя

Fig. 1. The initial temperature signal of the engine stator steel

Отношение сигнал/шум ($U_c/U_{ш}$) в этом сигнале составляет величину порядка 1, что наложало достаточно серьезные требования к входным портам контроллера АСУ ТП, принимающим такие зашумленные сигналы. Для достижения заданной абсолютной погрешности измерения температур статора во всем диапазоне, для обработки таких сигналов в программируемом контроллере авторами была применена как аппаратная, так и программная фильтрация. Кроме аппаратных RC фильтров на входах модулей был введен программный параметрический алгоритм реализации фильтра нижних частот с регулируемой частотой среза в диапазоне от 0,1 до 10 Гц.

В любой системе автоматизации используется большое количество различных электро-механических контактов, главным недостатком которых является произвольные повторные коммутации после срабатывания контактной пары. Это явление и стали называть – дребезгом контактов и борются с ним с самого начала, как только появились элементы автоматизированных систем. В зависимости от размеров, массы, материала и конструкции контактной системы время дребезга (время от первого соприкосновения контактов до затухания механических колебаний и установления стабильного контактирования) составляет величину от 0,5 до 2 мс у миниатюрных герконов и до сотен мс у мощных контакторов.

Практически вся цифровая электроника воспринимает дребезг как целую череду сигналов. Эти ложные коды вполне могут привести к неправильной работе и даже выходу из строя чувствительной электроники. Опять же исходя из собственной практики, измеренные значения времени дребезга контактов конечных выключателей электрифицированных задвижек в реальной системе составили величину не более 10 мс.

Еще одна существенная наводка присутствует практически во всех сигналах АСУ ТП – сетевая помеха 50 Гц: синфазная, достигающая величины десятков и даже сотен вольт на длин-

ных линиях связи и противофазная – небольшой амплитуды, возникающая от несимметричности линии связи. Эта помеха оказывает влияние в основном только на аналоговые сигналы.

Параметры линии связи (сопротивление, индуктивность и емкость – RLC) не являются случайными, как в случае с ЭМП, и определяются при проектировании системы в соответствии с установленными техническими условиями применения аппаратных средств, используемых при построении АСУ ТП. Однако при тиражировании таких систем, использовании кабельной продукции различной номенклатуры, способов прокладки кабелей, выполнении электрических соединений разными способами и с разным качеством на практике приводит к девиации параметров линий связи.

Виды имитируемых воздействий

Types of Simulated Impacts

	Типы портов	2-х проводные			3-х– 4-х	4-х– 6-и	Цепи питания		Порты цифровой связи (RS-485)	Порты заземления
		Дискретный (12В/сух. конт)	Сигнальный (напряжение, ток, сопротивления, термодвойной)	Частотный (12В, 0-60 Гц, 0-200Гц):	Сигнальный (Термометр-сопротивления)	Сигнальный (тензопреобразователь)	1-фазные 230В 50Гц	Постоянного тока 12В/24В		
Виды имитируемых воздействий:										
Провалы, кратковременные прерывания сигнала		+					+	+		
Дребезг контакта		+		+						
Импульсные помехи и «розовый шум»		+	+	+	+	+	+	+	+	
Гармонические помехи 50 Гц синфазные			+		+	+			+	+
Гармонические помехи 50 Гц дифференциальные			+		+	+				
Отклонение напряжения питания							+	+		
Положительные выбросы напряжения питания							+	+		
Изменение частоты питающего напряжения							+			
Провалы напряжения сигнала в пределах допуска									+	
Логически искажения целостности пакета данных									+	
Электростатические разряды										+
Параметры линии связи (RLC)		+	+	+	+	+	+	+	+	

Обобщая все приведенное выше и исходя из опыта работы по созданию АСУ ТП, а также с учетом экспертных оценок, из ГОСТ Р 51317.2.5-2000 авторами были выбраны наиболее характерные для различных портов технических средств АСУ ТП виды ЭМП и дополнены

другими видами искажений сигналов, необходимых для исследования работоспособности аппаратной и программной частей АСУ ТП (см. табл.).

Эти ЭМП и было решено реализовать в проектируемом имитационном отладочном комплексе путем их генерации и возможностью введения в порты ввода-вывода тестируемых технических средств кондуктивным способом с учетом вариации параметров линий связи (RLC).

Структурная схема имитационного отладочного комплекса

Для выполнения задачи по доработке существующего имитационного отладочного комплекса был разработан дополнительный блок, который вошел в состав отладочного комплекса и стал элементом его структуры. Блок получил название – формирователь кондуктивных помех и параметров линии связи (ФКП). Формирователь достаточно просто вписывается в структуру существующего отладочного комплекса и подключается между имитатором первичных сигналов и тестируемым контроллером АСУ ТП. На практике это может производиться, например, простой перекоммутацией сигналов на клеммных колодках отладочного комплекса.

На рис. 2 приведена усовершенствованная структурная схема имитационного отладочного комплекса с ФКП в своем составе, предназначенного для полнофункционального тестирования контроллеров АСУ ТП с возможностью моделирования и имитирования в лабораторных условиях параметров линий связи и помеховой обстановки, с максимальным приближением к тем условиям, в которых в дальнейшем на объекте будет эксплуатироваться действующий образец испытываемой системы.

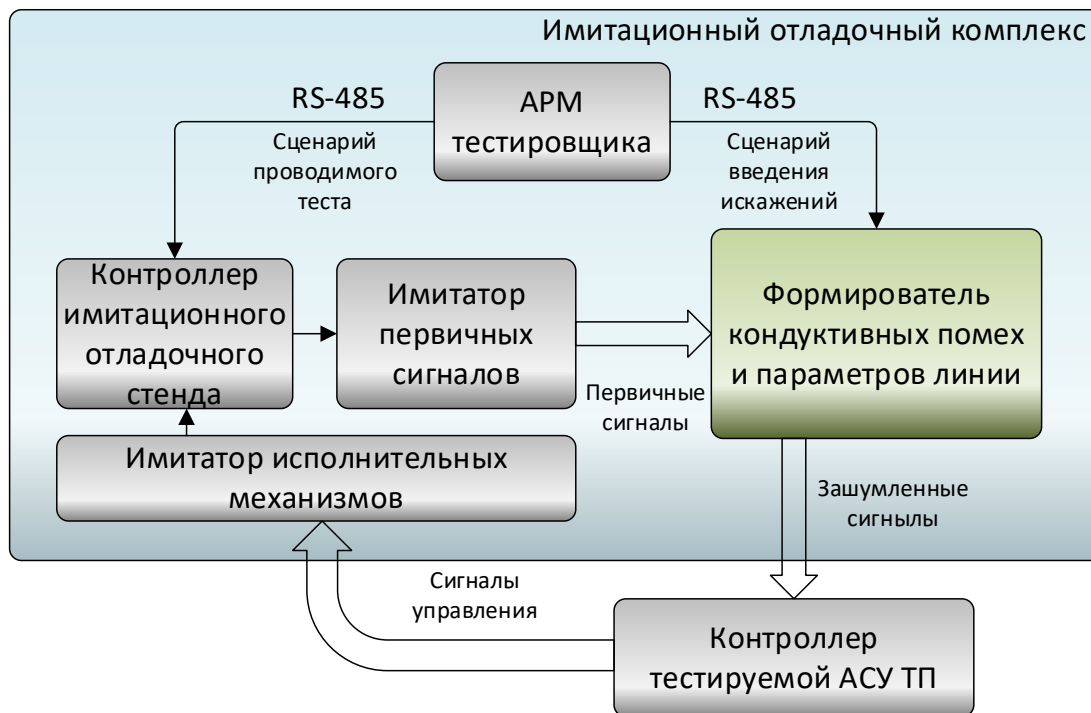


Рис. 2. Усовершенствованная структурная схема имитационного отладочного комплекса

Fig. 2. Improved block diagram of the simulation debugging complex

Формирователь кондуктивных помех и параметров линии

Формирователь кондуктивных помех и параметров линии предназначен для формирования, наиболее характерных помех, согласно приведенной выше таблице, наложения их на фор-

мируемые отладочным комплексом первичные сигналы с разными весовыми коэффициентами, задаваемыми от АРМ тестировщика, а также для имитации влияния параметров линий связи на первичные сигналы, т. е. создания на входе тестируемого контроллера АСУ ТП зашумленных сигналов. ФКП представляет собой функционально законченный блок, имеющий в своем составе набор имитаторов всех видов выбранных помех, для всех первичных сигналов, генерируемых в отладочном комплексе. ФКП подключается в разрыв цепей ввода-вывода тестируемого контроллера АСУ ТП в имитационном отладочном комплексе и управляется по линии интерфейса RS-485 от АРМ тестировщика (рис. 2; 3). ФКП имеет в своем составе полнофункциональный контроллер с индикатором и клавиатурой, в который от АРМ тестировщика загружаются заранее выбранные и установленные виды помех, а также их уровни, для каждого генерируемого комплексом сигнала и для каждого теста. Ввиду того, что параметры линий связи и виды формируемых шумов могут быть различны для каждого первичного сигнала, было принято решение для каждого типа сигнала реализовать отдельный модуль для формирования помех и имитации параметров линии - адаптер-формирователь помех (АФП). По сути дела, АФП предназначены для обеспечения аддитивного сложения формируемых в них помех с различными типами первичных сигналов: дискретных, тока, напряжения, сопротивления, термодар, сигналов тензомостовых датчиков, цифровых линий связи, питания и др. и, таким образом, на входе тестируемого контроллера получается суперпозиция чистого сигнала и всех имитируемых помех отладочного комплекса.

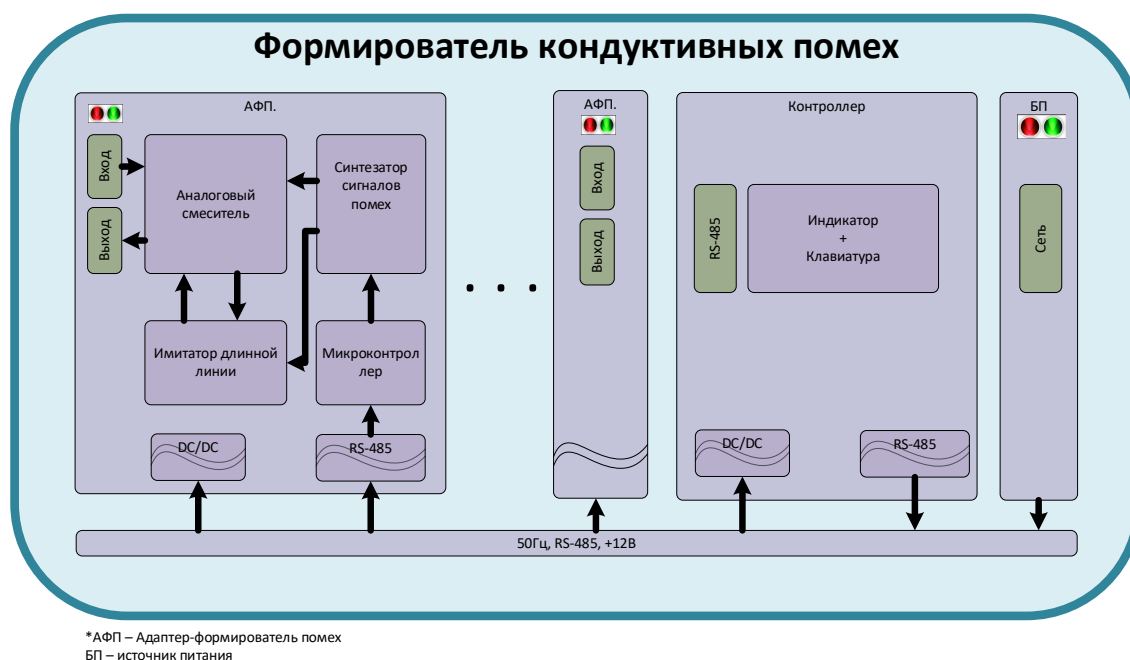


Рис. 3. Структурная схема формирователя кондуктивных помех и параметров линии

Fig. 3. Block diagram of the conductive interference shaper and line parameters

По внутренней шине ФКП RS-485 данные о параметрах формируемых помех для каждого сигнала передаются в микроконтроллеры АФП, которые устанавливают режимы работы синтезатора сигналов помех и имитатора длинной линии. Исходную сетевую помеху 50 Гц можно взять от источника питания ФКП или сгенерировать программно в его контроллере и пустить по шине. Для каждого первичного сигнала отладочного комплекса в АФП программно можно установить свой уровень каждой помехи и свои параметры линии связи.

Таким образом, конфигурация ФКП необходимым набором АФП зависит от количества и номенклатуры входов тестируемого контроллера.

Клавиатура и индикатор контроллера ФКП служат дополнительно в качестве резервной системы управления. При отключенном АРМ или его отсутствии, в ручном режиме работы комплекса, с клавиатуры контроллера ФКП в адаптерах-формирователях помех можно устанавливать и изменять необходимые уровни шумов, сетевой и импульсных помех, а также параметров линий связи для каждого канала тестируемого контроллера с визуальным их контролем на индикаторе или осциллографе.

Конструктивно ФКП может быть выполнен, как один из вариантов исполнения, в виде отдельного блока стандарта 3 U Евромеханика, с шинами питания и интерфейса RS-485, через которые осуществляется питание и управление АФП. АФП унифицированы по габаритным и установочным размерам и могут устанавливаться в качестве модулей в любую позицию блока, затем их типы и местоположение в блоке прописывается в контроллере.

АФП в своем составе имеет:

- блок входных контактов (соединитель) для подключения первичных сигналов (Вход);
- блок выходных контактов (соединитель) для подключения сформированных сигналов к тестируемому контроллеру АСУ ТП (выход);
- микроконтроллер;
- программируемый синтезатор сигналов помех, формирующий сигналы установленной интенсивности «розового шума» и случайной последовательности импульсных помех;
- имитатор длинной линии;
- аналоговый смеситель, с помощью которого сформированные помехи и параметры линии связи подключаются к линиям первичных сигналов;
- последовательный интерфейс RS-485;
- вторичный блок питания.

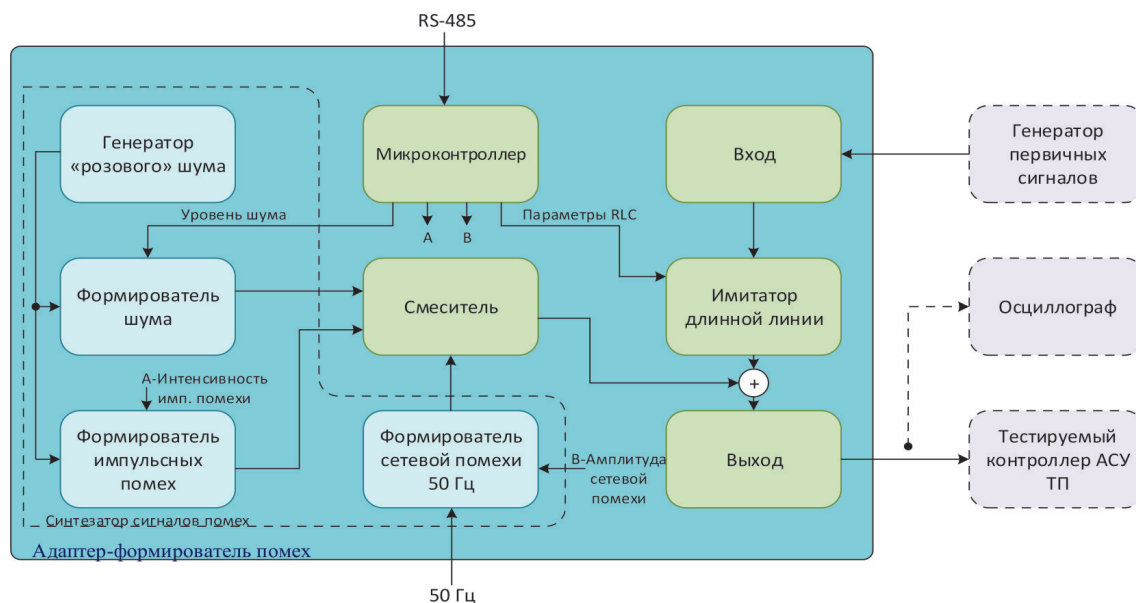


Рис. 4. Структурная схема АФП для аналогового сигнала

Fig. 4. Block diagram of an adapter interference shaper for an analog signal

Перечень необходимых для разработки АФП определялся исходя из реальных сигналов, внедренных на объектах АСУ ТП, и типов первичных сигналов, генерируемых в имитационном отладочном комплексе:

- АФП в дискретных линиях передачи сигналов, с напряжением 12 В постоянного тока;
- АФП в линиях передачи токового сигнала 4-20 мА;
- АФП в линиях передачи напряжения 0-1 В;
- АФП в линиях передачи сигналов от термометров-сопротивления;
- АФП в линиях передачи сигналов от тензопреобразователей (тензомостов);
- АФП в линиях передачи сигналов от термопар;
- АФП в линиях частотных сигналов напряжением 12В;
- АФП в линиях RS-485.

На рис. 4 приведена структурная схема АФП для аналогового сигнала с более подробным представлением синтезатора сигналов помех (ССП). Программируемый синтезатор сигналов помех предназначен для формирования широкополосного шума, сетевой наводки 50 Гц и импульсных помех случайной последовательности. Для каждого первичного сигнала отладочно-го комплекса в АФП можно программно установить и зафиксировать индивидуальный уровень помех для каждого теста.

Основой ССП является генератор «розового шума» – «белого шума» с ограниченным спектром. В природе процессов в чистом виде «белого шума» не существует, но при определенных предположениях реальные помехи можно описать моделью «белого шума». Ввиду того, что автокорреляционная функция «белого шума» представляет собой дельта – функцию, «белый шум» часто называют дельта-коррелированным случайным процессом. Отсутствие корреляции между любыми, сколь угодно близкими значениями «белого шума» означает бесконечно большую скорость изменения процесса, т. е. энергетический спектр «белого шума» является равномерным в диапазоне частот от 0 до ∞ , таким образом и средняя мощность «белого шума» также не ограничена. На практике же спектр «белого шума» на входах информационно-измерительных систем, при проведении их тестирования, можно смело ограничить верхней частотой в единицы МГц, т. к. входные каналы систем имеют ограниченную полосу пропускания, как за счет применяемой элементной базы, так и за счет схемотехнических решений.

В качестве источника «белого шума» в полосе частот от 20 Гц до 1 МГц, (генератора «розового шума») целесообразно использовать специализированный стабилитрон – кремниевый генератор шума КГ401В, имеющий спектральную плотность напряжения шума, при постоянном рабочем токе 50 мкА, не менее $30 \text{ мкВ}/\sqrt{\text{Гц}}$, т. е. в указанной полосе частот напряжение шума на выходе генератора будет порядка 30 мВ. Для генерации шумов в более широкой полосе спектра – от 20 Гц до 2,5 МГц, применяют КГ401А, правда при этом напряжение шума на выходе будет ниже – около 10 мВ. Кроме этого специализированного прибора в литературе есть рекомендации к применению и некоторых обычных стабилитронов: КС168; Д814В; 1N759, которые апробированы в реальных схемах и показали неплохие результаты в качестве генераторов шума при работе с микротоками.

Формирователь шума – это широкополосный и быстродействующий усилитель с высоким входным сопротивлением (не менее 10 МОм), чтобы не шунтировать генератор «розового шума», и регулируемым коэффициентом усиления, управляемым от микроконтроллера, для установки необходимого для данного теста уровня шума. Важными требованиями к усилителю, необходимыми для передачи дельта- функций шума, являются: входное сопротивление – не менее 10 МОм; полоса единичного усиления – не менее 15 МГц; скорость нарастания выходного напряжения, определяемая из выражения $V = 6,28 * U_{\text{вых.пик}} * F_{\text{макс.}}$ – не менее 15,7 В/мкс ($U_{\text{вых.пик}} = 1 \text{ В}$, $F_{\text{макс.}} = 2,5 \text{ МГц}$); время установления выходного напряжения – не более 400 нс. Среди операционных усилителей, широко представленных сегодня на рынке, для этих целей наибольшего внимания заслуживает операционный усилитель КР544УД2АТ (НПО «Восток»), с характеристиками, полностью удовлетворяющими нашим требованиям. Регулируемый коэффициент передачи формирователя реализуется использованием в обратной связи цифрового потенциометра с I2C интерфейсом, например, 256-позиционного AD5245 от Analog Devices (цифрового переменного резистора MCP41XXX с интерфейсом SPI компа-

нии Microschip или AD511X, DS18XX, MAX54XX). Этот прибор выполняет те же функции, что и механические потенциометры или переменные резисторы, но имеет более качественные параметры типа разрешающей способности, постоянство положения движка и температурного коэффициента сопротивления. Установка значения сопротивления производится микроконтроллером через I²C интерфейс, который позволяет как считывать, так и записывать данные.

Формирователь импульсной помехи – это компаратор (дифференциальный усилитель с большим коэффициентом усиления) с регулируемым от микроконтроллера порогом. На один вход поступает «розовый шум» с генератора, на второй вход – напряжение порога. В зависимости от уровня выставленного порога на выход формирователя проходят только верхняя часть «розового шума»: чем ниже порог, тем большая часть шумов проходит на выход и значит тем больше на выходе интенсивность импульсной помехи, т. е. выше хаотически повторяющаяся частота ее появления. При высоких порогах на выходе формирователя можно добиться крайне редкого появления импульсных помех. В качестве операционного усилителя опять же можно использовать КР544УД2АТ (К154УД4), регулировку порога достаточно просто осуществить при помощи цифрового потенциометра с I²C интерфейсом AD5245.

Формирователь сетевой помехи 50 Гц – усилитель с регулируемым коэффициентом усиления выполняется по аналогии со схемой формирователя шума и на тех же элементах.

На смесителе проще и целесообразнее смешивать только сетевую помеху 50 Гц и шумы с формирователя шума, при этом на выходе смесителя будет сигнал схожий с реальным сигналом (см. рис. 1). Этот сигнал *кондуктивным* способом вводится в линию с первичным аналоговым сигналом отладочного комплекса. Таким же способом в линию вводится и импульсная помеха. В результате такого преобразования на нагрузке входного порта тестируемого контроллера будет суперпозиция всех имитируемых отладочным комплексом сигналов.

Чтобы имитировать линию связи разной длины от источника сигнала до входа контроллера, необходимо задавать разные значения параметров R , L , C в зависимости от длины линии. Зависимость здесь прямая – чем более длинная линия, тем большие значения принимают ее параметры. По приведенным в литературе сведениям, погонная емкость межблочных кабелей определяется конструкцией, размерами и свойствами диэлектрика, и лежит в пределах 10...100 пФ/м, погонная индуктивность кабеля зависит от его геометрии и тоже невелика, речь идет о величинах 0,1...1 мкГн/м. Омическое сопротивление проводника определяется его сечением и материалом, из которого он изготовлен. Например, один метр медной проволоки сечением 1 мм² имеет сопротивление 0,0172 Ом.

В качестве ориентира для расчета параметров имитатора линии возьмем широко применяемый в АСУ ТП кабель ВВГнг-Ls – 0,5 мм², который и авторы использовали в реальных системах в качестве кабелей питания. Для этого кабеля в литературе приводятся измеренные и вычисленные значения погонной емкости и индуктивности, которые составляют 100 пФ/м и 0,54 мкГн/м соответственно. Сопротивление постоянному току двух жил пары – примерно 70 Ом/км. Для кабеля марки ТПП, который также часто используется в АСУ ТП в качестве сигнальных кабелей, значения погонной емкости составляют также от 50 до 60 нФ/км, т.е. величины того же порядка, что и для кабеля ВВГнг, поэтому его данные и примем для разработки имитатора длинной линии. Для упрощения задачи остановимся на четырех вариантах длины линии и ее параметрах, что для оценки ее влияния на генерируемый сигнал и на вход тестируемого контроллера будет достаточно.

Вариант 1: 0 м; $R = 0$ Ом; $L = 0$ мкГн; $C = 0$ нФ.

Вариант 2: 100 м; $R = 7,0$ Ом; $L = 54$ мкГн; $C = 1$ нФ.

Вариант 3: 500 м; $R = 35,0$ Ом; $L = 270$ мкГн; $C = 5$ нФ.

Вариант 4: 1000 м $R = 70,0$ Ом; $L = 540$ мкГн; $C = 10$ нФ.

На рис. 5 приведена эквивалентная схема имитатора длинной линии для реализации всех четырех его вариантов с разными значениями R , L , C , устанавливаемыми от микроконтроллера, который управляет реле, аналоговыми ключами или коммутаторами, подключая нужные

элементы к линии. При выборе элементов коммутации необходимо обращать внимание на их сопротивление в открытом состоянии, которое возможно придется учитывать в общем омическом сопротивлении линии. Параметры линии R , L , C выбираются из номинальных рядов соответствующих элементов, при этом нужно иметь в виду, что при применении в качестве индуктивностей дросселей типа ДМ, необходимо опять же учитывать их омическое сопротивление постоянному току, которое согласно справочным данным равно: 3,8 Ом для ДМ-0,1-56 мкГн; 5,5 Ом для ДМ-0,1-280 мкГн; 15 Ом для ДМ-0,1-500 мкГн.

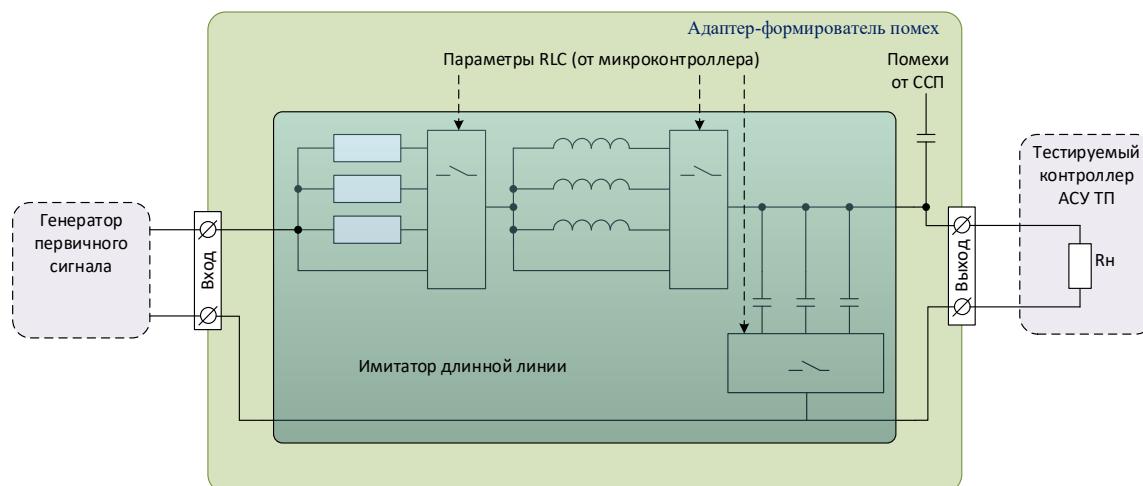


Рис. 5. Имитатор длинной линии в структуре АФП
Fig. 5. Long line simulator in the an adapter interference shaper structure

Заключение

В работе представлена структурная схема полнофункционального имитационного отладочного комплекса, предназначенного для тестирования программируемых контроллеров АСУ ТП в лабораторных условиях, но максимально приближенных к реальным условиям эксплуатации на объекте, с учетом большой зашумленности сигналов от датчиков, воздействия сетевой помехи (50 Гц) и импульсных помех большой интенсивности, а также с учетом вариации параметров линии связи.

Проведен анализ существующих электромагнитных помех в сигналах АСУ ТП, выделены наиболее характерные помехи, для имитации которых разработана структурная схема ФКП, входящего в состав имитационного отладочного комплекса, и позволяющего имитировать помеховую обстановку для тестируемого контроллера путем ввода формируемых помех в сигнальные линии связи кондуктивным способом. Подробно проработана структура АФП для аналоговых сигналов с описанием основных его компонентов, выданы рекомендации по выбору элементной базы.

Практическая значимость выполненной работы заключается в том, что она позволит повысить результативность проведения комплекса тестовых лабораторных испытаний создаваемых систем, с целью уменьшения трудоемкости и сокращения сроков пуско-наладочных работ при их внедрении на объекте.

Список литературы

1. Гаркуша В. В., Собстель Г. М., Сурудин С. П., Яковлев В. В., Гилев В. М., В. И., Пищик Б. Н. Автоматизированная система управления технологическими процессами турбокомпрессорной станции. Проблемы информатики, 2009. № 3. С. 85–93.

2. **Гаркуша В. В., Пищик Б. Н., Михеев В. П., Потатуркин О. И.** Автоматизированная система управления технологическими процессами тепловой станции // *Теплофизика и аэромеханика*, 2006. № 2. С. 315–321.
3. **Благодарный А. И., Гусев О. З., Журавлев С. С., Золотухин Е. П., Каратышева Л. С., Колодей В. В., Михальцов Э. Г., Чейдо Г. П., Шакиров Р. А., Шакиров С. Р.** Автоматизированная система контроля и управления ленточными конвейерами на угольных шахтах // *Горная промышленность*. 2008. № 5 (81). С. 38–44.
4. **Благодарный А. И., Гусев О. З., Журавлев С. С., Зензин А. С., Золотухин Е. П., Каратышева Л. С. и др.** Автоматизированная система наблюдения, оповещения и поиска персонала при авариях в шахтах // *Горная промышленность*. 2009. № 1. С. 34–38.
5. **Журавлев С. С., Окольников В. В., Рудометов С. В., Шакиров С. Р.** Применение подхода «модельно-ориентированного проектирования» к созданию АСУ ТП опасных промышленных объектов // *Вестник НГУ. Серия: Информационные технологии*, 2018. Т. 16, № 4. С. 56–67. DOI 10.25205/1818-7900-2018-16-4-56-67
6. **Журавлев С. С.** Имитационный программно-аппаратный комплекс для тестирования АСУ ТП предприятий горнодобывающей промышленности: Дис. ... канд. техн. наук, Федеральное государственное бюджетное научное учреждение «Федеральный исследовательский центр информационных и вычислительных технологий» (ФИЦ ИВТ), Новосибирск. 2020. 16 с.
7. **Акимов А. А., Козлов М. В., Кудеяров Ю. А., Паньков А. Н., Раевский И. А., Стефанов А. Ю., Стефанов Ю. В.** Стенд для тестирования (испытаний) программного обеспечения средств измерений // *Законодательная и прикладная метрология*. 2008. № 6. С. 25–27.
8. **Попадько В. Е., Барашкин Р. Л., Антипов О. Д., Зуев С. А., Северенко В. С.** Имитационный стенд для тестирования алгоритмов управления объектов нефтегазовой отрасли // Конференция «Компьютерные измерительные технологии». М. 2015. С. 229–233.
9. **Акимов А. А., Козлов М. В., Кудеяров Ю. А., Паньков А. Н., Раевский И. А., Стефанов А. Ю., Стефанов Ю. В.** Стенд для тестирования (испытаний) программного обеспечения средств измерений // *Законодательная и прикладная метрология*. 2008. № 6. С. 25–27.
10. **Андрянов И. Н.** Индустрия 4.0 на практике: виртуальный стенд объекта автоматизации // *Промышленные контроллеры и АСУ*. 2020. № 3. С. 3–9.
11. **Гурина Л. А.** Электромагнитные помехи и методы защиты от них: Учебное пособие. Благовещенск: Амурский гос. ун-т, 2006. – 104 с.
12. ГОСТ 30805.16.2.1-2013 Совместимость технических средств электромагнитная. Требования к аппаратуре для измерения параметров промышленных радиопомех и помехоустойчивости и методы измерений. Часть 2-1. Методы измерений параметров промышленных радиопомех и помехоустойчивости. Измерение кондуктивных радиопомех. Дата введения: 01.01.2014.
13. **Вербин В. С.** Разработка методологической базы для исследования и обеспечения помехоустойчивости управляющих систем и устройств на информационных объектах : диссертация ... канд. техн. наук. Москва, Московский энергетический институт, 2005. 142 с.

References

1. **Garkusha V. V., Sobstel G. M., Surodin S. P., Yakovlev V. V., Gilev V. M., Zapryagaev V. I., Pishchik B. N.** Automated control system for technological processes of a turbocompressor station. *Problems of Informatics*, 2009. No. 3. (in Russ.)
2. **Garkusha V. V., Pishchik B. N., Mikheev V. P., Potaturkin O. I.** Automated control system for technological processes of a thermal station. *Thermophysics and Aeromechanics*, 2006. No. 2, pp. 315–321. (in Russ.)

3. **Zhuravlev S. S., Blagodarny A. I., Gusev O. Z., Zhuravlev S. S., Zolotukhin E. P., Karatysheva L. S., Kolodey V. V.** Automated system for control and management of belt conveyors in coal mines. *Mining*, 2008. No. 5 (81), pp. 38–44. (in Russ.)
4. **Blagodarny A. I., Gusev O. Z., Zhuravlev S. S., Zenzin A. S., Zolotukhin E. P., Karatysheva L. S.** Automated system for monitoring, notification and personnel search in case of accidents in mines. *Mining industry*, 2009. No. 1, pp. 34–38. (in Russ.)
5. **Zhuravlev S. S., Okolnishnikov V. V., Rudometov S. V., Shakirov S. R.** Application of the “Model-Based Design” approach to the creation of automated process control systems for hazardous industrial facilities. *Vestnik NSU. Series: Information Technologies*, 2018. Vol. 16, no. 4, pp. 56–67. DOI 10.25205/1818-7900-2018-16-4-56-67 (in Russ.)
6. **Zhuravlev S. S.** Simulation software and hardware complex for testing automated process control systems for mining enterprises: dissertation of candidate of technical sciences, Federal State Budgetary Scientific Institution “Federal Research Center for Information and Computing Technologies” (FRC ICT), Novosibirsk, 2020. 16 p. (in Russ.)
7. **Akimov A. A., Kozlov M. V., Kudiyarov Yu. A., Pankov A. N., Raevsky I. A., Stefanov A. Yu., Stefanov Yu. V.** Stand for testing (trials) software for measuring instruments. *Legislative and applied metrology*, 2008. No. 6, pp. 25–27. (in Russ.)
8. **Popad’ko V. E., Barashkin R. L., Antipov O. D., Zuev S. A., Severenko V. S.** Simulation stand for testing control algorithms for oil and gas industry facilities. Conference “COMPUTER MEASURING TECHNOLOGIES”. M., 2015. Pp. 229–233. (in Russ.)
9. **Akimov A. A., Kozlov M. V., Kudiyarov Yu. A., Pankov A. N., Raevsky I. A., Stefanov A. Yu.** Stand for testing (trials) software for measuring instruments. *Legislative and applied metrology*, 2008. No. 6, pp. 25–27. (in Russ.)
10. **Andriyanov I. N.** Industry 4.0 in practice: a virtual stand of an automation object. *Industrial controllers and automated control systems*, 2020. No. 3, pp. 3–9. (in Russ.)
11. **Gurina L. A.** Electromagnetic interference and methods of protection against them // Study guide. Blagoveshchensk: Amur State University, 2006. 104 p.
12. GOST 30805.16.2.1-2013 Electromagnetic compatibility of technical means. Requirements for equipment for measuring the parameters of industrial radio interference and noise immunity and measurement methods. Part 2-1. Methods for measuring the parameters of industrial radio interference and noise immunity. Measurement of conducted radio interference. Date of introduction: 01.01.2014. (in Russ.)
13. **Verbin V. S.** Development of a methodological base for the study and ensuring the noise immunity of control systems and devices at information objects: dissertation of candidate of technical sciences. Moscow, Moscow Energy Institute, 2005. 142 p. (in Russ.)

Информация об авторах

Гаркуша Владимир Владимирович, заведующий сектором измерительных систем и приборостроения центра инженерно-технического обеспечения, Федеральный исследовательский центр информационных и вычислительных технологий (Новосибирск, Россия)
РИНЦ Author ID: 557936
SCOPUS Author ID: 8253890500

Журавлев Сергей Сергеевич, кандидат технических наук, научный сотрудник лаборатории автоматизированных систем, Федеральный исследовательский центр информационных и вычислительных технологий (Новосибирск, Россия)
РИНЦ SPIN: 7484-8635
Web of Science Researcher ID: E-7348-2014
SCOPUS Author ID: 57206421044

Шакиров Станислав Рудольфович, кандидат физико-математических наук, исполняющий обязанности директора, Конструкторско-технологический институт научного приборостроения Сибирского отделения Российской академии наук (Новосибирск, Россия)

РИНЦ SPIN: 1929-2332

Web of Science Researcher ID: U-4958-2018

SCOPUS Author ID: 57196404292

Яковлев Владимир Владимирович, младший научный сотрудник лаборатории индустриальной информатики, Федеральный исследовательский центр информационных и вычислительных технологий (Новосибирск, Россия)

РИНЦ SPIN: 2311-8180

Web of Science ResearcherID: C-7406-2017

SCOPUS Author ID: 57197688579

Information about the Authors

Vladimir V. Garkusha, Head of the measuring systems and instrumentation sector of the laboratory of engineering and technical support, Federal Research Center for Information and Computational (Novosibirsk, Russian Federation)

RSCI Author ID: 557936

SCOPUS Author ID: 8253890500

Sergey S. Zhuravlev, Candidate of Technical Sciences, Researcher, Laboratory of Automated Systems, Federal Research Center for Information and Computational Technologies (Novosibirsk, Russian Federation)

RSCI SPIN: 7484-8635

Web of Science Researcher ID: E-7348-2014

SCOPUS Author ID: 57206421044

Stanislav R. Shakirov, Candidate of Physical and Mathematical Sciences, acting director, Technological Design Institute of Scientific Instrument Engineering of Siberian Branch of RAS (Novosibirsk, Russian Federation)

RSCI SPIN: 1929-2332

Web of Science ResearcherID: U-4958-2018

SCOPUS Author ID: 57196404292

Vladimir V. Yakovlev, Junior Researcher, Laboratory of Industrial Informatics, Federal Research Center for Information and Computational Technologies (Novosibirsk, Russian Federation)

RSCI SPIN: 2311-8180

Web of Science ResearcherID: C-7406-2017

SCOPUS Author ID: 57197688579

Статья поступила в редакцию 01.06.2022;

одобрена после рецензирования 25.09.2022; принята к публикации 25.09.2022

The article was submitted 01.06.2022;

approved after reviewing 25.09.2022; accepted for publication 25.09.2022

Научная статья

УДК 004.9

DOI 10.25205/1818-7900-2022-20-3-29-37

Развитие информационной среды СО РАН в условиях санкционных ограничений и информационной войны против РФ

**Ольга Анатольевна Клименко¹, Елена Владимировна Рычкова²
Игорь Владимирович Шабальников³**

¹⁻³Федеральный исследовательский центр информационных и вычислительных технологий
Новосибирск, Россия

¹o.klimenko@sb-ras.ru

²helen@ict.nsc.ru, <https://orcid.org/0000-0001-9692-8441>

³shabalnikov@ict.sbras.ru, <https://orcid.org/0000-0001-9821-854X>

Аннотация

Рассматривается информационная среда Сибирского отделения Российской академии наук (СО РАН), которая начала складываться с конца девяностых двадцатого века с созданием корпоративного сайта, сайтов институтов и НГУ. Сейчас информационная среда СО РАН объединяет более 100 научных организаций и вузов Сибири, служит площадкой, на которой представлен научно-технический потенциал Востока России. В связи с санкционным давлением и информационной войной против России возникли неотложные задачи. В статье предлагается использовать информационную среду СО РАН для публикации отечественных баз открытых данных, которые заменят зарубежные информационные системы с открытыми данными типа Figshare и Zenodo. В качестве примера приводится раздел «Импортонезависимость», который создан на Портале СО РАН в мае 2022 года и содержит открытые данные.

Ключевые слова

корпоративный сайт, открытые данные, метаданные, импортонезависимость

Для цитирования

Клименко О. А., Рычкова Е. В., Шабальников И. В. Развитие информационной среды СО РАН в условиях санкционных ограничений и информационной войны против РФ // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С.29–37. DOI 10.25205/1818-7900-2022-20-3-29-37

Development of the Information Environment of SB RAS under Sanctions and Information Warfare against the Russian Federation

Olga A. Klimenko¹, Elena V. Rychkova², Igor V. Shabalnikov³

¹⁻³Federal Research Center for Information and Computational Technologies
Novosibirsk, Russian Federation

¹o.klimenko@sb-ras.ru

²helen@ict.nsc.ru, <https://orcid.org/0000-0001-9692-8441>

shabalnikov@ict.sbras.ru, <https://orcid.org/0000-0001-9821-854X>

Abstract

The purpose of the work is to develop the information environment of the Siberian Branch of the Russian Academy of Sciences (SB RAS). The information environment of SB RAS has begun to take shape since the end of the 1990s with the creation of a corporate website, websites of institutes and the NSU website. Now the information environment of SB RAS unites more than 100 scientific organizations and universities of Siberia, serves as a platform where the

© Клименко О. А., Рычкова Е. В., Шабальников И. В., 2022

scientific and technical potential of the East of Russia is presented. Because of the recently introduced sanctions and the information warfare against Russia, some urgent tasks have arisen – in particular, to replace the information systems with open data such as Figshare and Zenodo. The article proposes to present a catalog of domestic open data databases on the SB RAS Portal. As an example, the section “Import-independence” that was created in May 2022 and contains open data is presented.

Keywords

corporate website, open data, metadata, import independence

For citation

Klimenko O. A., Rychkova E. V., Shabalnikov I. V. Development of the Information Environment of SB RAS under Sanctions and Information Warfare against the Russian Federation. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 29–37. DOI 10.25205/1818-7900-2022-20-3-29-37

Введение

Сибирское отделение РАН – крупнейший интегратор и основной эксперт научно-исследовательских, научно-образовательных, опытно-конструкторских и производственных организаций Востока России. Корпоративный сайт Сибирского отделения Российской академии наук – портал СО РАН¹ был создан в 1996 году по инициативе академика Ю. И. Шокина [1–3]. Первый вариант сайта был разработан под руководством чл.-корр. РАН А. М. Федотова в Институте вычислительных технологий СО РАН. Принципы, на которых портал СО РАН работает и сейчас, были заложены в начале нулевых на заседаниях семинара ИВТ СО РАН по информационным технологиям. При создании корпоративного сайта СО РАН одной из первоочередных задач было создание базы данных «Организации и сотрудники СО РАН» [4]. Как сделать идеальную кадровую базу, какую персональную информацию нужно использовать, как связать персону и место работы, все эти вопросы обсуждались на семинарах и совещаниях. Выводом было то, что идеальной базы не существует, задача решается для конкретных организаций по-разному. Отметим, что на данный момент база данных «Организации и сотрудники СО РАН»² является основной, в ней аккумулируется информация о руководстве институтов СО РАН, членах РАН, профессорах РАН, которые работают в Сибирском отделении, хранятся сведения о 180 организациях и более чем 2500 персонах.

Другой важной инициативой, предложенной разработчиками портала и их коллегами из ряда институтов СО РАН, стало создание виртуального музея науки и техники СО РАН [5]. Виртуальный музей³ представляет собой объединение различных электронных коллекций. Коллекции находятся в Новосибирске, Томске, Иркутске, Красноярске и других городах Сибири. Информация о каждой коллекции вносится в каталог, который расположен на центральном сервере в Новосибирске. Каталогизируются коллекции с помощью международной схемы метаданных, рубрикатора информационных ресурсов и классификаторов.

Существующая информационная среда

В настоящее время структура портала СО РАН представляет собой систему, в которой каждый элемент (новость, событие или документ) имеет метаданные. Одним из атрибутов метаданных является направления научных исследований [6; 7]. Благодаря этому можно анализировать, как распределяются новости на портале СО РАН, по данному атрибуту: в 2021 году было опубликовано 1900 новостей, связанных с направлениями научных исследований и 434 публикации с категорией «документ». По сравнению с 2020 годом в процентном отношении увеличилось количество новостей по направлениям «Науки о Земле», «Нанотехнологии и информационные технологии», «Экономические науки» (см. рис. 1).

¹ URL: <https://www.sbras.ru>.

² URL: <https://www.sbras.ru/ru/sbras/db>.

³ URL: <https://www.sbras.ru/ru/prj/museum>.



Рис. 1. Распределение новостей по направления научных исследований

Fig. 1. Distribution of news by direction of scientific research

На рис. 2 показано количество опубликованных в 2020 и 2021 годах новостей (в процентах от общего количества новостей за год) по каждому из направлений науки (МИ – Математика и информатика; ФН – Физические науки; НИТ – Нанотехнологии и информационные технологии; ЭММиПУ – Энергетика, машиностроение, механика и процессы управления; ХН – Химические науки; БН – Биологические науки; НЗ – Науки о Земле; ГН – Гуманитарные науки; ЭН – Экономические науки; МН – Медицинские науки; СХН – Сельскохозяйственные науки).

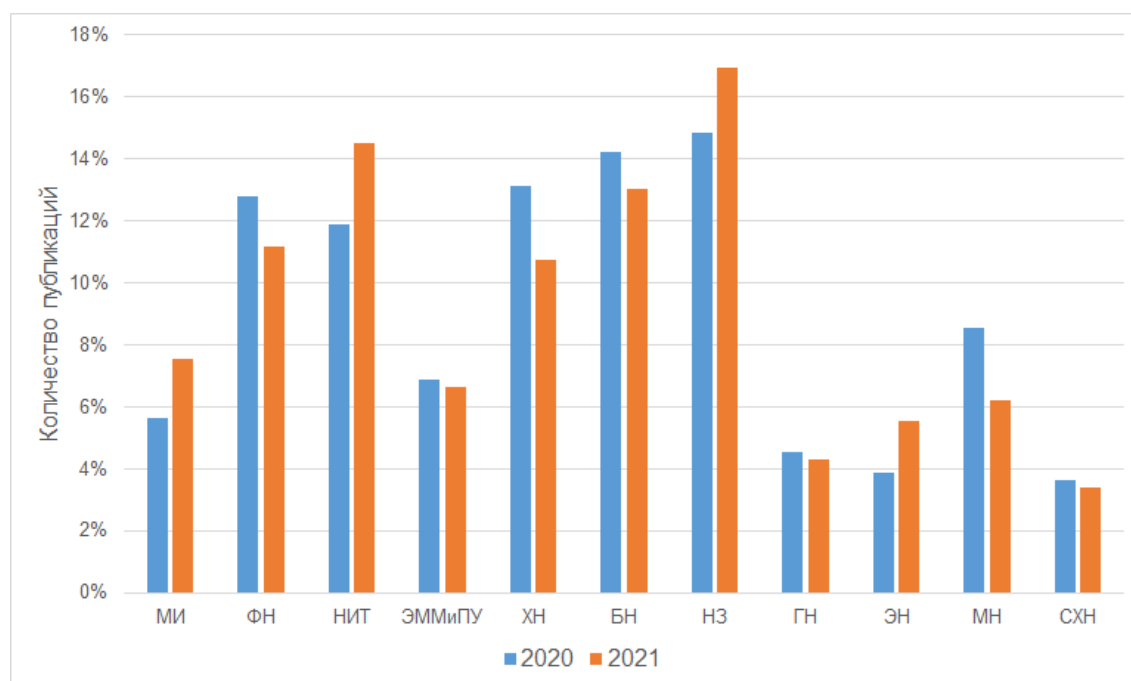


Рис. 2. Количество новостных публикаций в 2020 и 2021 годах

Fig. 2. Number of news publications in 2020 and 2021

В том случае, когда содержание новости представляет интеграционный характер исследований, ей возможно присвоить не более двух направлений науки. Сравнение количества опубликованных в 2020 и 2021 годах новостей (в процентах от общего количества новостей за год)

для сочетания направлений науки отображено на рис. 3. На диаграмме показаны только те пары направлений науки, у которых количество новостей было более 1 % за год.

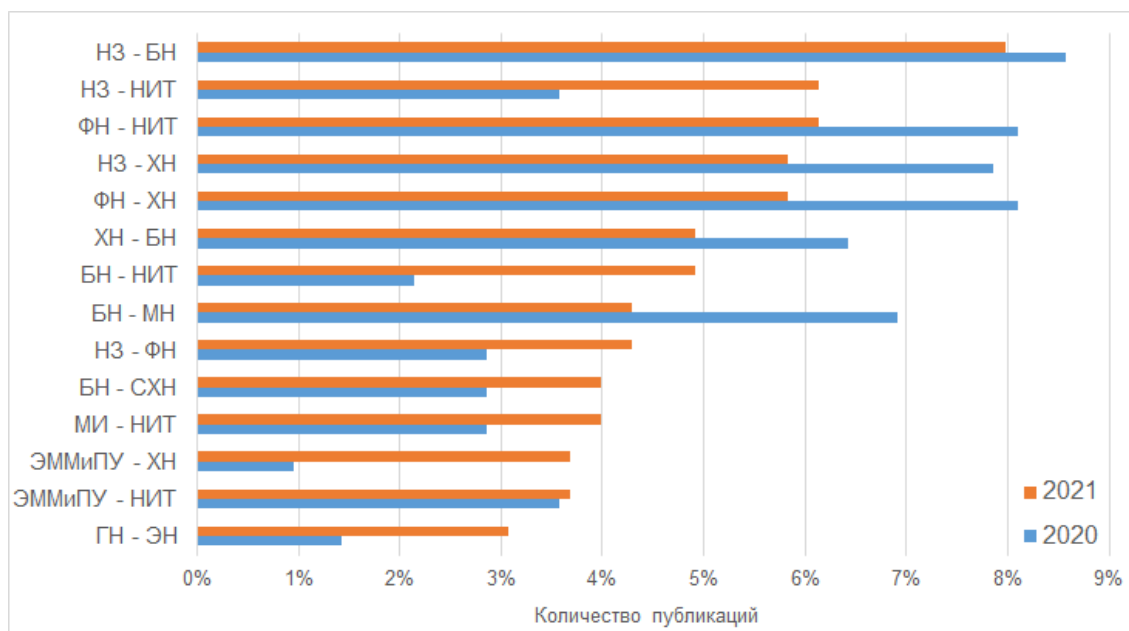


Рис. 3. Количество новостных публикаций в 2020 и 2021 годах для сочетания направлений научных исследований

Fig. 3. Number of news publications in 2020 and 2021 for a combination of two scientific directions

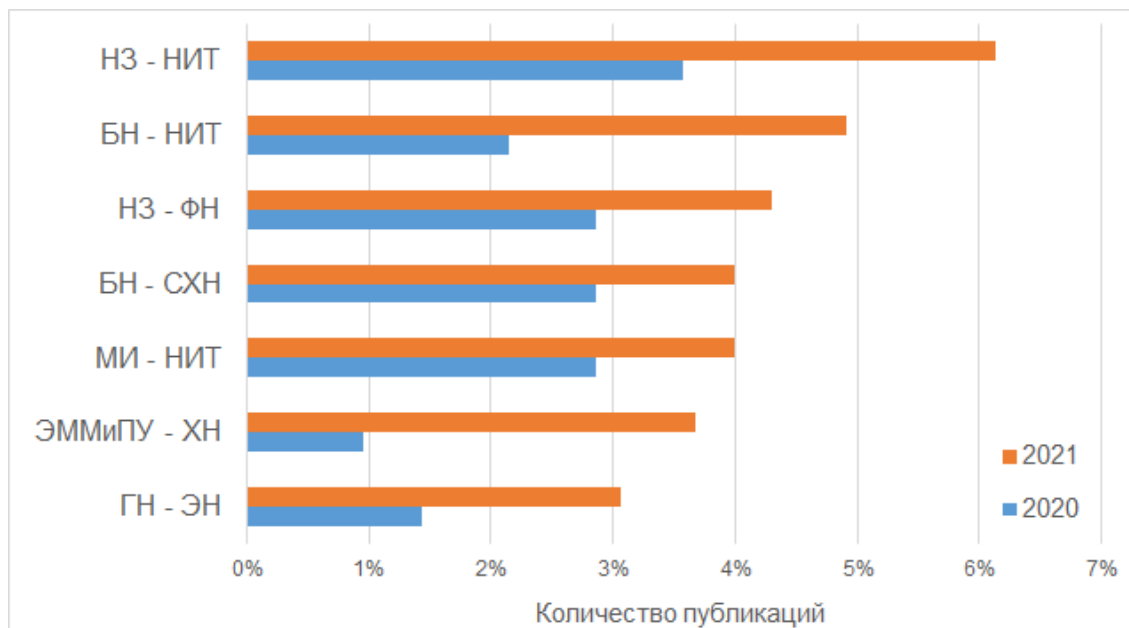


Рис. 4. Сочетания направлений науки, показавшие рост количества новостных публикаций в 2021 году по сравнению с 2020 годом

Fig. 4. Combinations of two scientific directions that showed an increase in the number of news publications in 2021 compared to 2020

Семь сочетаний направлений науки (см. рис. 4) показали рост количества новостных публикаций в 2021 году по сравнению с 2020 годом (в процентах от общего количества новостей за год).

В 2021 году продолжился тренд проникновения информационных технологий и математического моделирования во все науки, что выразилось в росте публикаций на портале СО РАН на стыке наук. Новым стал рост публикаций по экологической тематике и «зеленой» энергетике.

В начале 2021 года был образован новый подраздел «Пресс-служба Президиума СО РАН»⁴, к которому было отнесено 30 публикаций. В разделе «Инновационные разработки»⁵ размещены полные тексты трех выпусков издания «Наука и технологии Сибири». На портале СО РАН хранится и пополняется архив издания «Наука в Сибири»⁶, за 2021 год был 5371 просмотр данной страницы.

В 2021 году в разделе «Организации и сотрудники СО РАН» были созданы два новых подраздела. Первый подраздел: «Персональный состав Сибирского отделения с 1957 года»⁷. В нем представлена информация о членах Академии с момента основания Сибирского отделения АН СССР в 1957 году и по сегодняшний день. Можно увидеть по каким специальностям проходили выборы в члены-корреспонденты и действительные члены Академии. Информация о выборах в члены Академии сгруппирована по годам. Второй подраздел – «Почетные доктора СО РАН»⁸, где перечислены ученые зарубежных стран, которым присвоено звание «Почетный доктор Сибирского отделения РАН» (начиная с 2001 года).

В 2022 году был существенно дополнен и раздел «Академический час для школьников»⁹, содержащий сведения о 78 лекциях (с 2011 года и по настоящее время), с которыми выступали ведущие ученые СО РАН в рамках проекта СО РАН «Академический час для школьников».

Изменение модели в сторону более тесной связи с внешними информационными потоками

Информация на Портал СО РАН (и из него) поступает по разнообразным каналам (см. рис. 5). Например, с сайтов институтов СО РАН дублируется информация про научные разработки, анонсы научных конференций. В обязательном порядке в разделе «Документы» размещается официальная информация, касающаяся научной сферы, полученная с сайтов Президента РФ и Правительства РФ. С официальных сайтов регионов Сибири извлекаются новости от руководителей регионов и научно-технических советов (например, Красноярский краевой фонд поддержки научной и научно-технической деятельности) про конкурсы и гранты, касающихся науки и высшего образования, экономического развития, научного сопровождения, программы, стратегические документы. Активно идет взаимодействие портала СО РАН со СМИ (издания «Наука в Сибири», «Поиск», «Наука в деталях», «Наука из первых рук»). Размещаются новости Академпарка, НГУ и других вузов Сибири о совместных научных разработках, научных школах и конференциях; базовых школах РАН.

Важным элементом портала СО РАН, где ежегодно (начиная с 1995 года) размещается концентрированная информация о выдающихся результатах и разработках, ведущихся на переднем фронте науки, является раздел «Отчеты о деятельности СО РАН»¹⁰. Особенностью этого раздела является то, что его содержимое в части лучших научных результатов, представляемых в отчет СО РАН за каждый год, проходит многоступенчатый экспертный отбор: на ученых советах институтов и вузов, в Объединенных ученых советах по направлениям наук. Оконча-

⁴ URL: <https://www.sbras.ru/press>.

⁵ URL: <https://www.sbras.ru/ru/innovation>.

⁶ URL: https://www.sbras.ru/ru/HBC_archive.

⁷ URL: https://www.sbras.ru/ru/members_sbras_history.

⁸ URL: https://www.sbras.ru/ru/honorary_doctors_sbras.

⁹ URL: <https://www.sbras.ru/ru/academichour>.

¹⁰ URL: <https://www.sbras.ru/ru/cmn/reports>.

тельный выбор осуществляет председатель СО РАН, отбирая самые интересные и перспективные результаты.

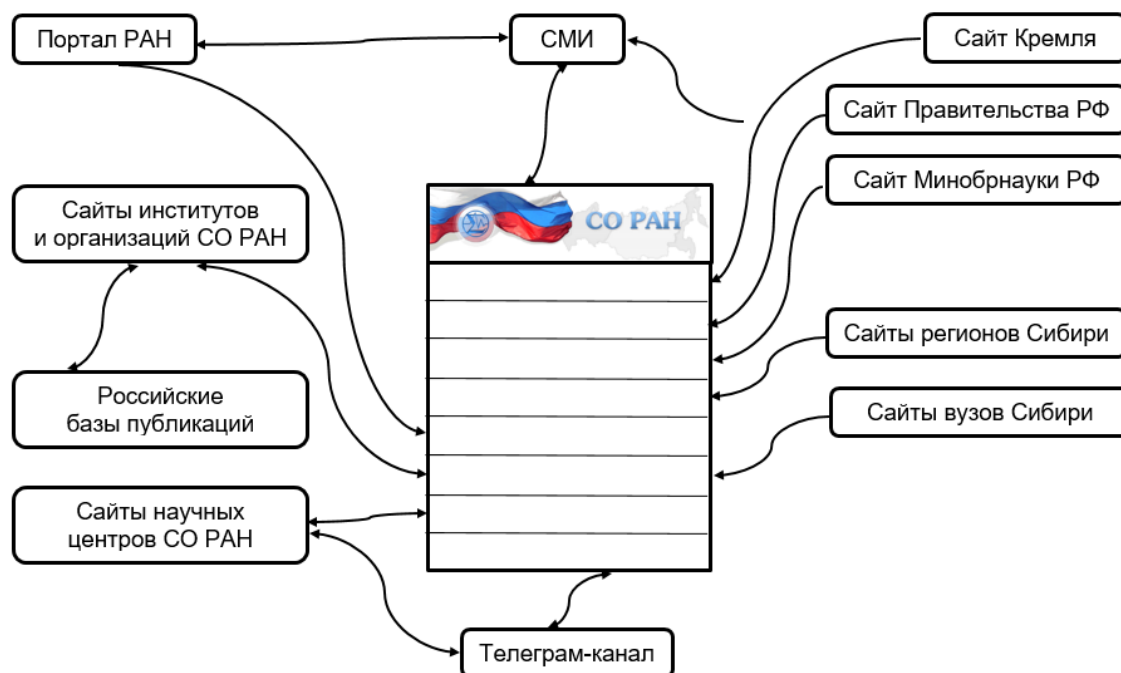


Рис. 5. Портал СО РАН: каналы получения и обмена информацией
Fig. 5. Portal of the SB RAS: channels for receiving and exchanging information

В последние три года была отмечена тенденция увеличения заходов на портал СО РАН пользователями из социальных сетей, телеграм-каналов, которые имеются у РАН, Министерства науки и высшего образования РФ, СМИ, видных ученых, руководителей федерального и региональных уровней.

Импортозамещение и импортонезависимость

В 2022 году изменилась мировая информационная среда, резко повысились требования к устойчивому функционированию российских сайтов и сетей. Под вопросом стала надежность обновляемого свободно распространяемого программного обеспечения (ПО), возможность покупки серверов, систем хранения данных, сетевого оборудования. В соответствии с Указом Президента РФ № 250¹¹, запрещается учреждениям и госкомпаниям с 1 января 2025 года использовать средства защиты информации, произведенные в недружественных странах.

Все это надо учитывать при выборе пути развития российских корпоративных систем и информационной среды СО РАН, в частности. Можно выделить три больших задачи, одна связана с выбором оборудования и ПО, другая – с обеспечением безопасности и надежности работы, третья – с информационным наполнением корпоративной среды.

Рассмотрим третью задачу. В связи с тем, что против РФ введены санкции, нацеленные на изоляцию, разрушение экономики, задержку научно-технического развития, возникает целый ряд задач импортозамещения и импортонезависимости. В частности, создание отече-

¹¹ Указ Президента РФ от 01.05.2022 № 250 «О дополнительных мерах по обеспечению информационной безопасности Российской Федерации». URL: <http://www.kremlin.ru/acts/news/68322>.

ственных баз открытых данных, которые заменят зарубежные информационные системы с открытыми данными типа Figshare¹² и Zenodo¹³.

В качестве начальной задачи можно рассматривать размещение на Портале СО РАН списка открытых данных, имеющихся в наличии у научных организаций СО РАН.

Под открытыми данными мы понимаем упорядоченную текстовую или цифровую информацию, сгенерированную в институтах СО РАН или вузах Сибири, которой любой человек или организация может воспользоваться на безвозмездной основе с целью анализа и обработки. В качестве примера можно привести публикацию открытых данных различными государственными агентствами и структурами: Федеральной службой государственной статистики¹⁴, Министерством науки и высшего образования РФ¹⁵.

Предполагается, что с исходными файлами будут осуществляться различные преобразования, которые заранее невозможно предугадать. Поэтому файлы публикуют в простых и привычных форматах. Так, например, если это статистические таблицы то, обычно, это либо текстовые файлы в формате CSV, либо Excel (XLSX). Для изображений это форматы GIF, JPEG, PNG. Важно, чтобы помимо собственно данных в набор входил и файл с метаописанием (автор, контактные данные, дата создания, правовая информация, область знаний и др.).

Открытые данные могут использоваться в учебных или исследовательских целях. Например, исследователь с помощью своей программы и своими алгоритмами обрабатывает данные. При публикации результата, исследователь дает описание алгоритма и указывает на каких открытых данных он проводил исследования. Это позволяет другому исследователю, использующему ту же программу и те же данные воспроизвести результат или улучшить его.

Примером открытых данных может быть раздел «Импортнезависимость»¹⁶, который создан на портале СО РАН в мае 2022 года. В этом разделе разработки сгруппированы по категориям (Макрорегион Сибирь; Информационные технологии; Микроэлектроника; Сырье; Добыча нефти и газа; Лекарства и медицинские изделия; Научное приборостроение; Химические реактивы для исследований; Семена, Птицеводство; Животноводство; Технологии для агрокомплекса; Технологии промышленности; Экология; Продукты питания). Если разработка отнесена к двум или трём категориям, то она присутствует в каждой из этих категорий. Перечисленные категории представлены в виде веб-страниц, на которых для каждой разработки приведены ее название, назначение, решаемые задачи и ссылка на страницу института-разработчика в базе данных «Организации и сотрудники СО РАН». На главной странице раздела «Импортнезависимость» также приведена ссылка на zip-файл с исходными данными в формате Excel (XLSX) и их метаописанием. Таблица Excel состоит из 4 колонок и содержит 328 строк текстовой информации.

Заключение

В условиях санкционных ограничений и информационной войны против России необходимо решать задачу создания и публикации в открытом доступе отечественных баз данных, которые не только заменят зарубежные информационные системы, но и дадут исследователям, руководителям всех уровней, экспертам, общественности достоверную информацию пригодную для анализа и трансформации. Информационная среда СО РАН охватывает академические институты и вузы, может использоваться для представления каталога и самих баз открытых данных.

¹² URL: <https://figshare.com/>.

¹³ URL: <https://zenodo.org/>.

¹⁴ URL: <https://rosstat.gov.ru/opendata/>.

¹⁵ URL: <https://minobrnauki.gov.ru/opendata/>.

¹⁶ URL: https://www.sbras.ru/ru/import_substitution.

Список литературы

1. **Шокин Ю. И., Федотов А. М., Никульцев В. С.** Вопросы развития Сети Интернет Новосибирского научного центра // Вычислительные технологии. 1997. Т. 2, № 3. С. 97–103.
2. **Шокин Ю. И., Федотов А. М., Жижимов О. Л., Федотова О. А.** Эволюция информационных систем: от WEB-сайтов до систем управления информационными ресурсами // Вестник НГУ. Серия: Информационные технологии. 2015. Т. 13, № 1. С. 117–134.
3. **Шокин Ю. И., Федотов А. М., Жижимов О. Л.** Технологии создания распределенных информационных систем для поддержки научных исследований // Вычислительные технологии. 2015. Т. 20, № 5. С. 251–274.
4. **Шокин Ю. И., Федотов А. М., Барахнин В. Б.** Проблемы поиска информации. Новосибирск: Наука. 2010. С. 130–133.
5. **Шокин Ю. И., Ламин В. А., Федотов А. М., Барахнин В. Б., Жижимов О. Л., Мазов Н. А., Пищик Б. Н., Покровский Н. Н., Рычкова Е. В.** Виртуальный музей науки и техники СО РАН // В сборнике: Информационно-телекоммуникационные ресурсы СО РАН. Материалы выездного заседания координационного научного совета СО РАН по целевой программе. Российская академия наук, Сибирское отделение; Иркутский научный центр. 2003. С. 118–125.
6. **Клименко О. А., Рычкова Е. В., Шабальников И. В.** Корпоративный портал Сибирского отделения РАН: модель, аналитика, вебометрика // Информационные и математические технологии в науке и управлении. 2017. № 2(6). С. 44–50.
7. **Клименко О. А., Рычкова Е. В., Шабальников И. В.** Портал СО РАН как основа знаний о научной корпорации // Новые информационные технологии в образовании и науке. 2021. Т. 4. С. 44–47.

References

1. **Shokin Y. I., Fedotov A. M., Nikul'tsev V. S.** On development of Internet network in Novosibirsk Research Centre. Computational Technologies, 1997. Vol. 2, no. 3, pp. 97–103. (in Russ.)
2. **Shokin I. Yu., Fedotov A. M., Zhizhimov O. L., Fedotova O. A.** The evolution of information systems: from websites to information resource management systems. Vestnik NSU. Series: Information Technologies, 2015. Vol. 13, no. 1, pp. 117–134. (in Russ.)
3. **Shokin Y. I., Fedotov A. M., Zhizhimov O. L.** Technologies for designing of distributed information systems to support research. Computational Technologies, 2015. Vol. 20, no. 5, p. 251–274. (in Russ.)
4. **Shokin Y. I., Fedotov A. M., Barakhnin V. B.** Problemy poiska informatsii [Information search problems], Novosibirsk: Nauka, pp. 130–133. (in Russ.)
5. **Shokin Y. I., Lamin V. A., Fedotov A. M., Barakhnin V. B., Zhizhimov O. L., Mazov N. A., Pishchik B. N., Pokrovskiy N. N., Rychkova E. V.** Virtualnyy muzey nauki i tekhniki SO RAN [Virtual Museum of Science and Technology of the SB RAS]. In: *Informatsionno-telekommunikatsionnye resursy SO RAN. Materialy vyezdnogo zasedaniya koordinatsionnogo nauchnogo soveta SO RAN po tselevoy programme*. RAS, SB RAS, Irkutsk Scientific Center, 2003, pp. 118–125. (in Russ.)
6. **Klimenko O. A., Rychkova E. V., Shabalnikov I. V.** Corporate portal of the Siberian Branch of the Russian Academy of Sciences: Model, Analytics, Webometrics. Information and mathematical technologies in science and management, 2017. No. 2(6), pp. 44–50. (in Russ.)
7. **Klimenko O. A., Rychkova E. V., Shabalnikov I. V.** The SB RAS Portal as a Basis of Knowledge about the Scientific Corporation. New Information Technologies in Education and Science, 2021. Vol. 4, pp. 44–47. (in Russ.)

Информация об авторах

Клименок Ольга Анатольевна, кандидат физико-математических наук, старший научный сотрудник, Федеральный исследовательский центр информационных и вычислительных технологий (Новосибирск, Россия)

Рычкова Елена Владимировна, кандидат физико-математических наук, младший научный сотрудник, Федеральный исследовательский центр информационных и вычислительных технологий (Новосибирск, Россия)

Шабальников Игорь Владимирович, ведущий специалист, Федеральный исследовательский центр информационных и вычислительных технологий (Новосибирск, Россия)

Information about the Authors

Olga A. Klimenko, Candidate of Sciences (Physics and Mathematics), Senior Researcher, Federal Research Center for Information and Computing Technologies (Novosibirsk, Russian Federation)

Elena V. Rychkova, Candidate of Sciences (Physics and Mathematics), Junior Researcher, Federal Research Center for Information and Computing Technologies (Novosibirsk, Russian Federation)

Igor V. Shabalnikov, Leading Specialist, Federal Research Center for Information and Computing Technologies (Novosibirsk, Russian Federation)

Статья поступила в редакцию 23.06.2022;

одобрена после рецензирования 06.10.2022; принята к публикации 06.10.2022

The article was submitted 23.06.2022;

approved after reviewing 06.10.2022; accepted for publication 06.10.2022

Научная статья

УДК 004.032.26

DOI 10.25205/1818-7900-2022-20-3-38-50

Метод для восстановления аудиосигнала с помощью свёрточных нейронных сетей

Кристина Игоревна Дементьева¹, Антон Андреевич Ракитский²

^{1,2}Сибирский государственный университет телекоммуникаций и информатики
Новосибирск, Россия

²Новосибирский государственный университет
Новосибирск, Россия

¹k.morozova@cyber.sibsutis.ru, <https://orcid.org/0000-0001-8650-3749>

²rakitsky.anton@gmail.com, <https://orcid.org/0000-0002-1224-4278>

Аннотация

В работе представлено исследование возможности восстановления искажённого аудио-сигнала. На основе полученных ранее результатов использования методов глубокого машинного обучения разработана концепция нейронной сети, предназначенная для коррекции искаженного звукового сигнала. На базе первоначально полученной разработано несколько новых архитектур нейронных сетей, ориентированных на восстановление аудио-сигнала. В статье приведены описания разработанных архитектур с теоретическим обоснованием возможности их применения. Представленные архитектуры были протестированы для решения задачи восстановления партии конкретного инструмента в музыкальной композиции, из которой она была удалена.

Ключевые слова

аудиосигнал, восстановление сигнала, методы машинного обучения, нейронные сети, регрессия

Для цитирования

Дементьева К. И., Ракитский А. А. Метод для восстановления аудиосигнала с помощью свёрточных нейронных сетей // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 38–50. DOI 10.25205/1818-7900-2022-20-3-38-50

Method for Restoring Audio Signal Using Convolutional Neural Networks

Kristina. I. Dementyeva¹, Anton A. Rakitskiy²

^{1,2}Siberian State University of Telecommunications and Information Sciences
Novosibirsk, Russian Federation

²Novosibirsk State University
Novosibirsk, Russian Federation

¹k.morozova@cyber.sibsutis.ru, <https://orcid.org/0000-0001-8650-3749>

²rakitsky.anton@gmail.com, <https://orcid.org/0000-0002-1224-4278>

Abstract

This paper describes a research of the restoring distorted audio signal possibility. Based on the previously obtained results of using deep machine learning methods, the concept of a neural network to correct a distorted audio signal has been developed. On the basis of the originally obtained results, several new neural network architectures were developed, focused on the audio signal restoring. The paper contains descriptions of the developed architectures with a theoretical substantiation of the possibility of their application. The presented architectures were tested to solve the problem of restoring the part of a specific instrument in a musical composition where it was removed. The results of testing the developed architectures of neural networks are presented in several forms.

© Дементьева К. И., Ракитский А. А., 2022

Keywords

audio signal, signal recovery, machine learning methods, neural networks, regression

*For citation*Dementyeva K. I., Rakitsky A. A. Method for Restoring Audio Signal Using Convolutional Neural Networks. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 38–50. DOI 10.25205/1818-7900-2022-20-3-38-50

Введение

Среди областей обработки сигналов можно выделить такие, как: наложение эффектов, распознавание и устранение шума, усиление сигнала, улучшение качества и частотных характеристик, восстановление в случае потерь, восстановление удаленной части сигнала и др. Восстановление направлено на исправление искаженной части таким образом, чтобы она была наиболее близкой к оригиналу.

Искажения разного рода могут возникать в аудиосигнале при его формировании, передаче, записи или обработке. Например, при записи музыкальных партий в студии возможно возникновение сбоев в работе звукозаписывающего оборудования, что приводит к потере части аудиосигнала. Не исключено, что сходящий с конвейера носитель информации аудио-сигнала имеет дефекты, которые приведут к повреждению сигнала. При распространении звуковой волны в какой-либо среде также могут возникнуть искажения (например, телефония, эхолокация и т. д.), вызванные как особенностями самой среды, так и внешним воздействием. В процессе эксплуатации носителя аудиосигнала также могут появиться новые повреждения, приводящие к потере данных. Существуют и другие возможности повреждения аудиосигнала. В данной работе предлагается новый способ исправления таких искажений, базирующийся на свёрточных нейронных сетях.

В последующих главах формулируется модель искажения, которая наиболее адаптирована для применения к ней нейронных сетей, описан процесс исследования структуры аудиосигнала, рассмотрены различные архитектуры нейронных сетей и представлена архитектура, которая позволяет решить задачу восстановления сигнала с высокой точностью. В общем случае, исследование конкретных практических задач восстановления сигнала влияет на потенциальную возможность дальнейшей разработки инновационного программного обеспечения, позволяющего доносить до пользователя истинную форму сигнала. В разделах «Особенности реализации» и «Анализ результатов» разработанные архитектуры исследуются в рамках применения к задаче восстановления ударной партии в музыкальной композиции, из которой она была удалена.

1. Анализ состояния предметной области

В настоящее время известен ряд методов восстановления сигналов, в частности: в работе [1] исследована задача восстановления аудио на основе заполнения недостающих сегментов аудио. Исследователи в рамках данной работы предлагают улучшения в наложении звука, направленные на компенсацию потери в аудиосигнале.

В исследовании [2] предложены методы восстановления аудиофайлов, основанные на нейронных сетях с прямой связью и долговременной памятью, а также описаны возможности автоматизации задачи восстановления сигнала с помощью предлагаемых методов восстановления, использующих глубокие нейронные сети.

Подходы к идентификации и восстановлению структурированных данных на основе методов глубокого обучения описаны в [3–5]. Реконструкция звуковых сигналов, искаженных щелчками и потрескиваниями, методами сжатия и восстановления представлена в [6].

В исследовательской статье [7] предлагается алгоритм на основе многослойной свёрточной нейронной сети для улучшения звукового сигнала. Частотные характеристики звукового сигнала были извлечены как набор частот, которые кратко описывают общую форму спектраль-

ной огибающей. Поскольку аудиосигнал изменяется с очень постоянной скоростью, предполагается, что аудиосигнал не сильно изменяется за короткие промежутки времени. Предлагаемая нейронная сеть принимает входные данные от разных кадров из звукового сигнала, загрязненного шумом, для обучения и тестирования.

В отличие от описанных выше методов, в данной работе исследуется возможность восстановления звукового сигнала с помощью регрессионного анализа. В частности, рассматривается проблема восстановления партии ударных в музыкальном произведении, где она была удалена, и способы ее решения на основе методов машинного обучения.

2. Описание модели искажения и восстановления аудиосигнала

Восстановление сигнала – реконструкция истинной формы сигнала, несущего полезное сообщение после его искажения зашумлением либо удалением какой-либо его части.

Предположим, что при восстановлении аудиосигнала можно следовать следующей схеме: используя кратковременное преобразование Фурье для получения таких характеристик сигнала, как синусоидальная частота и фазовое содержание локальных участков сигнала по мере его изменения во времени и, разделив звуковой сигнал на небольшие промежутки времени, можно представить его в виде «изображения». Благодаря произведенным действиям, «изображения» сигнала имеют пространственные закономерности во временной и частотной области, что позволяет обучить искусственный интеллект обнаруживать отсутствие составляющих аудиосигнала на конкретном частотном интервале в конкретный момент времени.

В исходном сигнале покадрово выполняется преобразование Фурье и, таким образом, определяется частотный спектр каждого кадра аудиосигнала. Опишем это преобразование подробно: каждая часть исходного сигнала подвергается преобразованию Фурье для получения m частотных диапазонов компонент сигнала из рассматриваемого кадра; затем отсекается половина полученного спектра, поскольку для входного сигнала с действительным знаком спектр в диапазоне от $1/2T$ до $1/T$ просто отражает спектр от 0 до $1/2T$ (где $1/T$ – частота дискретизации) и, соответственно, не несет полезной нагрузки.

Таким образом, получен набор кадров, который представляет информацию о значениях всех частотных диапазонов в конкретных временных интервалах. В каждом конкретном кадре необходимо восстановить недостающую часть ударника или выявить, что она там не нужна; делать это нужно с учетом зависимости от соседних кадров.

Рассмотрим некоторый аудиосигнал длительностью t секунд, записанный с частотой дискретизации R , в котором необходимо восстановить какую-то его составляющую (например, в музыкальной композиции была вырезана партия ударника). Разделим этот сигнал на n равных временных интервалов (кадров), которые будут содержать m значений сигнала. К каждому кадру рассматриваемого сигнала применяется преобразование Фурье.

Пусть $X_1, X_2, \dots, X_n$ – векторы со значениями исходного сигнала, разделенными на кадры, а функция $f(X)$ – преобразование Фурье кадра X . Тогда $f(X_1), f(X_2), \dots, f(X_n)$ – это комплексные векторы, полученные после преобразования Фурье каждого из кадров исходного сигнала. Можно представить совокупность преобразований всех кадров в виде матрицы комплексных чисел $F = f_1, \dots, f_n$ с размерностью (n, m) , где $f_i = f(X_i)$.

Обозначим $\bar{F}$ как матрицу комплексных чисел размерностью (n, m) , которая является дополнением к матрице F , необходимой для восстановления сигнала. Тогда $F_r = F + \bar{F}$, где F_r – это матрица значений преобразования Фурье для каждого кадра восстановленного сигнала. Чтобы получить сам восстановленный аудиосигнал, к F_r следует применить обратное преобразование Фурье.

Предполагается, что $\bar{F}$ имеет некоторую функциональную зависимость от F . Опишем функцию зависимости дополнения к F от исходного искаженного F как $Z: F \rightarrow \bar{F}$. В рамках данной работы будем предполагать, что такая зависимость Z существует. Тогда решение задачи

восстановления звукового сигнала F_r сводится к выявлению зависимости Z . Исходя из выше-сказанного, можно попытаться выявить эту сложную функциональную зависимость с помощью нейронных сетей.

Переходя к конкретной задаче, предположим, что k -й частотный пул t -го кадра $a_{k,t}$ зависит от всех других частотных пулов всех кадров,

$$a_{k,t} = \sum_{i=0}^n \sum_{j=0}^m G_{i,j,k,t}(a_{i,j}) + \varepsilon,$$

где $G_{i,j,k,t}$ – некоторая неизвестная нелинейная функция, описывающая влияние $a_{i,j}$ на $a_{k,t}$, а ε – случайная составляющая (это значение должно быть довольно маленьким и описывать некоторые недостаточные отклонения в представленной зависимости). Поскольку функции $G_{i,j,k,t}$ неизвестны, их логично аппроксимировать с помощью нейронных сетей.

3. Исследование особенностей восстановления аудиосигнала

Работа с исследованием аудиосигнала имеет ряд особенностей: задачу восстановления аудиосигнала следует рассматривать как задачу линейной регрессии комплексных чисел, размер исходных файлов достаточно велик для быстрой обработки, следует рассматривать каждый канал звукового файла отдельно для более точной обработки.

3.1. Задача линейной регрессии комплексных чисел

При записи звукового сигнала с помощью микрофона, он улавливается и преобразуется в электрический сигнал. Затем преобразованный сигнал посылается звуковой карте на компьютере, которая преобразует сигнал в числа, эти числа называются сэмплами. Сэмплы представляют собой информацию о том, как записанный сигнал звучит в определённые моменты времени. Например, для создания цифрового аудиосигнала, качеством схожим с обычной записью на компакт-диске, должна поступать информация о 44 100 сэмплах в секунду. Количество сэмплов, полученных в секунду, называется частотой сэмплирования или частотой дискретизации. Размер каждого отдельного сэмпла в свою очередь тоже влияет на качество записываемого звука, этот размер называется разрешением. Чем больше разрешение, тем выше качество звука. Для цифровой звукозаписи с таким же качеством, как на компакт-диске, каждый сэмпл будет размером 16 бит.

После считывания аудиосигнала данные представляют собой числа типа `int16`. Для получения более конкретных характеристик для последующей обработки сигнала необходимо знать фазу и амплитуду сигнала на конкретном частотном интервале в конкретный период времени, для этого необходимо, чтобы сигнал подвергся преобразованию Фурье. Данное преобразование возвращает аудиосигнал, представляющий собой данные типа `complex128`, амплитуда сигнала определяется как модуль комплексного числа, фаза – как аргумент.

Таким образом, задача восстановления сигнала сводится к следующей: имеются числовые данные в виде комплексных чисел, нужно восстановить число, показывающее характеристику сигнала на заданной частоте в определенный момент времени. То есть задача линейной регрессии в данном случае представляет собой линейную регрессию комплексных чисел.

3.2. Объем входных данных

Исходные данные представлены в формате WAV (WAVE, WAV, от англ. waveform – «в форме волны»), использование данного формата аудиосигнала позволяет использовать данные в несжатом виде и без потерь, в то время как более привычный для пользователя формат

MP 3 содержит сжатые данные с потерями, что позволяет хранить ту же самую аудиоинформацию в меньшем размере файла.

Использование данных в несжатом формате подразумевает работу с большим объёмом данных. Каждый аудиосигнал имеет следующие характеристики: количество каналов, частота дискретизации (показывает, сколько раз в секунду производится выборка аудио), разрешение (количество бит, используемых для кодирования каждого значения выборки) и продолжительность. Для хранения одной многоканальной стереозаписи длительностью 3 минуты с частотой дискретизации 44,1 кГц и разрешением 16 бит потребуется порядка 30 Мбайт. Для качественного обучения моделей глубокого машинного обучения могут потребоваться десятки и сотни аудиозаписей, что при сложных архитектурах моделей будет означать большое количество времени, требуемое для обработки информации.

3.3. Многоканальная аудиоинформация

Зачастую используется стерео аудиосигнал. Стереофонический звук состоит из двух уникальных каналов. При прослушивании это позволяет получить объёмное звучание, при котором есть ощущение направленности звука и его расположения. На практике же стереозвук представляет собой сигнал, что для записи которого используются два микрофона, каждый из которых улавливает немного отличающийся сигнал.

Некоторые кодеки фактически разделяют левый и правый каналы, сохраняя их в отдельных блоках в своей структуре данных. Таким образом, при исследовании стереосигнала мы имеем два моносигнала с небольшими различиями. Это приводит к тому, что, если необходимо получить более точное предсказание стереосигнала, нужно рассматривать и обрабатывать обе его составляющие отдельно.

4. Особенности реализации

Частная задача восстановления сигнала – восстановление партии ударника в аудиосигнале (музыкальном произведении). Актуальность и сложность исследования заключается в том, что частотные диапазоны ударных на фоне всего частотного разложения сигнала трудно определить, поэтому процесс восстановления должен основываться на всем исходном частотном диапазоне.

Создание партии ударных – это уникальный процесс, в котором человек придумывает ритмический рисунок песни, но он так или иначе будет соответствовать некоторым ритмическим законам и будет зависеть от других инструментов (или, наоборот, партии инструментов будут зависеть от партии ударника). Таким образом, весь частотный диапазон аудиосигнала несет в себе значимые признаки для последующего восстановления барабанной партии, которая является дополнением к исходному сигналу без нее.

Исходные данные в данной работе – аудиосигнал с удаленной барабанной партией и аудиосигнал с присутствующей партией.

Различия оригинального аудиосигнала от того же аудиосигнала с удаленной барабанной партией в частотных интервалах можно наблюдать после применения к каждому сигналу разложения Фурье. На рис. 1а представлена спектрограмма композиции исполнителя 30 Seconds to Mars, на рис. 1б – спектрограмма той же композиции с удаленной барабанной партией. Можно наблюдать, что в первом случае при наличии барабанной партии некоторые изображения частотных интервалов более выраженные, яркие, полный частотный спектр аудиосигнала выглядит более дополненным, чем второй случай с удаленной барабанной партией.

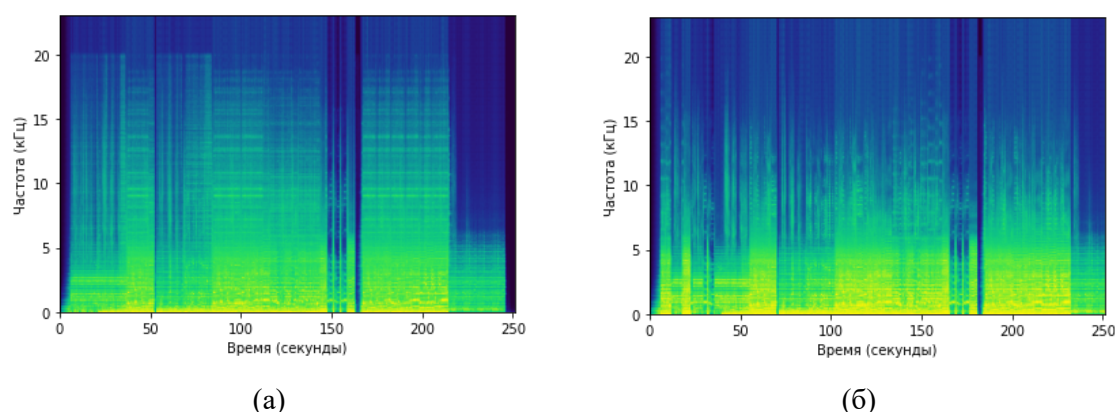


Рис. 1. Спектрограмма композиции исполнителя 30 Seconds to Mars оригинал (а) и с удаленной барабанной партией (б)

Fig. 1. Spectrogram of the composition of the artist 30 Seconds to Mars original (a) and without a drum part (б)

Изначально аудиосигналы поступают на обработку в формате WAV. После считывания в данном формате данные представляют собой массив чисел типа `int16`. Затем необходимо разделить исходный стереосигнал на два моносигнала, чтобы обеспечить детальную обработку каждой составляющей сигнала отдельно, и произвести преобразование Фурье для каждого канала сигнала. Во время преобразования следует обратиться к нормализации данных.

Нормализация данных – это операция преобразования входной информации (признаков), которая выполняется на этапе подготовки данных (при генерации признаков) в машинном обучении, при которой значения признаков приводятся к некоторому заданному диапазону. После нормализации все числовые значения входных признаков приведены к одинаковой области их изменения – некоторому узкому диапазону, что обеспечит корректную работу вычислительных алгоритмов. На рис. 2 показаны частотные спектры аудио-сигнала после преобразования Фурье, при создании которого использовалась (а) и не использовалась нормализации (б).

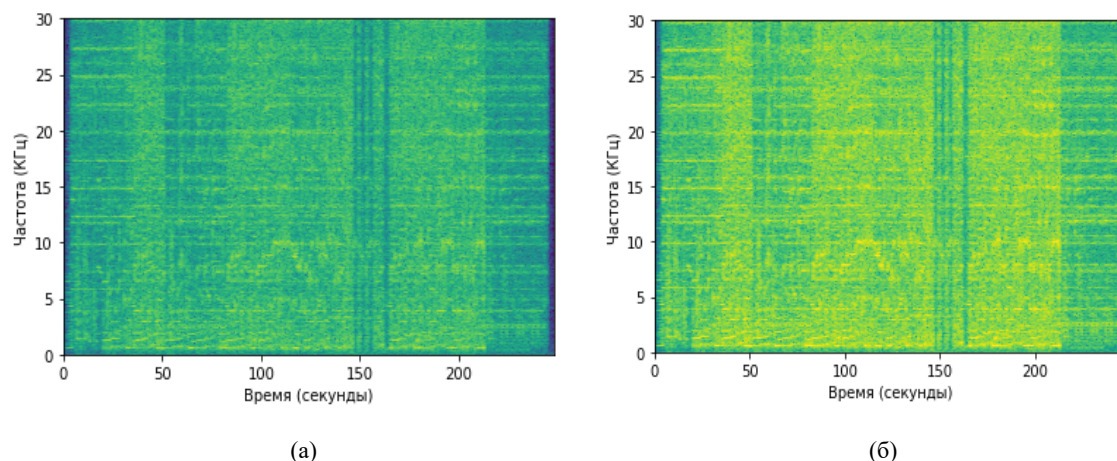


Рис 2. Спектрограмма аудиосигнала после преобразования Фурье, при создании которого использовалась (а) и не использовалась нормализации (б)

Fig. 2. Spectrogram of the audio signal after the Fourier transform, which was created using (a) and not using normalization (б)

После преобразований Фурье данные представляют собой многомерный массив размером $(512, n, 2)$, где n – количество временных интервалов (кадров) аудиосигнала, а 2 – показатель того, что комплексные числа рассматриваются как реальная и мнимая часть отдельным значением.

Задачу восстановления удаленной части аудиосигнала предлагается решать с помощью свёрточной нейронной сети. Данный тип методов глубокого обучения является одним из самых популярных на данный момент, а активное применение получил, начиная с 2012 года. Главное преимущество свёрточных нейронных сетей состоит в том, что они автоматически извлекают нужные признаки и коэффициенты без какого-либо человеческого наблюдения. Также эффективны в вычислениях, в них используются специальные операции свертки и объединения, а также выполняются объединенное использование параметров.

Для восстановления одного кадра рассматриваются 25 последовательных кадров (предположим, что этого количества достаточно, чтобы установить все сложные зависимости во времени и между разными частотными диапазонами). Итак, из 25 кадров восстанавливаем 1 кадр посередине. Первоначальное решение заключалось в восстановлении всех 512 частотных диапазонов входного аудиосигнала с помощью одной нейронной сети. В данном случае использовалась архитектура свёрточной нейронной сети, которая восстанавливает (за счет реализованного скользящего окна с шагом в один временной интервал) весь частотный диапазон на одном конкретном временном интервале, работая с размерами (512, 25, 2) на входе и (512, 1, 2) на выходе.

Данная архитектура сети имеет около 336 миллионов параметров, поэтому как обучить, так и работать с такой сетью – довольно трудоемкая задача. Тем не менее, получить качественные результаты не удалось, поскольку сеть создавала на выходе зашумленную версию исходного аудиосигнала. Для преодоления проблемы было решено изменить конструкцию нейронной сети, добавив на выходе слой, содержащий исходный сигнал, соответствующий результирующему кадру. Также было решено сократить количество используемых параметров до ~ 4 миллионов. Оказалось, что результаты становятся близкими к предыдущему эксперименту – сеть генерировала зашумленный исходный сигнал без ожидаемых изменений.

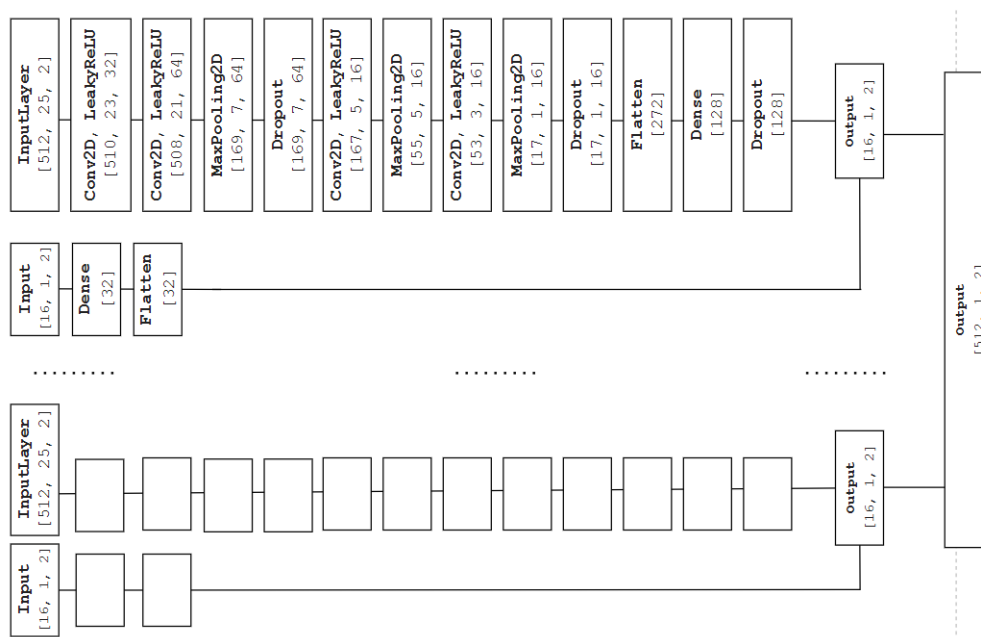


Рис 3. Схема объединения нейронных сетей, разработанных на непересекающихся частотных поддиапазонах

Fig. 3. Scheme for combining neural networks developed on non-overlapping frequency subranges

После получения представленных выше результатов архитектура была изменена таким образом, чтобы количество параметров было уменьшено до 70 тысяч (см. рис.3). Кроме того,

внесены следующие изменения: теперь весь частотный диапазон из 512 полос делится на части, и каждая эта часть обрабатывается отдельной нейронной сетью. Таким образом, каждая нейронная сеть обрабатывает только свою часть всего частотного диапазона. Предполагалось, что такой подход должен повысить точность восстановления. Если количество нейронных сетей равно 32, – то каждая сеть на выходе восстанавливает 16 полос из 512. Важно отметить, что на вход подаются все 512 частотных полос, но каждая сеть в данном случае восстанавливает только 16 поло. Затем результат работы всех 32 нейронных сетей складывается для получения всего частотного диапазона из 512 полос. Конечный этап – сложить полученный результат с исходным сигналом. Таким образом, выполняется не чистая модификация сигнала, а сложение исходного сигнала с восстановленными недостающими элементами.

При разработке архитектуры нейронной сети в качестве слоя активации был выбран LeakyRelu. В общем случае функция активации определяет выходное значение нейрона в зависимости от результата взвешенной суммы входов и порогового значения. Входные данные, поступившие на слой, умножаются на веса полносвязного или свёрточного слоя, а результат передаётся в функцию активации. Выбранная функция LeakyRelu имеет преимущество относительно большинства подобных функций активации – отрицательные значения обрабатываются благодаря небольшому наклону в левой полуплоскости. Математически LeakyRelu объяснена следующим образом:

$$f(x) = \begin{cases} 0.01x, & x < 0 \\ x, & x \geq 0 \end{cases}$$

В качестве функции потерь применена функция среднеквадратичного отклонения. Функция потерь в теории статистических решений характеризует потери при неправильном принятии решений на основе наблюдаемых данных. Другими словами, использование функции потерь позволяет получить оценку того, насколько хорошо алгоритм моделирует набор данных – если прогнозы полностью неверны, функция потерь выдаст большее число, если они достаточно хороши – будет выведено меньшее число. Чтобы вычислить среднеквадратичную ошибку необходимо вычислить разницу между предсказанными и исходными данными, возвести ее в квадрат и усреднить по всему набору данных. Цель использования функции потерь – отследить и минимизировать это значение, чтобы получить лучшую линию, проходящую через все точки при рассмотрении данных на координатной плоскости.

В процессе компилирования модели нейронной сети используется оптимизатор. Оптимизаторы – это алгоритмы или методы, используемые для изменения атрибутов нейронной сети, таких как веса и скорость обучения, с целью уменьшения потерь [8]. В рамках данного исследования использовался оптимизатор Adam [9] – адаптивный алгоритм оптимизации скорости обучения.

В качестве дополнительного метода улучшения результатов можно рассмотреть возможность увеличения количества нейронных сетей, а, следовательно, уменьшения частотного диапазона, обрабатываемого каждой сетью (например, использование 64 сетей и, соответственно, 8 диапазонов частот для обрабатываемого диапазона 512). В результате полученные частотные дополнения исходного аудиосигнала становятся более выраженными, значения восстановленного аудиосигнала приближаются к оригинальному (исходному) из-за того, что коэффициенты сети становятся настроенными на более конкретный набор неизвестных.

5. Анализ результатов

На рис. 4 и рис. 6–9 показаны спектрограммы исходного звукового сигнала (с барабанной частью) и звукового сигнала, полученного с помощью нейронной сети, соответственно. На этих рисунках можно даже визуально заметить схожесть исходного и восстановленного

сигнала, что является показателем качества метода. Кроме того, для лучшего сравнения на рис. 5 приведена аналогичная спектрограмма звукового сигнала с удаленной барабанной частью.

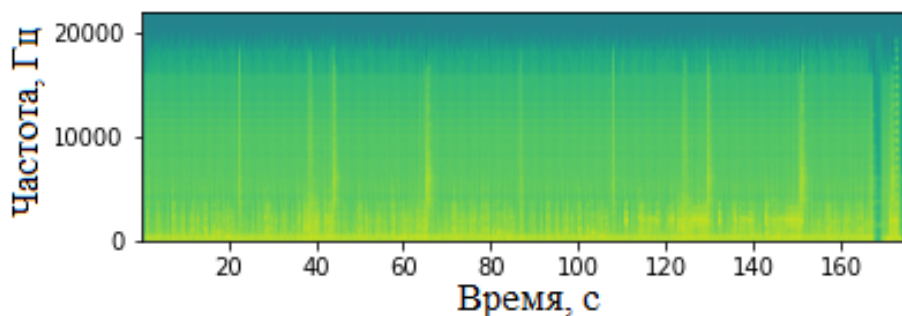


Рис 4. Спектрограмма исходного аудиосигнала с присутствующей барабанной партией

Fig. 4. Spectrogram of the original audio signal with drum part

Для улучшения результатов было увеличено количество нейронных сетей, а значит, уменьшен частотный диапазон, обрабатываемый каждой сетью (например, используя 64 сети и, соответственно, 8 частотных диапазонов из 512 для каждой сети). Полученные частотные дополнения исходного аудиосигнала становятся более выраженными, восстановленный аудиосигнал становится ближе к оригиналу из-за того, что коэффициенты сети становятся настроенными на более конкретный набор неизвестных. На рис. 6–9 показаны спектрограммы исходного аудиосигнала (с барабанной частью) и аудиосигнала, полученного с помощью нейронной сети с 16, 32, 64, 128 сетями соответственно. На этих рисунках можно даже визуально заметить схожесть исходного и восстановленного сигнала, что является показателем качества метода.

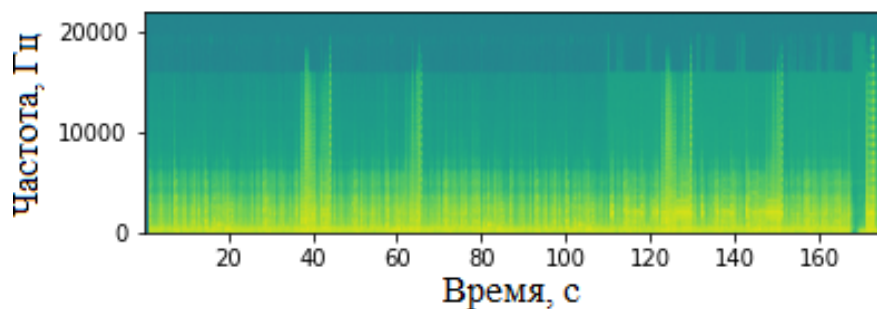


Рис 5. Спектрограмма исходного аудиосигнала с удалённой барабанной партией

Fig. 5. Spectrogram of the original audio signal without drum part

Таким образом, как видно на рис. 6–9, при увеличении количества нейронных сетей результат работы приближается к ожидаемому, но при этом увеличивается уровень шума аудиосигнала.

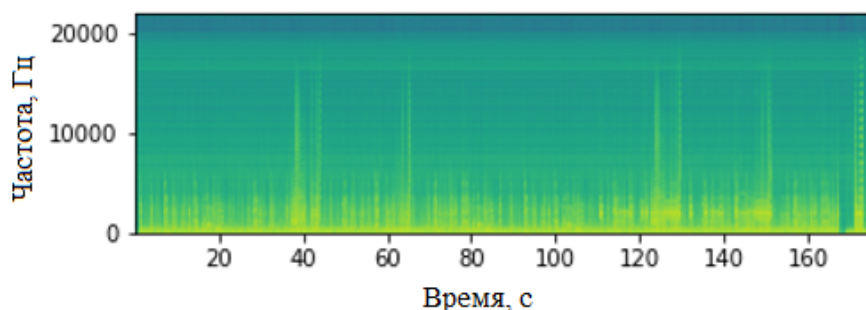


Рис 6. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 16 нейронных сетей

Fig. 6. Spectrogram of the resulting audio signal with the reconstructed drum part based on 16 neural networks

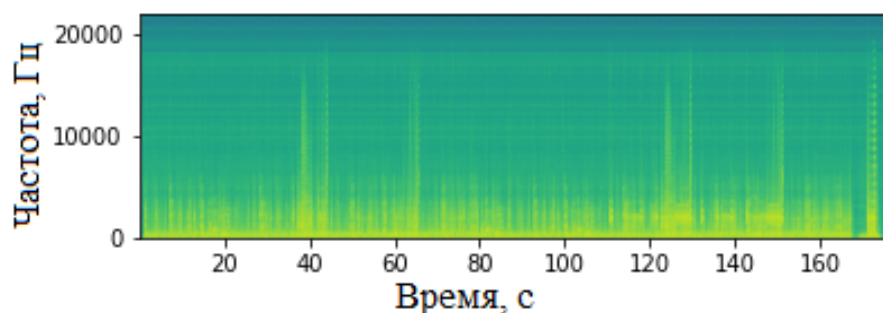


Рис 7. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 32 нейронных сетей

Fig. 7. Spectrogram of the resulting audio signal with the reconstructed drum part based on 32 neural networks

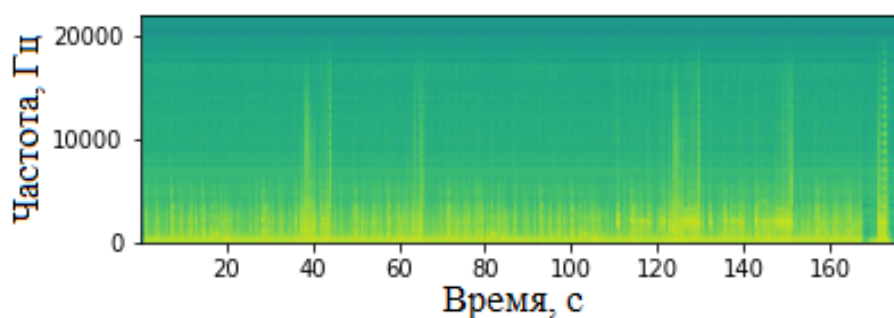


Рис 8. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 64 нейронных сетей

Fig. 8. Spectrogram of the resulting audio signal with the reconstructed drum part based on 64 neural networks

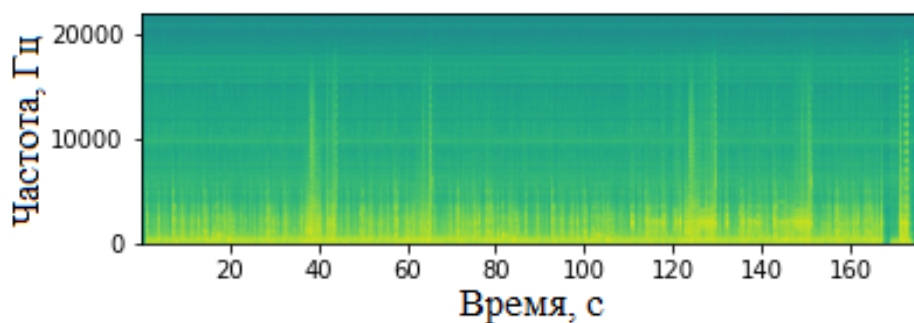


Рис 9. Спектрограмма результирующего аудиосигнала с восстановленной барабанной партией на основе 128 нейронных сетей

Fig. 9. Spectrogram of the resulting audio signal with the reconstructed drum part based on 128 neural networks

Для оценки эффективности применения исследуемых методов для поставленной задачи предлагается исследовать в совокупности несколько представлений результатов: спектрограммы восстановленных аудиосигналов, значения метрик MSE, MAE, и median absolute error (см. табл.), а также результат в аудиоформате. Анализируя полученные данные, можно прийти к выводу, что лучший результат получен при количестве нейронных сетей равном 32. Значения только метрики в нашем случае не информативно для оценки результата предсказания, так

как неясно, появились ли в аудиосигнале барабаны или возник кадровый шум. Именно поэтому, чтобы сделать вывод об эффективности используемых методов, следует принимать все представления результатов в совокупности.

Таким образом, как видно на рис. 6–9, при увеличении количества нейронных сетей результат работы приближается к ожидаемому, но при этом увеличивается уровень шума аудиосигнала.

Метрики для валидации результата
Metrics for result validation

	16	32	64	128
mean squared error	0,017456	0,003772	0,009480	0,020648
mean absolute error	0,032182	0,014861	0,023788	0,036868
median absolute error	0,003299	0,003031	0,003105	0,004277

Заключение

Полученные в ходе работы результаты позволяют утверждать, что при наличии сигнала с искажением или с удаленной частью существует возможность построить сложную регрессионную модель для решения задачи восстановления оригинального сигнала при помощи сверточных нейронных сетей. В предыдущих разделах показано, что существует взаимосвязь между поврежденными и оставшимися данными, которую можно аппроксимировать для восстановления сигнала.

В результате была разработана архитектура составной нейронной сети, состоящей из нескольких более простых сверточных сетей, количество которых может варьироваться. На основании полученных результатов можно утверждать, что при увеличении количества нейронных сетей в рамках решения той же задачи эффективность восстановления сигнала также увеличивается. Более того, полученные результаты показали высокий потенциал использования сверточных нейронных сетей для решения такой сложной задачи, как восстановление аудиосигнала.

Список литературы

1. **Mokry O., Rajmic P.** Audio inpainting: Revisited and reweighted //IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020. Vol. 28. Pp. 2906–2918.
2. **Heo H. S., So B. M., Yang I. H., Yoon S. H., Yu, H. J.** Automated recovery of damaged audio files using deep neural networks //Digital Investigation. 2019. Vol. 30. Pp. 117–126.
3. **Kwon B., Gong M., Huh J. and Lee S.** Identification and Restoration of LZ77 Compressed Data Using a Machine Learning Approach. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018, pp. 1787–1790. DOI: 10.23919/APSIPA.2018.8659755
4. **Mousavi A., Patel A. B., Baraniuk R. G.** A deep learning approach to structured signal recovery. 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, 2015, pp. 1336–1343. DOI: 10.1109/ALLERTON.2015.7447163
5. **Zhang Z., Rao B. D.** Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning. IEEE Journal of Selected Topics in Signal Processing. 2011. Vol. 5, no. 5. Pp. 912–926. DOI: 10.1109/JSTSP.2011.2159773

6. **Bellasi D., Maechler P., Burg A., Felber N., Kaeslin H., Studer C.** Live demonstration: Real-time audio restoration using sparse signal recovery. 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, 2013, pp. 659–659. DOI: 10.1109/ISCAS.2013.6571929.
7. **Adler A. et al.** Audio inpainting //IEEE Transactions on Audio, Speech, and Language Processing. 2011. Vol. 20, no. 3. Pp. 922–932.
8. **Doshi S.** Various Optimization Algorithms For Training Neural Network. Towards Data Science. 2019 Jan. 13.
9. **Diederik P. Kingma and Jimmy Lei Ba.** Adam : A method for stochastic optimization. 2014. arXiv:1412.6980v9

References

1. **Mokrý O., Rajmic P.** Audio inpainting: Revisited and reweighted //IEEE/ACM Transactions on Audio, Speech, and Language Processing. 2020. Vol. 28. Pp. 2906–2918.
2. **Heo H. S., So B. M., Yang I. H., Yoon S. H., Yu H. J.** Automated recovery of damaged audio files using deep neural networks // Digital Investigation. 2019. Vol. 30. Pp. 117–126.
3. **Kwon B., Gong M., Huh J., Lee S.** Identification and Restoration of LZ77 Compressed Data Using a Machine Learning Approach. 2018 Asia-Pacific Signal and Information Processing Association Annual Summit and Conference (APSIPA ASC), Honolulu, HI, USA, 2018, pp. 1787–1790. DOI: 10.23919/APSIPA.2018.8659755
4. **Mousavi A., Patel A. B., Baraniuk R. G.** A deep learning approach to structured signal recovery. 2015 53rd Annual Allerton Conference on Communication, Control, and Computing (Allerton), Monticello, IL, 2015, pp. 1336–1343. DOI: 10.1109/ALLERTON.2015.7447163.
5. **Zhang Z., Rao B. D.** Sparse Signal Recovery with Temporally Correlated Source Vectors Using Sparse Bayesian Learning. IEEE Journal of Selected Topics in Signal Processing, 2011. Vol. 5, no. 5, pp. 912–926. DOI: 10.1109/JSTSP.2011.2159773.
6. **Bellasi D., Maechler P., Burg A., Felber N., Kaeslin H., Studer C.** Live demonstration: Real-time audio restoration using sparse signal recovery. 2013 IEEE International Symposium on Circuits and Systems (ISCAS), Beijing, 2013, pp. 659–659. DOI: 10.1109/ISCAS.2013.6571929
7. **Adler A. et al.** Audio inpainting //IEEE Transactions on Audio, Speech, and Language Processing, 2011. Vol. 20, no. 3, pp. 922–932.
8. **Doshi S.** Various Optimization Algorithms For Training Neural Network. Towards Data Science. 2019 Jan.13.
9. **Diederik P. Kingma and Jimmy Lei Ba.** Adam: A method for stochastic optimization. 2014. arXiv:1412.6980v9

Информация об авторах

Дементьева Кристина Игоревна, ассистент кафедры прикладной математики и кибернетики, аспирант, Сибирский государственный университет телекоммуникаций и информатики (Новосибирск, Россия)

Антон Андреевич Ракитский, кандидат технических наук, доцент кафедры прикладной математики и кибернетики, Сибирский государственный университет телекоммуникаций и информатики (Новосибирск, Россия); научный сотрудник, Новосибирский государственный университет (Новосибирск, Россия); старший научный сотрудник, Институт вычислительной техники СО РАН (Новосибирск, Россия)

Information about the Authors

Kristina I. Dementyeva, Assistant of the Department of Applied Mathematics and Cybernetics, Postgraduate student, Siberian State University of Telecommunications and Informatics (Novosibirsk, Russian Federation)

Anton A. Rakitskiy, Candidate of Technical Sciences, Associate Professor of the Department of Applied Mathematics and Cybernetics, Siberian State University of Telecommunications and Informatics (Novosibirsk, Russian Federation); Researcher, Novosibirsk State University (Novosibirsk, Russian Federation); Senior Researcher, Institute of Computer Engineering SB RAS (Novosibirsk, Russian Federation)

*Статья поступила в редакцию 10.06.2022;
одобрена после рецензирования 07.10.2022; принята к публикации 07.10.2022
The article was submitted 10.06.2022;
approved after reviewing 07.10.2022; accepted for publication 07.10.2022*

Научная статья

УДК 004.415

DOI 10.25205/1818-7900-2022-20-3-51-64

Автоматизированная система контроля формирования индикаторов достижения компетенций при изучении основ алгебры логики

**Елена Анатольевна Сидорова¹, Анна Владимировна Долгова²,
Светлана Петровна Железняк³**

Омский государственный университет путей сообщения
Омск, Россия

¹armsid@mail.ru, <https://orcid.org/0000-0001-5312-7564>

²9059238462@mail.ru, <https://orcid.org/0000-0002-8201-8823>

³zhsp120866@yandex.ru, <https://orcid.org/0000-0001-8957-2678>

Аннотация

Эффективное изучение дисциплины «Информатика» невозможно без широкого применения информационных технологий. Одним из основополагающих разделов информатики является алгебра логики. Применение традиционных методов тестирования обучающихся для оценки степени освоения этой темы имеет ряд определенных недостатков, что обуславливает необходимость создания других интерактивных средств для этой цели. Статья посвящена особенностям разработки и функционирования автоматизированной системы «Алгебра логики», которая предназначена для контроля формирования индикаторов достижения компетенций, полученных при изучении основ алгебры логики. Разработанная система позволяет автоматически сформировать индивидуальный вариант задания, содержащий задачи трех типов. Первое задание включает в себя построение таблиц истинности для трех логических функций, второе требует построить логическую схему, третье задание ориентировано на упрощение логического выражения. Архитектура системы включает в себя восемь независимых модулей, реализующих синтез нового комплекта заданий и его оформление, декомпозицию полученных логических функций, построение таблиц истинности, анализ и оценку результатов. В статье раскрыта концепция работы каждого модуля. Подробно описана работа модулей синтеза и анализа логических функций, приведены укрупненные графические схемы алгоритмов их функционирования. Рассмотренная автоматизированная система за время эксплуатации зарекомендовала себя положительно, она позволяет объективно оценить теоретические знания пользователей и их умения применять известные способы построения таблиц истинности. К достоинствам системы относятся небольшой объем занимаемой на диске памяти, возможность функционирования в компьютерных сетях и на локальных компьютерах, интуитивно понятный интерфейс, разнообразие генерируемых заданий и возможность автоматической проверки результатов их выполнения.

Ключевые слова

автоматизированная система, алгебра логики, модуль, интерфейс, графическая схема алгоритма

Для цитирования

Сидорова Е. А., Долгова А. В., Железняк С. П. Автоматизированная система контроля формирования индикаторов достижения компетенций при изучении основ алгебры логики // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 51–64. DOI 10.25205/1818-7900-2022-20-3-51-64

Automated System of Competencies Indicators' Formation Control during the Basic Boolean Algebra Studying

Elena A. Sidorova¹, Anna V. Dolgova²,
Svetlana P. Zheleznyak³

Omsk State Transport University
Omsk, Russian Federation

¹armsid@mail.ru, <https://orcid.org/0000-0001-5312-7564>

²9059238462@mail.ru, <https://orcid.org/0000-0002-8201-8823>

³zhsp120866@yandex.ru, <https://orcid.org/0000-0001-8957-2678>

Abstract

The effective study of Computer Science is impossible without the widespread use of information technology. Boolean algebra is one of the most fundamental sections of computer science. There is a list of disadvantages of this testing method for the evaluation of the master's degree students' knowledge of this topic. This results in development of other interactive tools. This article is devoted to the development and operation features of the automated system "Boolean algebra", which was designed to control the formation of the competencies' achievement indicators obtained during studying Boolean algebra. The developed system automatically creates task option and contains three types of tasks. The first task includes construction of a truth table for three Boolean functions, the second requires Boolean circuit construction, and the third task is focused on Boolean expression simplifying. The system architecture consists of eight independent modules which realize the new task option synthesis and its design, Boolean functions decomposition, truth tables construction, analysis, and evaluation of the results. The article describes the concept of each module. It describes in detail the synthesis of Boolean functions, as well as the analysis modules and their functioning algorithms. The automated system under operation proved to be effective since it allows one to objectively evaluate user's theoretical knowledge and ability to apply these methods to the truth tables construction. There are also several advantages of the considered automated system: it requires little space on a hard disk, works in a network and on local computer, has intuitive interface, generates a large variety of tasks and performs their autochecking.

Keywords

automated system, Boolean algebra, module, interface, algorithm

For citation

Sidorova E. A., Dolgova A. V., Zheleznyak S. P. Automated System of Competencies Indicators' Formation Control during the Basic Boolean Algebra Studying. Vestnik NSU. Series: Information Technologies, 2022, vol. 20, no. 3, pp. 51–64. DOI 10.25205/1818-7900-2022-20-3-51-64

Введение

Информационные технологии (ИТ) являются неотъемлемой частью современного образовательного процесса. Их активное внедрение в учебных заведениях регламентируется Федеральным законом «Об образовании в Российской Федерации» от 29.12.2012 № 273-ФЗ [1], федеральными государственными образовательными стандартами высшего образования (ФГОС ВО) последнего поколения, ГОСТ Р 55751-2013 «Информационно-коммуникационные технологии в образовании. Электронные учебно-методические комплексы. Требования и характеристики» и другими документами. Широкое внедрение и эффективное использование ИТ во всех сферах деятельности предусматривает Единый план по достижению национальных целей развития Российской Федерации на период до 2024 года и на плановый период до 2030 года [2].

При реализации образовательных программ бакалавриата и специалитета в Омском государственном университете путей сообщения (ОмГУПС) кафедра «Информатика и компьютерная графика» (далее – кафедра) осуществляет образовательную деятельность более чем по 70 учебным планам, каждый из которых предусматривает изучение информатики.

На сегодняшний день ФГОС ВО не предъявляет конкретных требований к содержанию дисциплины «Информатика», она может включать в себя любые компоненты, обеспечивающие формирование у студентов необходимых для будущей профессиональной деятельности индикаторов достижения компетенций (ИДК). Например, при подготовке инженеров по специ-

альностям 23.05.03 «Подвижной состав железных дорог» и 23.05.05 «Системы обеспечения движения поездов» должны формироваться следующие ИДК: «УК-1.3: осуществляет систематизацию информации различных типов для анализа проблемных ситуаций. Вырабатывает стратегию действий для построения алгоритмов решения поставленных задач», «УК-1.4: владеет навыками программирования разработанных алгоритмов и критического анализа полученных результатов». В обоих случаях содержание ИДК предусматривает формирование у обучающегося знаний, умений и навыков для оценки исходной информации, корректной ее обработки и анализа полученных результатов. Для достижения этих целей при изучении информатики предусмотрено освоение раздела «Алгебра логики» (далее – АЛ), который является одним из основополагающих в теоретических основах информатики и неразрывно связан с разделами «Кодирование и измерение информации», «Логические основы построения компьютера», «Алгоритмизация и программирование». В учебном процессе по указанным специальностям изучение АЛ продолжается в дисциплинах «Математическое моделирование систем и процессов», «Электроника», «Теория дискретных устройств» и других.

Постановка задачи

Согласно имеющимся данным в Российской Федерации в 2021 году единый государственный экзамен (ЕГЭ) по информатике сдавали 94 тыс. человек, что составляет около 14 % выпускников учебных заведений. При этом из 1092 абитуриентов, зачисленных в ОмГУПС на технические направления подготовки и специальности, где по выбору в числе вступительных испытаний можно было представить итоги ЕГЭ по информатике, сделали это только 46 человек, что составляет всего 4 % от общего числа поступивших. Приведенные данные, а также результаты регулярных опросов студентов свидетельствуют о том, что значительное количество обучающихся имеют сложности при изучении информатики.

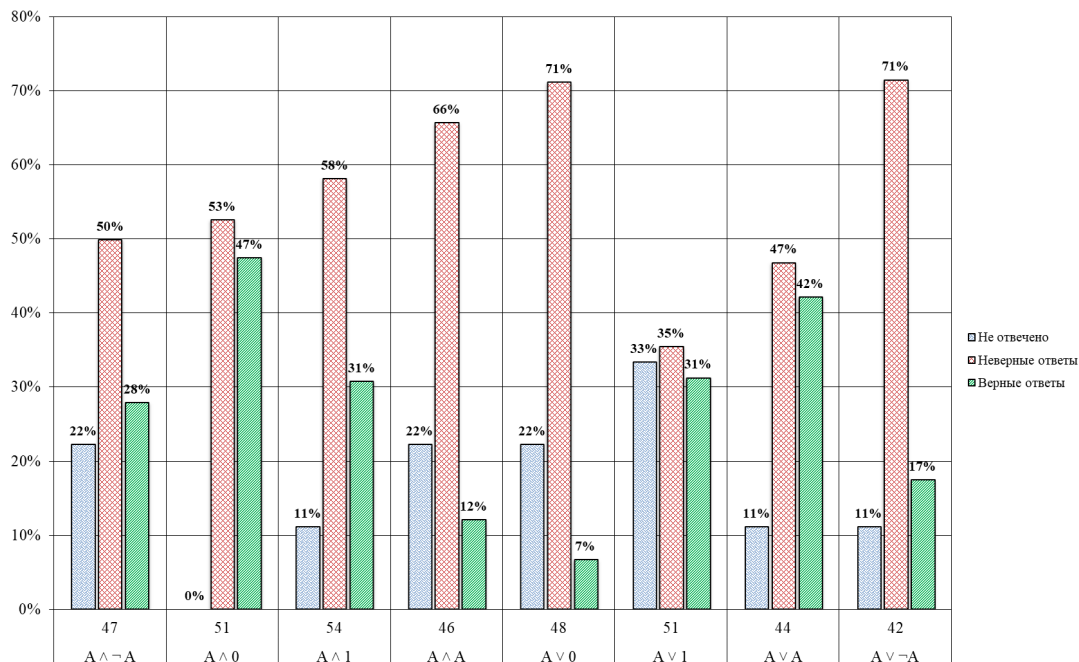


Рис. 1. Результаты выполнения тестовых заданий по АЛ при входном контроле студентов первого курса
Fig. 1. The results of Boolean algebra test tasks at the first-year students entrance control

В качестве входного контроля базовой подготовки студентов первого курса по информатике на кафедре проводится специальное компьютерное тестирование остаточных знаний, в которое включены и вопросы по АЛ, ориентированные на определение степени усвоения выпускниками школ простейших логических операций [3]. В 2021–2022 учебном году в таком тестировании приняли участие 590 человек. Формирование набора тестовых вопросов осуществлялось методом экспертных оценок. Он включает в себя задания для проверки знания закона идемпотентности, операций переменной с ее инверсией, операций с константами 0 и 1. Статистика результатов выполнения этих заданий представлена на рис. 1, где по оси абсцисс указан конкретный тип вопроса и общее число студентов, которым этот вопрос был представлен в сеансе тестирования. Высокий процент неотвеченных вопросов может быть обусловлен нехваткой времени или сложностью других вопросов в тесте. Анализ этих данных свидетельствует об очень низком уровне подготовки студентов первого курса по АЛ, что в свою очередь обуславливает возникновение трудностей в изучении не только информатики, но и связанных с ней дисциплин.

Входной контроль остаточных знаний и последующий текущий контроль успеваемости студентов осуществляется с помощью разработанной на кафедре и зарегистрированной в Реестре программ для ЭВМ «Универсальной тестово-обучающей программы контроля профессиональных знаний Test Master 2007 (версия 1.0)» [4]. В этой системе тестирования индивидуальный тестовый сеанс для каждого пользователя формируется на основе обширной базы данных по специальному алгоритму, обеспечивающему заданную тематическую структуру и пропорциональное наличие вопросов разного типа и сложности. База данных системы включает в себя свыше 70 тестовых вопросов по АЛ, детализированных по темам: «Высказывания и законы», «Логические выражения», «Таблицы истинности», «Логические схемы». Индивидуальный тестовый сеанс по АЛ содержит 10 вопросов, в полном объеме охватывающих изученный материал, на его выполнение отводится 20 минут. Выполнение теста позволяет определить результат освоения компетенций по дисциплине в рамках рассматриваемой темы, оцениваемый с помощью соответствующих ИДК. Результаты тестирования студентов по АЛ за период с 2013 по 2022 годы (рис. 2) показывают, что итоговый средний балл в целом превышает пороговый уровень 60 %, но при этом доля неуспевающих студентов остается достаточно высокой и составляет около 35 %.

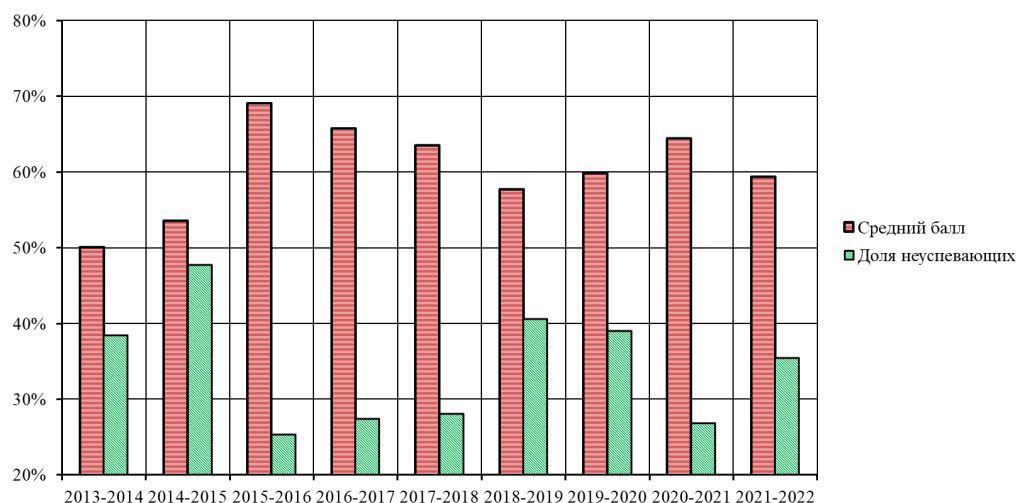


Рис. 2. Результаты текущего контроля успеваемости по АЛ за период с 2013 по 2022 г.
Fig. 2. The results of Boolean algebra academic performance current monitoring from 2013 to 2022

Для оценки уровня самоподготовки студентов к контрольному тестированию на кафедре организовано тренировочное тестирование [5]. Студент может проходить его многократно и по окончании каждого сеанса анализировать вопросы, на которые он дал верные или неверные ответы. Несмотря на большое разнообразие вопросов в базе данных, их количество ограничено. Это приводит к тому, что при тренировочном тестировании студенты зачастую не применяют для решения поставленных задач известные им алгоритмы, а стремятся лишь запомнить верные ответы. Кроме того, при тестировании большинство вопросов обычно содержат варианты возможных ответов, что не позволяет в полной мере оценить реальные знания обучающихся по причине наличия элемента случайности при выборе ими правильного ответа [6–8].

Для повышения эффективности учебного процесса и устранения недостатков тестирования при текущем контроле знаний предпочтительнее применять контрольную работу. Однако она, как правило, выполняется студентами письменно, что существенно увеличивает нагрузку на преподавателя в части проверки. Таким образом, актуальной становится задача создания такого программного продукта для обучения и аттестации студентов, который позволит сформировать индивидуальный вариант задания, оценить правильность применения студентами известных алгоритмов и при этом предоставить им возможность предварительно ознакомиться с типовыми заданиями и технологией работы с ними.

Архитектура автоматизированной системы

На основе накопленного педагогического опыта, анализа научной и учебно-методической литературы, изучения современных образовательных технологий и способов разработки электронных учебных тренажеров [9; 10] на кафедре разработана и внедрена в учебный процесс автоматизированная система «Алгебра логики» (АСАЛ), предназначенная для контроля формирования индикаторов достижения компетенций, полученных при изучении основ алгебры логики.

Согласно статистическим данным, полученным онлайн-инструментом статистики посетителей Statcounter, за период с февраля 2015 г. по март 2022 г. около 75 % всех стационарных компьютеров в мире и около 88 % в Российской Федерации используют операционную систему Windows¹. На компьютерах в локальной сети ОмГУПС также установлена лицензионная операционная система Windows, поэтому она выбрана в качестве платформы для разработки АСАЛ.

Программная реализация АСАЛ выполнена на языке программирования Visual Basic for Applications (VBA), выбор которого обусловлен его широким распространением (в рейтинге популярности языков программирования PYPL² по состоянию на март 2022 г. он занимает 16 место) и удобством применения для решаемой задачи. С целью обеспечения возможности автономной работы АСАЛ без дополнительного программного обеспечения как в локальной вычислительной сети ОмГУПС, так и на домашних компьютерах студентов, система реализована в среде Microsoft Excel в виде файла с расширением .xlsm.

АСАЛ формирует индивидуальный вариант задания (ИВЗ), включающий в себя три задачи разных типов. Первое задание ориентировано на построение таблиц истинности трех логических функций, второе задание требует построить логическую схему, третье задание заключается в упрощении логического выражения. Для реализации этих задач архитектура АСАЛ включает в себя восемь модулей (рис. 3).

¹ Desktop Operating System Market Share Worldwide // StatCounter. URL: <https://gs.statcounter.com/os-market-share/desktop/worldwide> (дата обращения: 24.04.2022).

² Popularity of Programming Language // PYPL. URL: <https://pypl.github.io/PYPL.html> (дата обращения: 24.04.2022).

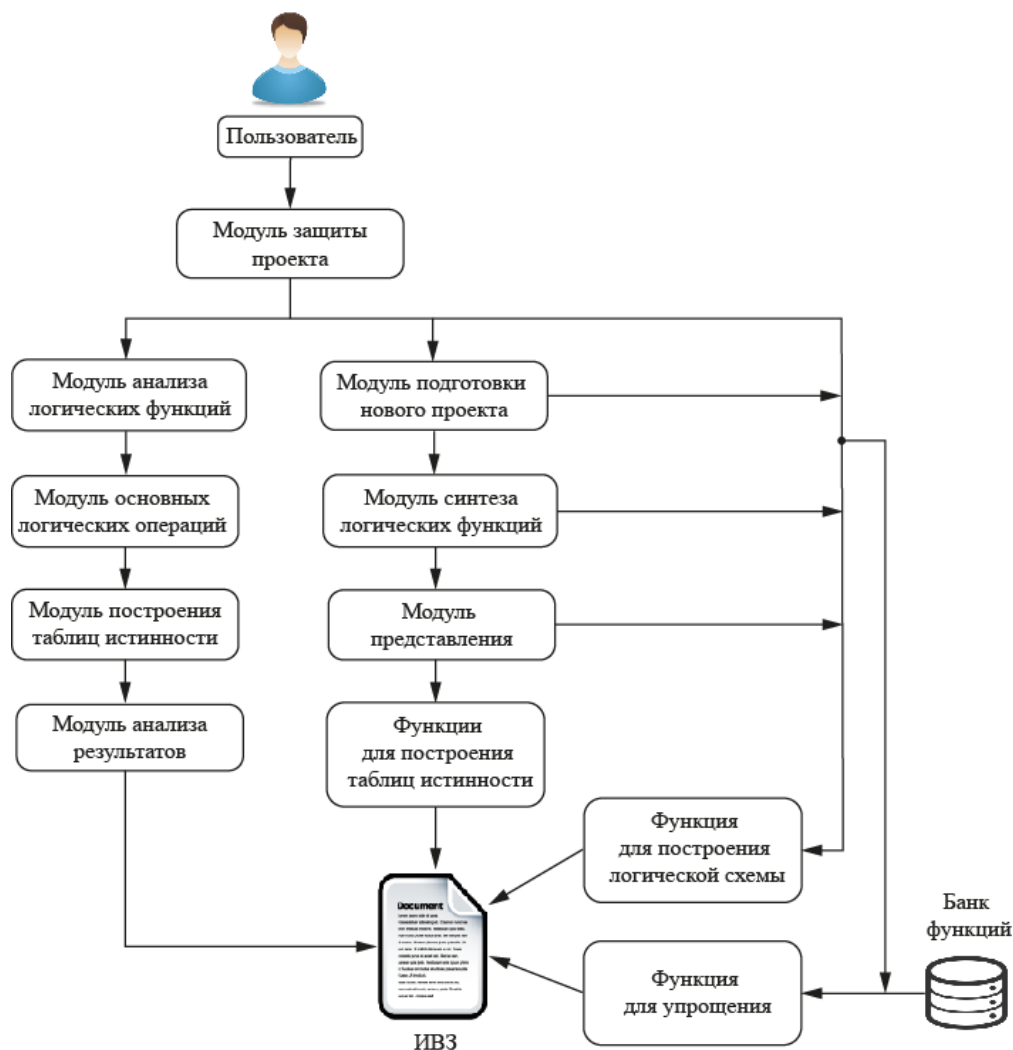


Рис. 3. Архитектура АСАЛ

Fig. 3. Automated system "Boolean Algebra" architecture

1. Модуль защиты проекта отвечает за защиту рабочей книги по время функционирования программы.
2. Модуль подготовки нового проекта предназначен для очистки требуемых областей рабочего листа перед синтезом нового комплекта заданий.
3. Модуль синтеза логических функций служит для формирования комплекта заданий по теме «Построение таблиц истинности логических выражений».
4. Модуль представления отвечает за оформление областей, отводимых для построения таблиц истинности конечным пользователем.
5. Модуль анализа логических функций предназначен для выделения в сформированном выражении основных логических операций.
6. Модуль основных логических операций реализует построение таблиц истинности, соответствующих логических операций.
7. Модуль построения таблиц истинности служит для формирования таблицы истинности заданного логического выражения (далее – эталонная ТИ).
8. Модуль анализа результатов предназначен для сопоставления значений, введенных конечным пользователем, с эталонной ТИ и формирования оценки.

Модули подготовки нового проекта, его представления и защиты

Интерфейс АСАЛ представляет собой рабочий лист книги Microsoft Excel, на котором расположены три элемента управления формой «Начать», «Завершить» и «Очистить».

Для обеспечения защиты структуры рабочей книги во время функционирования отдельных фрагментов программы и ИВЗ от несанкционированного воздействия пользователя предусмотрен модуль защиты. Этот модуль запускается каждый раз при нажатии на любую из кнопок интерфейса и предусматривает снятие и установку пароля для рабочего листа при выполнении запрограммированных действий. Для обеспечения ввода данных на рабочий лист конечным пользователем предусмотрены специальные диапазоны-исключения, для которых защита не устанавливается.

Для подготовки рабочей книги к синтезу нового комплекта заданий реализован модуль подготовки нового проекта, в результате работы которого производится очистка требуемых областей рабочего листа.

Оформление нового комплекта заданий осуществляется при работе модуля представления. При этом на листе Excel размещаются три контейнера, в которые конечный пользователь вводит исходные данные для работы модулей построения таблиц истинности и анализа результатов. Ширина каждого контейнера определяется числом переменных, входящих в соответствующую логическую функцию ИВЗ. Для контроля знания студентами порядка составления таблиц истинности логических выражений, высота всех контейнеров принята одинаковой, заведомо избыточной и составляет десять строк, а пользователь во время работы заполняет в каждом контейнере только те ячейки, которые необходимы для выполнения задания. При работе модуля представления также происходит оформление контейнеров границами и заливка столбца для ввода результирующих значений логической функции (рис. 4).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R
1	КР по теме "Алгебра логики"														Пуск 1			
2															Начало	11:01:12		
3																		
4																		
5	1. Составить таблицы истинности для трех логических функций:																	
6	(в таблицах последовательно без пропусков строк заполняются																	
7	только те ячейки, которые необходимы для выполнения задания)																	
8																		
9	1) $F = A \cdot \neg B + \neg C$						2) $F = \neg A + B$						3) $F = A + B + \neg C$					
10																		
11	A	B	C	F			A	B	F				A	B	C	F		
12																		
13																		
14																		
15																		
16																		
17																		
18																		
19																		
20																		
21																		
22																		
23																		
24																		
25	2. Построить логическую схему для функции $F = A + B + \neg C$.																	
26																		
27	3. Упростить логическое выражение $F_5 = A \cdot (B \cdot (\neg A + \neg B))$.																	

Рис. 4. Интерфейс АСАЛ
Fig. 4. Automated system "Boolean Algebra" interface

Для учета времени выполнения ИВЗ, а также предотвращения недобросовестной работы студентов и несанкционированного формирования ими нового ИВЗ при функционировании

модуля представления АСАЛ на рабочий лист выводятся счетчик числа запусков программы и время формирования текущего ИВЗ.

Модуль синтеза логических функций

Модуль синтеза логических функций реализует непосредственное формирование ИВЗ. Первое задание предусматривает проверку знания студентом порядка составления таблиц истинности функций из n переменных. Для формирования этих функций синтезируется массив msf из трех логических функций f : одна функция из двух переменных ($kf2 = 1$) и две функции из трех переменных ($kf3 = 2$). Для увеличения числа вариантов заданий логические функции f , представляющие собой текстовые последовательности, формируются случайным образом. При заполнении массива msf выполняется проверка логических функций на совпадение с уже имеющимися в массиве.

Укрупненный алгоритм функционирования модуля синтеза логических функций при формировании первого задания представлен на рис. 5.

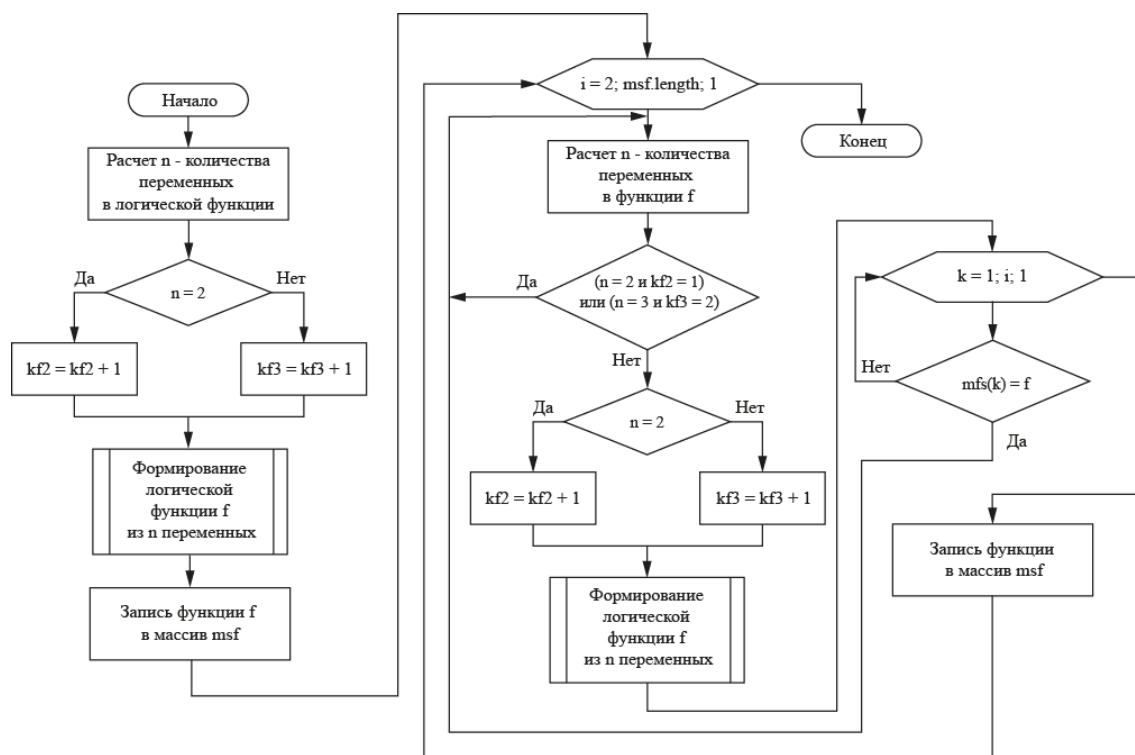


Рис. 5. Алгоритм синтеза логических функций

Fig. 5. Boolean functions synthesis algorithm

Для второго задания логическая функция выбирается случайным образом из сформированных в массиве msf логических функций из трех переменных.

Логическая функция для третьего задания выбирается случайным образом из заранее подготовленного банка функций.

Второе и третье задания ИВЗ на текущий момент выполняются студентом письменно и проверяются преподавателем вручную.

Модули основных логических операций и анализа логических функций

Модуль основных логических операций включает в себя три процедуры-функции, выполняющие операции инверсии (*inversia*), конъюнкции (*kon*) и дизъюнкции (*diz*). Аргументами этих процедур-функций являются массивы значений переменных.

Для декомпозиции логических выражений, полученных при работе модуля синтеза логических функций, и определения последовательности выполнения логических операций предусмотрен модуль анализа логических функций. Декомпозиция логической функции выполняется согласно алгоритму, представленному на рис. 6. Логическое выражение считывается в текстовую переменную из соответствующей ячейки на листе Excel. Для определения числа переменных в логическом выражении определяется позиция первого вхождения в строку символа «C». Если эта позиция отлична от нуля, то в логическом выражении формируются три массива для чтения значений переменных *A*, *B*, *C*, в противном случае формируются два массива для чтения значений переменных *A*, *B*.

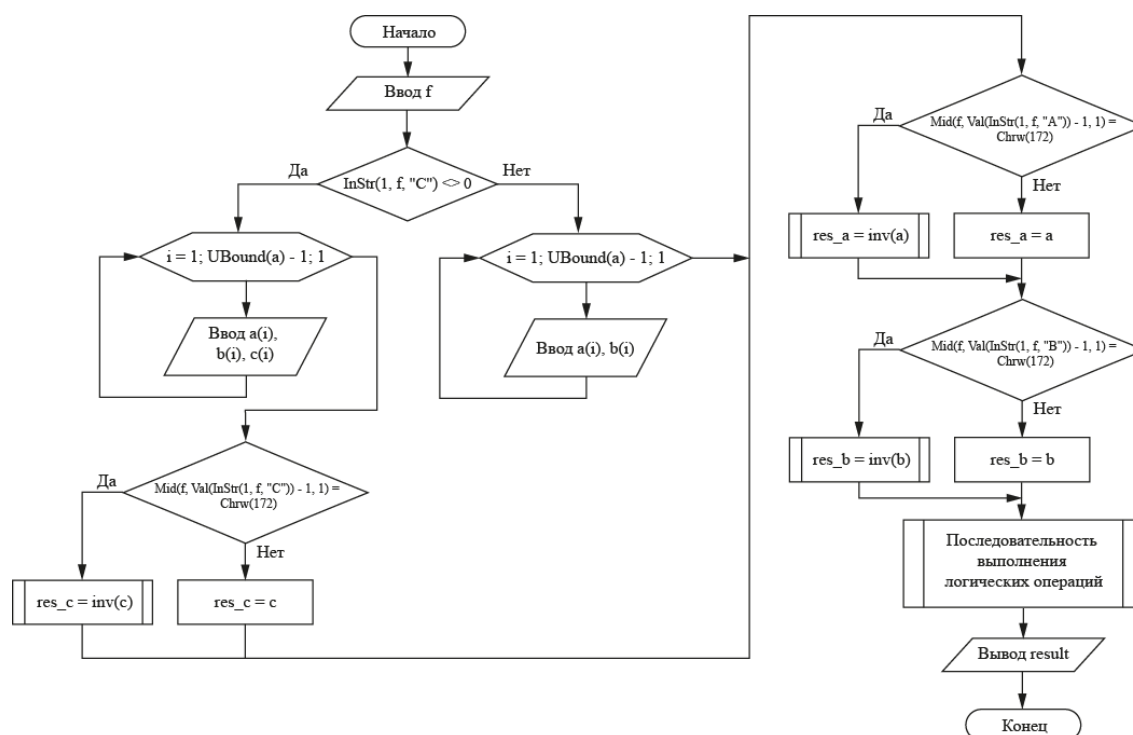


Рис. 6. Алгоритм декомпозиции логических функций
Fig. 6. Boolean functions decomposition algorithm

Согласно существующему приоритету логических операций первой выполняется операция инверсии. Положение этой операции в логическом выражении определяется сопоставлением ASCII кода символа, расположенного в позиции $n - 1$ (где n – позиция первого вхождения в строку символа анализируемой переменной), с ASCII кодом символа отрицания, который был выбран при формировании логической функции. При совпадении этих кодов посредством процедуры-функции *inversia* в массив значений соответствующей переменной записываются ее инверсные значения.

Последовательность выполнения операций конъюнкции и дизъюнкции устанавливается путем сопоставления позиций первого вхождения в строку логической функции символов с ASCII кодом, соответствующим знакам конъюнкции и дизъюнкции, которые были примене-

ны при формировании логического выражения. Укрупненный алгоритм определения последовательности выполнения логических операций приведен на рис. 7.

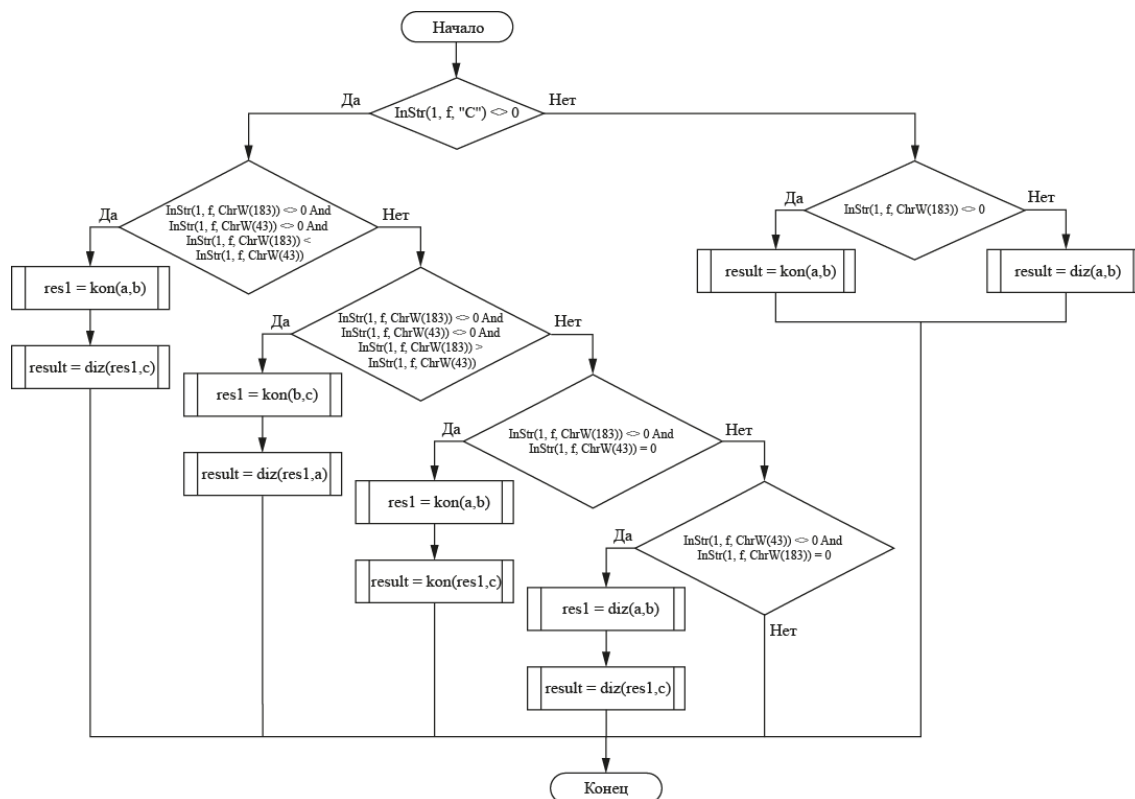


Рис. 7. Алгоритм определения последовательности выполнения логических операций

Fig. 7. Boolean operations execution sequence algorithm

Модули построения таблиц истинности и анализа результатов

Для анализа результатов выполнения задания предусмотрены модули построения таблиц истинности и анализа результатов. Необходимость вызова модуля построения таблиц истинности для заданного логического выражения определяется предварительной проверкой наличия в соответствующих ячейках необходимых для работы программы исходных данных.

При построении ТИ существует рекомендуемый порядок перечисления комбинаций исходных значений переменных. В АСАЛ ввод этих значений на лист Excel и соответствующих им значений функции может осуществляться пользователем в произвольном порядке. Для реализации этой возможности производится запись значений переменных, указанных пользователем на листе Excel, в отдельные массивы. Эти массивы сопоставляются с эталонными значениями, которые должны принимать переменные при построении ТИ. Дополнительно контролируется длина массива исходных значений. Если введенные значения исходных данных соответствуют эталонным значениям, то вызывается процедура построения эталонной ТИ.

Анализ результатов выполнения задания заключается в сопоставлении массивов значений функции F эталонной ТИ и значений, указанных пользователем в столбце F. При несовпадении этих массивов эталонные значения выводятся на лист Excel в столбец, расположенный справа от контейнера с исходными данными для проверки соответствующей логической функции.

Для оценки результатов работы пользователя в АСАЛ для каждой логической функции установлены весовые коэффициенты: для функции из трех переменных – 40 баллов, для функции из двух переменных – 20 баллов. По завершении выполнения задания 1 на лист Excel

помимо общего результата выводятся время окончания работы и ее продолжительность в минутах (см. рис. 8).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S								
1	КР по теме "Алгебра логики"														Пуск 1												
2	Начать				Завершить				Очистить				Начало 11:01:12														
3													Конец 11:04:19														
4													Мин. 3,1														
5	1. Составить таблицы истинности для трех логических функций:														Т. ист.	60											
6	(в таблицах последовательно без пропусков строк заполняются																										
7	только те ячейки, которые необходимы для выполнения задания)																										
8																											
9	1) $F = A \cdot \neg B + \neg C$				2) $F = \neg A + B$				3) $F = A + B + \neg C$																		
10																											
11	A	B	C	F					A	B	F					A	B	C	F								
12	0	0	0	1					0	0	1					0	0	0	1	1							
13	0	0	1	0					0	1	1					0	0	1	0	0							
14	0	1	0	1					1	0	0					0	1	0	1	1							
15	0	1	1	0					1	1	1					0	1	1	1	1							
16	1	0	0	1												1	0	0	1	1							
17	1	0	1	1												1	0	1	1	1							
18	1	1	0	1												1	1	0	1	1							
19	1	1	1	0												1	1	1	0	1							
20																											
21																											
22	Результат			+					+							-											
23	Балл			40					20							0											

Рис. 8. Результаты выполнения первого задания

Fig. 8. First task results

Результаты и выводы

Разработанная автоматизированная система обеспечивает возможность обучения и аттестации студентов, позволяет сформировать ИВЗ и оценить знания пользователей и их умения применять известные алгоритмы для построения ТИ. За пятилетний период функционирования АСАЛ зарекомендовала себя положительно. Основными достоинствами системы являются небольшой объем памяти, занимаемой на диске, возможность работы в компьютерных сетях и на локальных компьютерах, интуитивно понятный интерфейс, большое количество вариантов, формирующихся автоматически, возможность автоматической проверки результатов построения ТИ.

В перспективе планируется развитие системы по следующим направлениям: перенос интерфейса в исполняемый файл, добавление логических операций импликации и эквиваленции при формировании логических функций, автоматизация проверки задания по упрощению выражений, автоматизированное построение логической схемы.

Список литературы

1. **Российская Федерация. Законы.** Об образовании в Российской Федерации : Федеральный закон № 273-ФЗ : [принят Государственной думой 29 декабря 2012 года : одобрен

- Советом Федерации 26 декабря 2012 года] [Электронный ресурс] // Доступ из справ.-правовой системы КонсультантПлюс.
2. **Российская Федерация.** Единый план по достижению национальных целей развития Российской Федерации на период до 2024 года и на плановый период до 2030 года : [утвержден распоряжением Правительства РФ от 01.10.2021 N 2765-р : с изменениями от 24.12.2021] [Электронный ресурс] // Доступ из справ.-правовой системы КонсультантПлюс.
 3. **Евсеева Т. П.** Разработка тестовых заданий как один из методов технологии интерактивного обучения // Вестник Казанского технологического университета. 2014. Т. 17, № 5. С. 320–324.
 4. Свидетельство о государственной регистрации программы для ЭВМ № 2008614627 Российская Федерация. Универсальная тестово-обучающая программа контроля профессиональных знаний Test Master 2007: № 200861340 : заявл. 29.07.2008 : зарегистр. 25.09.2008.
 5. **Киселева Н. А.** К вопросу об использовании тренинговых занятий в образовательном процессе вуза [Электронный ресурс] // Педагогические науки. URL: <https://cyberleninka.ru/article/n/k-voprosu-ob-ispolzovanii-treningovyh-zanyatiy-v-obrazovatelnom-protsesse-vuza/viewer> (дата обращения 20.04.2021).
 6. **González A. G., Salgado D. R., J. García Sanz-Calcedo, C. Cruz García, J. Barrios Muriel, O. Lopez Perez, Alvarez Garsia F.J.** A teaching methodology for the real-time assessment of students' competencies related to manufacturing subjects using technology based on electronic devices. *Procedia Manufacturing*, 2019. Vol. 41. Pp. 579–586.
 7. **Будилова А. С.** Развитие коммуникативных умений студентов при обучении информационным технологиям на основе тренинга [Электронный ресурс] // Наука и перспективы. 2016. № 4. URL: <https://cyberleninka.ru/article/n/razvitie-kommunikativnyh-umeniy-studentov-pri-obuchenii-informatsionnym-tehnologiyam-na-osnove-treninga/viewer> (дата обращения 20.04.2021).
 8. **Посов И. А.** Программирование генераторов задач. [Электронный ресурс] // Компьютерные инструменты в образовании. 2010. № 3. С. 19–31. URL: <http://www.ipos.spb.ru/journal/content/1197/Программирование%20генераторов%20задач.pdf> (дата обращения 20.04.2021).
 9. **Ультан А. Е.** Компьютерная обучающая система «Алгебра логики» [Электронный ресурс] // Инновации в науке. URL: <https://cyberleninka.ru/article/n/kompyuternaya-obuchayushchaya-sistema-algebra-logiki-1> (дата обращения: 17.04.2022).
 10. **Сидорова Е. А., Долгова А. В., Железняк С. П.** Автоматизированная система синтеза структурированного учебного контента [Электронный ресурс] // Вестник НГУ. Серия: Информационные технологии. 2021. Т. 19, № 2. С. 102–114. DOI: 10.25205/1818-7900-2021-19-2-102-114

References

1. **The Russian Federation. Laws.** On education in the Russian Federation: Federal Law No. 273-FZ : [adopted by the State Duma on December 29, 2012 : approved by the Federation Council on December 26, 2012] [Online]. ConsultantPlus.
2. **The Russian Federation.** A unified plan to achieve the national development goals of the Russian Federation for the period up to 2024 and for the planned period up to 2030: [approved Decree of the Government of the Russian Federation of 01.10.2021 N 2765-r : as amended dated December 24, 2021] [Online]. ConsultantPlus.
3. **Evseeva T. P., Sabirova Yu. V.** Razrabotka testovykh zadaniy kak odin iz metodov tekhnologii interaktivnogo obucheniya [Development of test tasks as one of the methods of interactive

- learning technology]. *Bulletin of the Kazan Technological University*, 2014, vol. 17, no. 5, pp. 320–324. (in Russ.)
4. Certificate of state registration of the computer program No. 2008614627 Russian Federation. Universal test and training program for the control of professional knowledge Test Master 2007: No. 200861340: Appl. 07/29/2008: registered 09/25/2008. P. 1.
 5. **Kiseleva N. A.** K voprosu ob ispol'zovanii treningovykh zanyatii v obrazovatel'nom protsesse vuza [To the question of the use of training sessions in the educational process of the university] / N. A. Kiseleva. [Online]. *Pedagogicheskie nauki*. URL: <https://cyberleninka.ru/article/n/k-voprosu-ob-ispolzovanii-treningovykh-zanyatii-v-obrazovatelnom-protsesse-vuza/viewer> (accessed on 04/24/2022). (in Russ.)
 6. **González A. G., Salgado D. R., J. García Sanz-Calcedo, C. Cruz García, J. Barrios Muriel, O. Lopez Perez, Alvarez Garsia F. J.** A teaching methodology for the real-time assessment of students' competencies related to manufacturing subjects using technology based on electronic devices. *Procedia Manufacturing*, 2019, vol. 41, pp. 579–586.
 7. **Budilova A. S.** Development of communicative skills of students at training information technologies on the basis of training [Online]. *Science and Perspectives*, 2016, no. 4. URL: <https://cyberleninka.ru/article/n/razvitie-kommunikativnyh-umeniy-studentov-pri-obuchenii-informatsionnym-tehnologiyam-na-osnove-treninga/viewer> (accessed on 04/24/2022).
 8. **Posov I. A.** Programming of problem generators [Online]. *Computer tools in education*, 2010, no. 3, pp. 19–31. URL: [http://www.ipos.spb.ru/journal/content/1197/Программирование %20 генераторов%20задач.pdf](http://www.ipos.spb.ru/journal/content/1197/Программирование%20генераторов%20задач.pdf) (accessed on 04/24/2022).
 9. **Ul'tan A. E., Kravtsov D. A.** Komp'yuternaya obuchayushchaya sistema «Algebra logiki» [Computing education system «Boolean algebra»] [Online]. *Innovatsii v nauke*. URL: <https://cyberleninka.ru/article/n/kompyuternaya-obuchayushchaya-sistema-algebra-logiki-1> (date of treatment 04/24/2022). (in Russ.)
 10. **Sidorova E. A.** Automated system synthesis of structured educational content / E. A. Sidorova, A. V. Dolgova, S. P. Zheleznyak [Online]. *Vestnik NSU. Series: Information Technologies*, 2021. Vol. 19, no. 2, pp.102–114. DOI: 10.25205/1818-7900-2021-19-2-102-114 (in Russ.)

Информация об авторах

Сидорова Елена Анатольевна, доктор технических наук, доцент, заведующий кафедрой «Информатика и компьютерная графика», Омский государственный университет путей сообщения (Омск, Россия)
Researcher ID: B-5089-2017

Долгова Анна Владимировна, кандидат технических наук, доцент, доцент кафедры «Информатика и компьютерная графика», Омский государственный университет путей сообщения (Омск, Россия)
Researcher ID: D-9708-2019

Железняк Светлана Петровна, кандидат технических наук, доцент кафедры «Информатика и компьютерная графика», Омский государственный университет путей сообщения (Омск, Россия)
Researcher ID: AAM-5444-2021

Information about the Authors

Elena A. Sidorova, Doctor of Technical Sciences, Associate Professor, Head of the Department of Computer Science and Computer Graphics, Omsk State University of Railway Engineering (Omsk, Russian Federation)
Researcher ID: B-5089-2017

Anna V. Dolgova, Candidate of Technical Sciences, Associate Professor, Associate Professor of the Department of Computer Science and Computer Graphics, Omsk State University of Railway Engineering (Omsk, Russian Federation)
Researcher ID – D-9708-2019

Svetlana P. Zheleznyak, Candidate of Technical Sciences, Associate Professor of the Department of Computer Science and Computer Graphics, Omsk State University of Railways (Omsk, Russian Federation)
ResearcherID: AAM-5444-2021

*Статья поступила в редакцию 27.04.2022;
одобрена после рецензирования 17.08.2022; принята к публикации 17.08.2022
The article was submitted 27.04.2022;
approved after reviewing 17.08.2022; accepted for publication 17.08.2022*

Научная статья

УДК 004.912+004.8

DOI 10.25205/1818-7900-2022-20-3-65-76

Извлечение семантических отношений из текстов научных статей

**Ольга Юрьевна Тихобаева¹, Елена Павловна Бручес²,
Татьяна Викторовна Батура³**

^{2,3}Институт систем информатики им. А. П. Ершова СО РАН
Новосибирск, Россия

¹⁻³Новосибирский государственный университет
Новосибирск, Россия

¹otikhobaeva10@gmail.com

²bruches@bk.ru

³tatiana.v.batura@gmail.com, <https://orcid.org/0000-0003-4333-7888>

Аннотация

В современном мире количество научных публикаций, существующих в виде электронного текста, постоянно растет. В связи с этим задачи, связанные с обработкой текстов научных статей, становятся особо актуальными. Данная работа посвящена задаче извлечения семантических отношений между сущностями из текстов научных статей на русском языке, где в качестве сущностей выступают научные термины. Извлечение отношений может быть полезно в отдельных специализированных областях, таких как поисковые и вопросно-ответные системы, а также при составлении онтологий. В ходе работы нами был создан корпус научных текстов, состоящий из 136 аннотаций научных статей на русском языке, в которых выделены 353 отношения следующих типов: USAGE, ISA, TOOL, SYNONYMS, PART_OF, CAUSE. Данный корпус использовался нами для обучения моделей. Кроме того, мы реализовали алгоритм автоматического извлечения семантических отношений и протестировали его на уже существующем корпусе научных текстов RuSERRC. Для реализации алгоритма использовалась нейросетевая модель BERT. Мы провели ряд экспериментов, связанных с использованием векторов, полученных из различных языковых моделей, а также с двумя нейросетевыми архитектурами. Разработанный инструмент и размеченный корпус выложены в открытый доступ и могут быть полезны для других исследователей.

Ключевые слова

извлечение отношений, научные термины, разметка данных, языковые модели, обработка текстов

Для цитирования

Тихобаева О. Ю., Бручес Е. П., Батура Т. В. Извлечение семантических отношений из текстов научных статей // Вестник НГУ. Серия: Информационные технологии. Т. 20, № 3. С. 65–76. DOI 10.25205/1818-7900-2022-20-3-65-76

Extracting Semantic Relations from the Texts of Scientific Articles

Olga Y. Tikhobaeva¹, Elena P. Bruches²,
Tatyana V. Batura³

^{2,3} A.P. Ershov Institute of Informatics Systems SB RAS
Novosibirsk, Russian Federation

¹⁻³ Novosibirsk State University
Novosibirsk, Russian Federation

¹otikhobaeva10@gmail.com

²bruches@bk.ru

³tatiana.v.batura@gmail.com, <https://orcid.org/0000-0003-4333-7888>

Abstract

Nowadays, the number of scientific publications existing in the form of electronic text is constantly growing. As a result, the tasks related to the text processing of scientific articles become especially actual. This paper is dedicated to the task of extracting semantic relations between entities from the texts of scientific articles in Russian, where we consider scientific terms as entities. Relation extraction can be useful in some specialized areas, such as searching and question-answering systems, as well as in the compilation of ontologies. In our work, we have created a corpus of scientific texts consisting of 136 abstracts of scientific articles in Russian, in which 353 relations of the following types were highlighted: USAGE, ISA, TOOL, SYNONYMS, PART_OF, CAUSE. This corpus was used to train the machine learning models. In addition, we have implemented the automatic semantic relation extraction algorithm and tested it on the already existing corpus RuSERRC. The neural network model BERT was used to implement the algorithm. We've done a number of experiments using vectors derived from different language models, as well as two neural network architectures. The developed tool and the annotated corpus are publicly available and can be useful for other researchers.

Keywords

relation extraction, scientific terms, data annotation, language models, natural language processing

For citation

Tikhobaeva O. Yu., Bruches E. P., Batura T. V. Extracting Semantic Relations from the Texts of Scientific Articles. *Vestnik NSU. Series: Information Technologies*, 2022, vol. 20, no. 3, pp. 65–76. DOI 10.25205/1818-7900-2022-20-3-65-76

Введение

В современном мире количество научных публикаций, существующих в виде электронного текста, в частности в сети Интернет, постоянно растет. Каким образом человек может обрабатывать все эти данные? Популярная идея заключается в преобразовании неструктурированного текста в структурированный посредством разметки семантической информации. Однако большой объем и неоднородность данных делают аннотацию вручную почти невозможной. Было бы гораздо эффективнее, если бы компьютер аннотировал все данные со структурой, необходимой для решения тех или иных задач.

Очень часто в научных текстах нас интересуют именно семантические отношения между терминами. Их извлечение может быть полезно в отдельных специализированных областях, таких как поисковые и вопросно-ответные системы, а также при составлении онтологий. Именно возможности практического применения полученных структурированных данных послужили толчком к усиленному исследованию данной проблемы.

Задача извлечения семантических отношений (Relation Extraction, RE) заключается в том, что после распознавания отдельных сущностей (например, «мультимедийные технологии» и «учебный процесс») необходимо классифицировать отношения, существующие между ними (например, «USAGE (мультимедийные технологии, учебный процесс)») по контексту, в котором они находятся (например, «Показаны преимущества использования мультимедийных технологий в учебном процессе»).

На данный момент, задача остается сложной для любой предметной области, так как часто требует использования знаний вне текста (например, из баз знаний или полученных другим

путем), а также из-за отсутствия большого количества размеченных данных на русском языке для решения задачи RE.

В данной работе мы поставили перед собой две цели, во-первых, создание корпуса научных текстов с разметкой семантических отношений для русского языка, во-вторых, реализация алгоритма для извлечения семантических отношений.

Мы провели несколько экспериментов, попробовав различные подходы к проблеме извлечения семантических отношений, а затем сравнили наши результаты с результатами уже существующего алгоритма.

Обзор методов

Существуют различные подходы к решению основной задачи извлечения семантических отношений, то есть к классификации отношений между распознанными сущностями.

Самый ранний и простой из них – это подход, основанный на лексико-синтаксических шаблонах. Данный подход, подробно описанный в работе [1], заключается в выявление реальных шаблонов, которые прямо выражают те или иные семантические отношения в тексте и поиске экземпляров таких отношений при помощи этих шаблонов. Сами шаблоны могут представлять собой как просто строки, так и комбинации из частеречных тегов и лексических единиц, что характерно при работе с корпусами. Однако данный подход имеет свои недостатки, так как дает высокую точность, но низкую полноту, а также требует большого количества человеческих усилий для составления шаблонов или проверки автоматически выявленных.

Следующий подход для решения задачи извлечения семантических отношений – это метод, основанный на признаках. Данный подход можно отнести к статистическим методам извлечения отношений. Он подробно описан в работе [2]. Суть данного метода заключается в извлечении семантических и синтаксических признаков из текстов обучающих примеров, которые затем представляются классификатору в виде вектора признаков для обучения или классификации. К таким признакам авторы относят сущности, между которыми нужно установить отношение, слова между этими сущностями и их количество, типы этих сущностей, пути в дереве синтаксических связей (the syntactic parse tree) и в дереве зависимостей (the dependency tree). Основным недостатком данного подхода заключается в том, что он требует сложной синтаксической и семантической обработки текста: построения деревьев, определения типов сущностей и т. д.

Последний подход, который мы рассмотрим, это извлечение семантических отношений при помощи нейронных сетей. Исследования в этой области в основном сосредоточены на проектировании и использовании различных сетевых архитектур для захвата семантики отношений в тексте. В качестве входных данных в основном используются векторные представления слов и позиций их в тексте вместо вручную отобранных признаков. Так в работе [3] описана архитектура нейронной сети, которую авторы используют для извлечения отношений. Она включает в себя следующие три компонента: векторное представление слов, извлечение признаков и вывод. Система не нуждается в сложной синтаксической или семантической предварительной обработке данных, и на вход подаются предложения с двумя отмеченными существительными. Слова преобразуются в векторы, затем последовательно извлекаются лексические и синтаксические признаки, которые потом конкатенируются для формирования конечного вектора признаков. Наконец, для вычисления достоверности каждого отношения вектор признаков подается в softmax классификатор. Выход классификатора является вектором, размерность которого равна числу предопределенных типов отношений. Значение каждого измерения – это балл достоверности соответствующего отношения.

Описание данных

На данный момент существует корпус научных статей на русском языке RuSERRC¹ [4], который содержит 1680 текстов, 80 из которых размечены вручную, и еще 1600 не имеют разметки. В размеченных текстах выделено 4 типа информации: научные термины, семантические отношения между ними, вложенные сущности, а также связи выделенных терминов с сущностями из Wikidata. Для решения задачи RE в данном корпусе представлены 589 отношений следующих типов:

- USAGE – ‘х используется для/в у’: 330;
- ISA – ‘х является у’: 93;
- TOOL – ‘х позволяет создавать/изучать/и т. п. у’: 38;
- SYNONYMS – ‘х – это то же, что у’: 22;
- PART_OF – ‘х является частью у’: 87;
- CAUSE – ‘х является причиной у’: 19.

В предыдущей версии инструмента [5] применялся zero-shot подход, то есть решение задачи без предоставления материалов для обучения этому решению, а именно без обучения на русскоязычных данных. В этой работе мы попробовали улучшить качество модели за счет добавления вручную размеченных данных в обучающее множество.

Для получения таких данных мы дополнили существующий корпус RuSERRC еще 136 текстами, выделив в них 353 отношений следующих типов:

- USAGE – ‘х используется для/в у’: 126;
- ISA – ‘х является у’: 96;
- TOOL – ‘х позволяет создавать/изучать/и т. п.’: 54;
- SYNONYMS – ‘х – это то же, что у’: 35;
- PART_OF – ‘х является частью у’: 23;
- CAUSE – ‘х является причиной у’: 19.

Выделение отношений проходило следующим образом: проводился поиск выбранных для рассмотрения отношений между терминами в текстах научных статей, где ранее была проведена BIO разметка терминов. Мы рассматривали отношения между сущностями только в пределах одного предложения, и одна сущность могла участвовать в нескольких отношениях. Эти отношения были занесены в специальный файл в следующем формате: PART OF(16:18). Слово перед скобками в данной записи обозначает название типа отношений, а числа в скобках – начало первого и второго термина, между которыми установлено отношение.

Эксперименты

Извлечение отношений при помощи лексических шаблонов

Вначале мы применили подход, основанный на лексических шаблонах, указывающих на наличие того или иного отношения. Он заключается в следующем: из текстов, в которых проведена разметка терминов, извлекается контекст, находящийся между двумя терминами, проводится его лемматизация, то есть приведение к начальной форме, а затем происходит его сравнение с имеющимися лексическими шаблонами. Если находится совпадение, то между терминами устанавливается соответствующее отношение. Длина контекста не должна превышать шесть слов. Такое значение было получено экспериментально путем его изменения и сравнения качества модели. Также стоит отметить, что все лексические шаблоны были составлены вручную и включают в себя не только лексические единицы, но и знаки пунктуации. Всего для работы алгоритма используется 111 шаблонов, распределенных по типам отношений следующим образом:

- USAGE: 36;

¹ <https://github.com/iis-research-team/rusercc-dataset>

- ISA: 13;
- TOOL: 29;
- SYNONYMS: 5;
- PART_OF: 5;
- CAUSE: 23.

Примеры лексических шаблонов представлены в табл. 1.

Таблица 1

Примеры лексических шаблонов

Table 1

Examples of Lexical Patterns

Отношение	Примеры маркеров
CAUSE	вызывает; дает в результате
ISA	является; представляет собой
PART_OF	является частью; состоит из
SYNONYMS	иначе; или
TOOL	изучает; создает
USAGE	используется для; с помощью

Однако данный подход имеет свои недостатки и сложности в работе, например, отношения могут быть не выражены в тексте лексически. Также если есть термины A-B-C, то при рассмотрении пары A-C будет выбран общий контекст, включающий и A-B, и B-C, следовательно, не ясно, какую именно пару нужно связать отношением.

Извлечение отношений, как решение стандартной задачи классификации

Следующим шагом в экспериментах стала работа с различными нейросетевыми архитектурами и языковыми моделями.

В первом варианте архитектуры мы брали предложение, спецсимволами выделяли два токена, для которых нужно определить наличие и тип отношения, затем решали задачу классификации, в которой классами выступали типы отношений и один класс для отсутствия отношения между терминами. При работе с данной архитектурой мы использовали языковую модель bert-base-multilingual-cased [6]. Это модель трансформера, предобученная на данных 104 языков с самыми большими объемами Википедии. Схема данной архитектуры представлена на рис. 1.

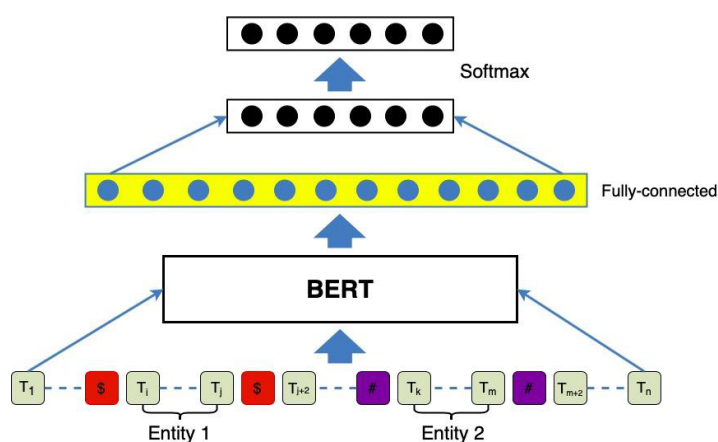


Рис. 1. Схема первой архитектуры
Fig. 1. Scheme of the first architecture

В ходе экспериментов с данной архитектурой мы пробовали обучать модель двумя разными способами. Вначале мы использовали текущую модель для извлечения отношений, обученную на англоязычных данных из корпуса SciERC [7], и дообучали ее на нашем новом русскоязычном корпусе. Затем мы попробовали обучить модель только на русскоязычных данных.

Извлечение отношений, как классификация с использованием CLS вектора

Теперь перейдем ко второму варианту нейросетевой архитектуры. Она устроена следующим образом: берется вектор специального токена CLS (считается, что он представляет собой вектор всего текста, который пришёл на вход) и вектора двух наших терминов. Затем эти три вектора конкатенируются, и полученный вектор подается в классификатор [8]. При работе с данной архитектурой мы попробовали использовать три различные языковые модели: bert-base-multilingual-cased, которая уже упоминалась ранее, rubert-base-cased от Deep Pavlov [9] и cointegrated/rubert-tiny2². Модель rubert-base-cased была обучена на русскоязычной части Википедии и новостных данных. Для инициализации этой модели использовалась многоязычная версия BERT. Модель cointegrated/rubert-tiny2 – это маленькая дистиллированная версия bert-base-multilingual-cased для русского и английского языка. Она была обучена на данных Yandex Translate corpus³, OPUS-100 [10] and Tatoeba⁴. Схема данной архитектуры представлена на рис. 2.

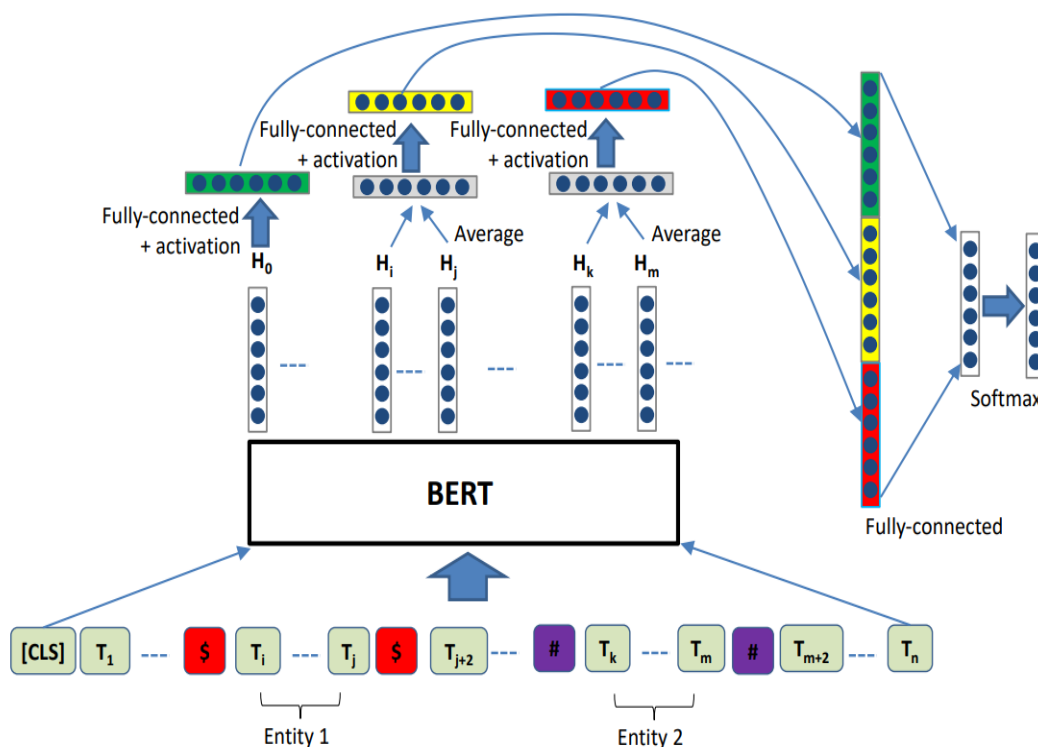


Рис. 2. Схема второй архитектуры
Fig. 2. Schema of second architecture

² <https://huggingface.co/cointegrated/rubert-tiny2>.

³ <https://translate.yandex.ru/corpus>.

⁴ <https://tatoeba.org/ru/>.

Кроме того, стоит упомянуть некоторые особенности обучения, которые заключаются в следующем: во-первых, для обучения модели мы использовали корпус русскоязычных текстов полностью, не разделяя его на обучающую и валидационную выборки, а наиболее подходящее по качеству количество эпох подбирали экспериментально. Это было сделано по причине того, что примеров некоторых отношений очень мало, и потому валидационная выборка была бы нерепрезентативной для определения качества модели; во-вторых, чтобы уменьшить дисбаланс между количеством примеров в классах, в обучающую выборку мы добавили только 50 % случайно выбранных пар терминов, исключая те, в которых расстояние между такими терминами более 10 токенов.

Результаты

Мы тестировали все варианты алгоритма на первоначальном корпусе RuSERRC, который не использовался при обучении моделей. Для оценки качества работы алгоритмов мы использовали стандартные метрики: точность, полнота, F1-мера.

По результатам работы алгоритма, основанного на лексических шаблонах, мы получили метрики, представленные в табл. 2.

Таблица 2

Метрики подхода на основе лексических шаблонов

Table 2

Metrics of Lexical Pattern's Approach

Отношение	Точность	Полнота	F1 – мера
CAUSE	0,07	0,05	0,06
ISA	0,18	0,19	0,19
PART_OF	0,17	0,14	0,15
SYNONYMS	0,23	0,82	0,35
TOOL	0,06	0,08	0,07
USAGE	0,21	0,39	0,27
NO-RELATION	0,96	0,92	0,94
macro-average	0,27	0,37	0,29

Эксперимент с первой архитектурой, в ходе которого мы дообучали текущую модель на новом русскоязычном корпусе, показал результаты, представленные в табл. 3. Поскольку корпуса SciERC и RuSERRC имеют разные наборы отношений, мы получали метрики только для интересующих нас отношений: HYPONYM-OF, PART-OF и USED-FOR. Данные метрики иллюстрируют комбинированный подход, при котором используются как лексические шаблоны, так и языковая модель.

Таблица 3

Метрики комбинированного подхода для дообученной модели
с англоязычными изначальными данными

Table 3

Metrics of the Combined Approach for the Pre-Trained Model
with English-Language Initial Data

Отношение	Точность	Полнота	F1 – мера
HYPONYM-OF	0,18	0,17	0,17
PART-OF	0,17	0,14	0,15
USED-FOR	0,16	0,39	0,23
NO-RELATION	0,96	0,9	0,93
macro-average	0,37	0,4	0,37

При попытке обучить ту же модель только на русскоязычных текстах, мы получили метрики, представленные в табл. 4. Эти метрики так же соответствуют комбинированному подходу.

Таблица 4

Метрики комбинированного подхода для модели,
обученной только на русскоязычных данных

Table 4

Metrics of the Combined Approach for the Model Trained Only in Russian-Language Data

Отношение	Точность	Полнота	F1 – мера
CAUSE	0,07	0,05	0,06
ISA	0,18	0,19	0,19
PART_OF	0,17	0,14	0,15
SYNONYMS	0,23	0,82	0,35
TOOL	0,06	0,08	0,07
USAGE	0,21	0,39	0,27
NO-RELATION	0,96	0,92	0,94
macro-average	0,27	0,37	0,29

При работе со второй архитектурой мы вначале получили метрики для каждой из языковых моделей. Лучшие метрики с макро-усреднением для каждой из моделей представлены в табл. 5.

Таблица 5

Метрики моделей bert-base-multilingual-cased, rubert-base-cased, rubert-tiny2

Table 5

Metrics for Bert-Base-Multilingual-Cased, Rubert-Base-Cased, Rubert-Tiny2 Models

Модель	Точность	Полнота	F1 – мера
bert-base-multilingual-cased	0,26	0,32	0,26
rubert-base-cased	0,26	0,34	0,27
rubert-tiny2	0,22	0,23	0,22

Затем мы получили метрики комбинированного подхода с использованием лексических шаблонов для каждой модели. Лучшие метрики с macro-усреднением для каждой из моделей представлены в табл. 6.

Таблица 6

Метрики комбинированного подхода для моделей bert-base-multilingual-cased, rubert-base-cased, rubert-tiny2

Table 6

Metrics of the Combined Approach for Bert-Base-Multilingual-Cased, Rubert-Base-Cased, Rubert-Tiny2 Models

Модель	Точность	Полнота	F1 – мера
bert-base-multilingual-cased	0,26	0,41	0,29
rubert-base-cased	0,29	0,35	0,28
rubert-tiny2	0,29	0,24	0,24

Лучшие метрики показал комбинированный вариант алгоритма, использующий вторую архитектуру и языковую модель bert-base-multilingual-cased. Мы сравнили результаты работы данного алгоритма с текущей версией инструмента, которая работает только с частью описанных в данной работе типов отношений. Метрики нашего алгоритма для сокращенного набора отношений, а также сравнение с текущей версией представлены в табл. 7.

Таблица 7

Метрики комбинированного подхода для модели bert-base-multilingual-cased для изначального набора отношений

Table 7

Metrics of the Combined Approach for Bert-Base-Multilingual-Cased Model with Initial Set of Relations

Отношение	Точность	Полнота	F1 – мера
HYPONYM-OF	0,27	0,37	0,31
PART-OF	0,15	0,13	0,14
USED-FOR	0,19	0,48	0,27
NO-RELATION	0,96	0,9	0,93
macro-average	0,39	0,47	0,41
macro-average (текущая версия инструмента)	0,38	0,36	0,34

Заключение

В ходе работы был создан корпус с разметкой семантических отношений для русскоязычных текстов научных статей, а также реализован алгоритм для автоматического извлечения семантических отношений. Код инструмента выложен в открытом доступе и может быть использован другими исследователями.⁵

Наши исследования показали, что для использования лексических маркеров при решении данной задачи недостаточно эксплицитного контекста, и отношения зачастую представлены в текстах неявно.

В экспериментах с использованием различных Transformer-архитектур было выявлено, что модель, использующая информацию о всём предложении (CLS-токен), существенно выигрывает в качестве.

Также мы сделали вывод, что добавление вручную размеченных данных в обучающий набор повышает общее качество системы извлечения отношений.

В дальнейшем мы планируем попробовать улучшить работу модели путем изменения параметров обучения, а также попробовать использовать другие языковые модели для сравнения результатов и нахождения лучшего варианта. Кроме того, мы собираемся расширить корпус вручную размеченных текстов, особенно уделив внимание малочисленным отношениям. Дополнительно мы хотим попробовать использовать модуль связывания сущностей для повышения качества извлечения отношений, так как в базе знаний представлены как сущности, так и их связи друг с другом, которые могут послужить основой для извлечения отношений.

Список литературы

1. **Auger A., Barrière C.** Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 2008. Vol. 14, no. 1. Pp. 1–19. DOI: 10.1075/term.14.1.02 aug
2. **Kambhatla N.** Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2004. pp. 178–181. DOI: 10.3115/1219044.1219066
3. **Zeng D., Liu K., Lai S., Zhou G., Zhao J.** Relation classification via convolutional deep neural network. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2014. Pp. 2335–2344.
4. **Bruches E., Pauls A., Batura T., Isachenko V.** Entity recognition and relation extraction from scientific and technical texts in Russian. *2020 Science and Artificial Intelligence conference (S.A.I.ence)*, IEEE, 2020. Pp. 41–45. DOI: 10.1109/s.a.i.ence50533.2020.9303196
5. **Bruches E., Mezentseva A., Batura T.** A system for information extraction from scientific texts in Russian, 2021. arXiv preprint arXiv:2109.06703.
6. **Devlin J., Chang M.W., Lee K., Toutanova K.** Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019. Vol. 1 (Long and Short Papers). Pp. 4171–4186. arXiv preprint arXiv:1810.04805. DOI: 10.18653/v1/N19-1423.
7. **Luan Y., He L., Ostendorf M., Hajishirzi H.** Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018. Pp. 3219–3232. DOI: 10.18653/v1/D18-1360

⁵ <https://github.com/iis-research-team/terminator#relation-extraction>

8. **Wu S., He Y.** Enriching pre-trained language model with entity information for relation classification. *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019. p. 2361-2364. DOI: 10.1145/3357384.3358119
9. **Kuratov Y., Arkhipov M.** Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, Moscow, May 29—June 1, 2019. arXiv preprint arXiv:1905.07213
10. **Zhang B., Williams P., Titov I., Sennrich R.** Improving massively multilingual neural machine translation and zero-shot translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Pp. 1628–1639, Online. arXiv preprint arXiv:2004.11867. DOI: 10.18653/v1/2020.acl-main.148

References

1. **Auger A., Barrière C.** Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*, 2008. vol. 14, no. 1, pp. 1–19. DOI: 10.1075/term.14.1.02aug
2. **Kambhatla N.** Combining lexical, syntactic, and semantic features with maximum entropy models for extracting relations. *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, 2004. Pp. 178–181. DOI: 10.3115/1219044.1219066
3. **Zeng D., Liu K., Lai S., Zhou G., Zhao J.** Relation classification via convolutional deep neural network. *Proceedings of COLING 2014, the 25th international conference on computational linguistics: technical papers*, 2014. p. 2335-2344.
4. **Bruches E., Pauls A., Batura T., Isachenko V.** Entity recognition and relation extraction from scientific and technical texts in Russian. *2020 Science and Artificial Intelligence conference (S.A.I.ence)*, IEEE, 2020. Pp. 41–45. DOI: 10.1109/s.a.i.ence50533.2020.9303196
5. **Bruches E., Mezentseva A., Batura T.** A system for information extraction from scientific texts in Russian, 2021. arXiv preprint arXiv:2109.06703
6. **Devlin J., Chang M.W., Lee K., Toutanova K.** Bert: Pre-training of deep bidirectional transformers for language understanding. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, Minnesota, 2019. Vol. 1 (Long and Short Papers), pp. 4171–4186. arXiv preprint arXiv:1810.04805. DOI: 10.18653/v1/N19-1423
7. **Luan Y., He L., Ostendorf M., Hajishirzi H.** Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018. Pp. 3219–3232. DOI: 10.18653/v1/D18-1360.
8. **Wu S., He Y.** Enriching pre-trained language model with entity information for relation classification. *Proceedings of the 28th ACM international conference on information and knowledge management*, 2019. Pp. 2361–2364. DOI: 10.1145/3357384.3358119
9. **Kuratov Y., Arkhipov M.** Adaptation of deep bidirectional multilingual transformers for Russian language. *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference "Dialogue 2019"*, Moscow, May 29—June 1, 2019. arXiv preprint arXiv:1905.07213
10. **Zhang B., Williams P., Titov I., Sennrich R.** Improving massively multilingual neural machine translation and zero-shot translation. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020. Pp. 1628–1639, Online. arXiv preprint arXiv:2004.11867. DOI: 10.18653/v1/2020.acl-main.148

Информация об авторах

Тихобаева Ольга Юрьевна, студентка, Новосибирский государственный университет (Новосибирск, Россия)

Бручес Елена Павловна, младший научный сотрудник, Институт систем информатики им. А. П. Ершова СО РАН (Новосибирск, Россия); старший преподаватель, Новосибирский государственный университет (Новосибирск, Россия)

Батура Татьяна Викторовна, кандидат физико-математических наук, доцент, заведующий лабораторией, Институт систем информатики им. А. П. Ершова СО РАН (Новосибирск, Россия); доцент, Новосибирский государственный университет (Новосибирск, Россия)
ORCID: 0000-0003-4333-7888

Information about the Authors

Olga Yur. Tikhobaeva, Student, Novosibirsk State University (Novosibirsk, Russian Federation)

Elena P. Bruches, Junior Researcher, A.P. Ershov Institute of Informatics Systems SB RAS (Novosibirsk, Russian Federation); Senior Lecturer, Novosibirsk State University (Novosibirsk, Russian Federation)

Tatiana Viktorovna Batura, PhD in Physics and Mathematics, Associate Professor, Head of Laboratory, A. P. Ershov Institute of Informatics Systems SB RAS (Novosibirsk, Russian Federation); Associate Professor, Novosibirsk State University (Novosibirsk, Russian Federation)

*Статья поступила в редакцию 19.05.2022;
одобрена после рецензирования 05.09.2022; принята к публикации 05.09.2022
The article was submitted 19.05.2022;
approved after reviewing 05.09.2022; accepted for publication 05.09.2022*

Правила оформления текста рукописи

Авторы представляют статьи на русском или английском языке объемом от 0,5 авторского листа (20 тыс. знаков) до 1 авторского листа (40 тыс. знаков), включая иллюстрации (1 иллюстрация форматом 190 × 270 мм = 1/6 авторского листа, или 6,7 тыс. знаков). Публикации, превышающие указанный объем, допускаются к рассмотрению только после индивидуального согласования с редакцией журнала.

Текст рукописи должен быть представлен в редколлегию в виде файла MS Word (.doc, .docx). Гарнитура Times New Roman, размер шрифта 11, межстрочный интервал 1, размеры полей – стандартные значения текстового редактора. Форматирование – выравнивание по ширине страницы, переносы слов включены, каждый новый абзац начинается с красной строки. Не допускается ручное форматирование абзацев (пробелами, лишними переводами строк, разрывами страниц).

Структура статьи

- Индекс УДК (универсальной десятичной классификации). Выравнивание по левому краю
- Название статьи. Выравнивание по центру, полужирный шрифт
- ФИО авторов (полностью). Выравнивание по центру, полужирный шрифт
- Места работы всех авторов. Выравнивание по центру, курсив
- Адреса электронной почты, ORCID авторов
- Аннотация статьи
- Ключевые слова, не более 10
- Благодарности, сведения о финансовой поддержке
- Название статьи **на английском языке**. Выравнивание по центру, полужирный шрифт
- ФИО авторов **на английском языке** (полностью). Выравнивание по центру, полужирный шрифт
- Места работы авторов **на английском языке**. Выравнивание по центру, курсив
- Аннотация статьи **на английском языке (Abstract)**, 200–250 слов
- Ключевые слова **на английском языке (Keywords)**, не более 10
- Благодарности, сведения о финансовой поддержке **на английском языке**, если есть соответствующий раздел на русском языке (**Acknowledgements**)
- Основной текст
- Список литературы / **References**
- Сведения об авторах

Требования к оформлению основного текста и иллюстративных материалов

Основной текст должен быть представлен в структурированном виде, рекомендуется использовать подзаголовки – например: Введение, Методика..., Выводы, Результаты, Заключение.

Подзаголовки отделяются и набираются полужирным шрифтом. В целях выделения частей текста и отдельных слов и словосочетаний допускается использование курсива или полужирного шрифта. Подчеркивание, разрядка, изменение основного кегля и выделение цветом не используются.

Иллюстрации к рукописи статьи должны быть приложены в виде отдельных файлов. При этом в тексте должно содержаться включенное изображение с указанием имени файла. Все иллюстрации, содержащие схемы, графики, алгоритмы и т. п., должны быть представлены в векторном виде (.ai, .eps, .cdr). Скриншоты и другие растровые изображения должны быть представлены в максимально высоком качестве, без каких-либо потерь и искажений (.jpg, .tif). Все иллюстрации должны иметь подрисуночную подпись – свое название. Надписи к таблицам и подписи к иллюстрациям приводятся **на двух языках (русском и английском)**.

Примеры:

Рис. 1. Диаграмма производительности...

Fig. 1. Performance diagram...

Таблица 1

Сравнение алгоритмов...

Table 1

Comparison of algorithms...

Нумерация последовательная и неразрывная от начала статьи. Не допускается использование других наименований, кроме «Рис.» / «Fig.», «Таблица» / «Table», и усложнение нумерации (например, «Рис. 3.2.»). Ссылка на иллюстрацию в тексте должна быть приведена в круглых скобках, например: (рис. 1), (табл. 1).

Формулы должны быть набраны с использованием редактора MathType либо встроенного редактора формул MS Word. Кегль основных символов – 11, греческие символы набираются прямым шрифтом, латинские – курсивом. Нумеруются только те формулы, на которые автор ссылается в тексте.

Abstract

Аннотация статьи на английском языке (Abstract) не должна быть дословным переводом русскоязычной аннотации. Раздел Abstract, как и основной текст, должен быть структурирован, в нем должно содержаться описание цели работы, методов исследования, научной значимости, выводов / результатов. Требуется качественный перевод на английский язык (при необходимости просим авторов обращаться к профессиональным переводчикам). **Объем Abstract 200–250 слов.**

Список литературы / References

Список литературы и список литературы на английском языке (References) размещаются в общем разделе. Рекомендуемое количество цитируемых в статье источников – не менее 10, в список желательно включать ссылки на актуальные работы по теме исследования, особенно в иностранных периодических изданиях.

В тексте статьи ссылки на литературу указываются цифрами в квадратных скобках, при необходимости указываются номера страниц, например: [2; 3. С. 15].

Список литературы нумеруется в порядке цитирования и оформляется в соответствии с ГОСТ Р 7.0.5-2008 на библиографическое описание (знаки тире в описании опускаются). Ссылки на неопубликованные работы, а также на Интернет-ресурсы (кроме электронных изданий, поддающихся библиографическому описанию) оформляются в виде сноски.

В Список литературы ссылки на источники следует включать на оригинальном языке опубликования. Каждый источник должен быть также оформлен на английском языке (References) по международному стандарту для публикаций в области информатики IEEE Style со следующими отличиями:

- инициалы авторов указываются после фамилии;
- название статьи не берется в кавычки, отделяется точкой;

- отсутствует союз «and» перед фамилией последнего автора;
- в диапазоне страниц – удвоенная «р» (например, «pp. 2–9»);
- год издания указывается после места издания (для книг) и сразу после названия журнала (для периодики).
- Перевод источника на английский язык:
- если источник имеет выходные данные на английском языке, то для формирования References **следует использовать именно эти данные**;
- если оригинальная публикация не содержит выходных данных на английском языке, то допускается транслитерация названия материала на латинский алфавит в сочетании с переводом на английский язык в квадратных скобках. В конце описания указывается, на каком языке написана эта работа, например, (in Russ.). При транслитерации можно воспользоваться Интернет-ресурсом <http://ru.translit.ru/>, рекомендуется выбрать стандарт BSI. Место издания не транслитерируется, указывается полностью на английском языке, например: Moscow. Название издательства / издателя, как правило, транслитерируется. Для журналов, у которых есть официальное название на английском языке, – использовать его (проверить на сайте журнала, или, например, в библиотеке WorldCat), если названия на английском языке нет, использовать транслитерацию по системе BSI. Не следует самостоятельно переводить названия журналов.

Если у цитируемого источника есть **цифровой идентификатор DOI** (<https://search.crossref.org/>), его требуется обязательно указывать в конце библиографической ссылки.

Примеры оформления ссылок. Каждый источник в том же пункте дублируется на английском языке (References).

Источник на русском языке, перевод на английский доступен в метаданных статьи

1. Журавлев С. С., Рудометов С. В., Окольников В. В., Шакиров С. Р. Применение модельно-ориентированного проектирования к созданию АСУ ТП опасных промышленных объектов // Вестник НГУ. Серия: Информационные технологии. 2018. Т. 16, № 4. С. 56–67. DOI 10.25205/1818-7900-2018-16-4-56-67

Zhuravlev S. S., Rudometov S. V., Okolnishnikov V. V., Shakirov S. R. Model-Based Design Approach for Development Process Control Systems of Hazardous Industrial Facilities. *Vestnik NSU. Series: Information Technologies*, 2018, vol. 16, no. 4, pp. 56–67. (in Russ.) DOI 10.25205/ 1818-7900-2018-16-4-56-67

Источник на английском языке. Оформляем согласно требованиям для References. Приводим только 1 раз.

2. Telnov V. I. Optimization of the Beam Crossing Angle at the ILC for E + e- and yy Collisions. *Journal of Instrumentation*, 2018, vol. 13, no. 03, pp. P03020–P03020. DOI 10.1088/1748-0221/13/03/p03020

Метаданные источника доступны только на русском языке

3. Жижимов О. Л., Федотов А. М., Шокин Ю. И. Технологическая платформа массовой интеграции гетерогенных данных // Вестник НГУ. Серия: Информационные технологии. 2013. Т. 11, вып. 1. С. 24–41.

Zhizhimov O. L., Fedotov A. M., Shokin Yu. I. Tekhnologicheskaya platforma massovoi integratsii geterogennykh dannykh [Technology Platform for the Mass Integration of Heterogeneous Data]. *Vestnik NSU. Series: Information Technologies*, 2013, vol. 11, no. 1, pp. 24–41. (in Russ.)

Сведения об авторах

Последний раздел статьи – информация об авторе / авторах **на русском и английском языках**:

- ФИО полностью, ученая степень, ученое звание;
- идентификаторы автора, такие как ResearcherID (всем авторам рекомендуется использовать данные сервисы для ведения актуального списка своих публикаций);
- контактный телефон (не публикуется).

Если статья представляется на английском языке, необходимо приложить перевод на русский язык названия, аннотации, ключевых слов, сведений об авторе.

Доставка материалов

Материалы предоставляются в редакцию по электронной почте inftech@vestnik.nsu.ru.

Порядок рецензирования

Все статьи сначала проходят проверку на заимствование и только после этого отправляются на рецензирование. Редакционный совет не допускает к публикации материал, если имеется достаточно оснований полагать, что он является плагиатом.

Тип рецензирования статей – двухуровневое, одностороннее анонимное («слепое»).

Для каждой статьи редколлегией выбираются рецензенты, научная деятельность которых связана с темой представленного материала. Ответственный секретарь журнала обращается к ним с просьбой дать экспертную оценку статье либо помочь организовать рецензирование.

Рецензии для журнала «Вестник НГУ. Серия: Информационные технологии» составляются по единой схеме и подразумевают оценку по следующим критериям: соответствие тематике журнала, оригинальность и значимость результатов, качество изложения материала.

Заполненный бланк рецензии высылается на электронный адрес редакции. В зависимости от экспертных заключений статья может быть принята редакционным советом к опубликованию, рекомендована автору к доработке (с последующим повторным рецензированием либо без него) или отклонена (с предоставлением автору мотивированного отказа). Автору на электронный адрес высылается текст рецензии без указания ФИО рецензента и его контактных данных.

Все рецензии хранятся в редакции журнала не менее 5 лет. Редколлегия журнала обязуется при поступлении соответствующего запроса направлять копии рецензий в Министерство науки и высшего образования Российской Федерации.