ВЕСТНИК
НОВОСИБИРСКОГО ГОСУДАРСТВЕННОГО УНИВЕРСИТЕТА

Научный журнал
Основан в ноябре 1999 года

Серия: Информационные технологии 2024. Том 22, № 4

СОДЕРЖАНИЕ

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)
Вестник НГУ. Серия: Информационные технологии. 2024. Том 22, № 4
Vestnik NSU. Series: Information Technologies, 2024, vol. 22, no. 4

V E S T N I K
NOVOSIBIRSK STATE UNIVERSITY

Scientific Journal
Since 1999, November
In Russian

Series: Information Technologies 2024. Volume 22, № 4

CONTENTS

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)
Вестник НГУ. Серия: Информационные технологии. 2024. Том 22, № 4
Vestnik NSU. Series: Information Technologies, 2024, vol. 22, no. 4

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

ISSN 1818-7900 (Print). ISSN 2410-0420 (Online)
Вестник НГУ. Серия: Информационные технологии. 2024. Том 22, № 4
Vestnik NSU. Series: Information Technologies, 2024, vol. 22, no. 4

Научная статья

УДК 004.624
DOI 10.25205/1818-7900-2024-22-4-5-16

О повышении качества выходных данных в вопросно-ответной системе, обрабатывающей климатическую информацию

Ольга Юрьевна Гавенко[1]
Наталья Александровна Шашок[2]

[1,2]Федеральный исследовательский центр информационных и вычислительных технологий, Новосибирск, Россия

[1]Новосибирский государственный университет
Новосибирск, Россия

[1]olga.yu.gavenko@mail.ru, https://orcid.org/0000-0003-3619-1120
[2]n.shashok@alumni.nsu.ru, https://orcid.org/0009-0007-3658-6110

Аннотация

Разработка вопросно-ответной системы (QA), обрабатывающей климатическую информацию, опирается на использование разнородных климатических данных в различных форматах (текстовые, числовые, графические, видео, аудио, географические и данные мониторинга). Обязательным элементом вопросно-ответной системы должен являться инструмент, позволяющий обрабатывать и анализировать подобные данные.

Процессы поиска и извлечения данных выступают центральной частью рассматриваемой системы, поскольку от них во многом зависит качество сгенерированного ответа. Точный способ извлечения данных имеет решающее значение для выходных данных системы QA, а также для проблем принятия решений, так как существуют ситуации, в которых LLM генерирует ответы, соответствующие контексту, но фактически являющиеся неверными и не соответствующими входным данным. Использование правильных метрик и алгоритмов для некоторых типов данных и неправильных для других может привести к превышению допустимого порога нерелевантных данных, что, в свою очередь, может снизить качество ответов. Дополненная поисковая генерация (Retrieval-augmented Generation, RAG) также может использоваться для оптимизации входных данных для этой задачи.

В работе рассматриваются различные алгоритмы извлечения данных и ранжирования документов, а также возможность использования ансамблей агентов LLM при разработке вопросно-ответной системы, обрабатывающей климатическую информацию.

Ключевые слова

оптимизация входных данных, вопросно-ответные системы, разработка RAG-системы, обработка мультимодальных данных

Для цитирования

Гавенко О. Ю., Шашок Н. А. О повышении качества выходных данных в вопросно-ответной системе, обрабатывающей климатическую информацию // Вестник НГУ. Серия: Информационные технологии. 2024. Т. 22, № 4. С. 5–16. DOI 10.25205/1818-7900-2024-22-4-5-16

On Increasing the Quality of the Climate Observations Question-Answering System's Output Data

Olga Yu. Gavenko[1], Natalia A. Shashok[2]

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

[1]Novosibirsk State University,
Novosibirsk, Russian Federation

[1]olga.yu.gavenko@mail.ru, https://orcid.org/0000-0003-3619-1120
[2]n.shashok@alumni.nsu.ru, https://orcid.org/0009-0007-3658-6110

Abstract

The development of the climate observations question-answer (QA) information system relies on heterogeneous climate data in various formats (text, numerical, graphic, video, audio, geographic and monitoring data). A mandatory element of such a system is a tool that allows processing and analyzing such data.

Searching and retrieving data is a central part of the system in question, since the quality of the generated answer heavily depends on it. The exact way the data is retrieved is critical to the output of a QA system as well as to decision-making problems, since there are situations in which the LLM generates a contextually appropriate but factually incorrect answers that do not match the input. Using correct metrics and algorithms for some data types and incorrect ones for others can cause the permissible threshold of irrelevant data to be exceeded, which in turn can cause the quality of the answers to decrease. Retrieval-augmented generation (RAG) systems can also be used to optimize input data for that task.

This work discusses various algorithms for data extraction and document ranking, as well as the possibility of using ensembles of LLM agents in development of the QA system that works with climate data.

Keywords

Input data optimization, question answering systems, RAG system development, multimodal data processing.

For citation

Gavenko O., Shashok N. On increasing the quality of the climate observations question-answering system's output data. *Vestnik NSU. Series: Information Technologies*, 2024, vol. 22, no. 4, pp. 5–16 (in Russ.) DOI 10.25205/1818-7900-2024-22-4-5-16

Введение

Разработка информационной системы типа «вопрос-ответ», обрабатывающей и анализирующей климатические данные, соответствует целевому направлению Климатической доктрины Российской Федерации от 26 октября 2023 г.[1], определяющей климатическую политику Российской Федерации.

Климатические данные могут быть получены как из внешних источников (систем мониторинга и глобальной сети), так и из доступных внутренних хранилищ, и могут существовать в различных форматах (текстовые, числовые, графические, видео-, аудио-, географические и данные мониторинга), при этом для обработки должны быть доступны не только имеющиеся в хранилищах данные, но и данные, поступающие в систему в непрерывном режиме. Обязательным элементом подобной информационной системы должен быть инструмент, позволяющий обрабатывать и анализировать разнородные динамические и статические данные с целью их использования в алгоритмах построения и генерации ответов для решения широкого круга задач, связанных с поддержкой принятия решений.

Очевидно, что вопросы, на которые вопросно-ответная система, обрабатывающая климатические данные, должна быть способна ответить, могут быть различной сложности. Так, вопрос, направленный на определение автора какой-либо конкретной научной работы, требует обработки только этой самой работы, если она существует; для ответа на вопрос о том, находится ли озеро Байкал в Уральских горах, достаточен будет один документ из достовер-

[1] Russia's new Climate Doctrine approved, http://www.en.kremlin.ru/acts/news/72598

ного источника, описывающий местоположение озера Байкал. Однако существуют вопросы, требующие изучения большого количества информации, в частности, вопросы, связанные с проведением сравнительного анализа каких-либо данных в течение некоторого периода времени. В качестве примера можно рассмотреть следующую задачу: определение совокупного количества CO_2, образующегося на конкретной территории в Российской Федерации в связи с работой промышленных предприятий региона. Для ответа на подобный вопрос может потребоваться изучение нескольких научных статей, в которых есть информация по каждому предприятию, но за отсутствием необходимых статей могут использоваться данные мониторинга и спутниковой съемки, выявляющей изменения содержания CO_2 в атмосфере за некоторый период.

Система может не предусматривать получения ответов на риторические вопросы и вопросы по более широкой тематике; представляется допустимым, что при попытке ответить на вопрос, не связанный с целевой тематикой системы, система может либо сгенерировать слабо связанный с вопросом ответ, либо полностью отказаться от его генерации. Подобное поведение вполне соответствует самому определению задачи автоматического нахождения ответа на вопросы [1]: это задача, которая направлена на генерацию правильного ответа на вопросы, специфичные для некоторой предметной области, на основе заданного контекста или базы знаний.

В эпоху развития глубокого обучения и больших языковых моделей задача генерации ответов может быть решена более сложными методами по сравнению с такими классическими методами, как прямое сопоставление, поиск ключевых слов, разметка частей речи и разбор фрагментов, использовавшимися ранее, начиная с 1970-х годов [2; 3]. Тем не менее основной критерий решения задачи не меняется: это построение правильного и корректного ответа на вопросы в предметной области и, возможно, вне предметной области, с помощью предварительной обработки входных данных и последующего использования результата для поиска в базе данных предварительно обработанных документов. Основное различие с более ранними разработками проявляется в использовании современных подходов к вычислению и предварительной обработке данных. Такие подходы включают преобразование или связывание частей документов различных типов с векторами и поиск частей документов или данных с нужным контекстом в общем векторном пространстве, а затем их использование для предоставления необходимого контекста для использования в некоторой большой языковой модели (LLM), такой как GPT[2], Gemini[3], Gemma[4], Falcon[5] и других.

Подход к решению задачи генерации ответа на вопрос, рассматриваемый в представленной работе – Retrieval Augmented Generation (RAG) [4] – заключается в следующем. На первом этапе используется система поиска векторной информации (IR) для получения документов, релевантных запросу пользователя, после чего применяется модуль извлечения информации (IE) для фильтрации данных, а именно отсечения нерелевантных запросу данных и осуществление выборки необходимых для предоставления контекста частей документов. На заключительном этапе полученная информация объединяется с запросом пользователя для формирования полного контекста, необходимого для корректной генерации ответа. При подобном подходе система IR обычно предоставляет пользователю возможность генерировать ответы самостоятельно на основе найденных документов. Модуль IE, а также LLM, генерирующая ответ, в свою очередь, могут автоматизировать задачу извлечения информации.

Модель LLM может использовать API поиска векторного хранилища для генерации более точных ответов, которые пользователь получает после генерации более краткого вопроса для LLM. Общий вид такого потока данных представлен на рис. 1.

[2] GPT-4 | OpenAI. https://openai.com/index/gpt-4/

[3] Gemini, https://gemini.google.com/

[4] Google AI Gemma open models - Gemini API, https://ai.google.dev/gemma

[5] Falcon LLM, https://falconllm.tii.ae

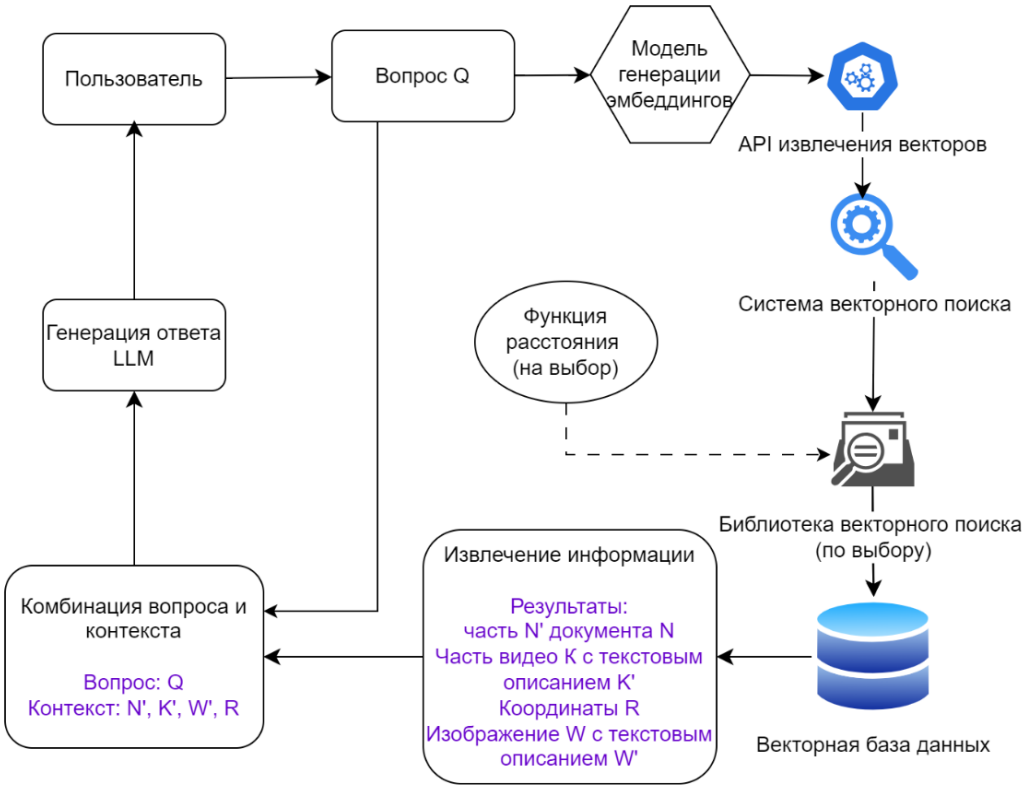


Рис. 1. Общий вид потока данных вопросно-ответной системы, обрабатывающей климатическую информацию
Fig. 1. A general view into the data flow within the retrieval augmented generation architecture for the climate observation question answering system

С учетом того, что объем данных, используемых разрабатываемой системой, может быть довольно большим, хранение данных в разных хранилищах и объединение их с помощью нескольких модулей с одним и тем же API-интерфейсом, зависящим от типа входных данных, на основе которых определялись эмбеддинги, представляется целесообразным, что изменяет поток данных, как показано на рис. 2.

Можно отметить несколько причин, по которым разделение хранилищ по типу или происхождению данных с последующим их объединением под одним API представляется обоснованным. Во-первых, общий размер набора данных не должен сильно ухудшать параметр скорости ответа системы на ввод запроса пользователем. Системы, подобные разрабатываемой системе обработки климатических данных, на данный момент используют весьма разный объем документов: от 35–50 тысяч отдельных утверждений[6] до миллионов документов, новостных каналов и сообщений из сети Интернет (например, MediSearch[7]). Разнообразные наборы данных, используемые для разработки подобных вопросно-ответных систем, такие как Cohere's Wikipedia Embeddings[8], CORD-19 [5], NewsQA[9], SQuAD [6] и другие, отличаются большим объемом данных, и со временем этот объем растет, поскольку эти наборы данных расширяются их разработчиками. Во-вторых, предоставление системе возможности выбирать, какие именно данные ей нужны, и указывать точный источник данных, с большой долей веро-

[6] SberQuAD (Sberbank Question Answering Dataset), https://paperswithcode.com/dataset/sberquad

[7] MediSearch, https://medisearch.io

[8] Wikipedia (en) embedded with cohere.ai multilingual-22-12 encoder, https://huggingface.co/datasets/Cohere/wikipedia-22-12-en-embeddings

[9] Dataset Card for NewsQA, https://huggingface.co/datasets/Maluuba/newsqa

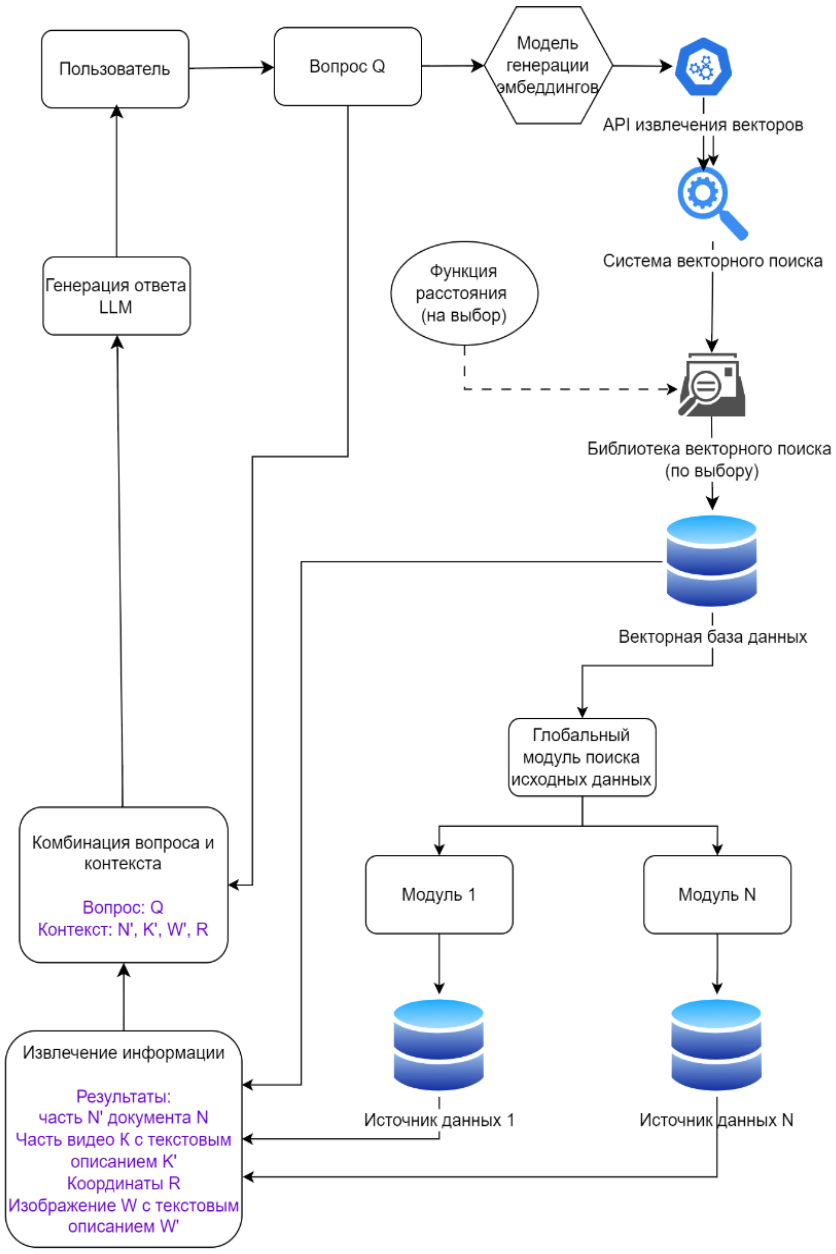


Рис. 2. Уточненный вид потока данных вопросно-ответной системы, обрабатывающей климатическую информацию
Fig. 2. A more detailed view into the data flow within the retrieval augmented generation architecture for the climate observation question answering system

ятности снизит нагрузку на сами источники данных, поскольку, при необходимости получить данные только из одного источника, запросы к остальным источникам осуществляться не будут. В-третьих, при административной поддержке системы также может быть полезно знать, какой именно источник данных содержит определенные данные.

Следует подчеркнуть, что описанный выше подход в его базовой постановке может быть недостаточным. Известна проблема так называемых «галлюцинаций», когда модель использует по большей части только «лучший», по своему «мнению», полученный результат данных без дополнительного анализа или верификации, и при этом генерирует фактически не-

верные ответы [7]. Примером таких «галлюцинаций» может быть текущее состояние ИИ LLM Google, который был обучен использовать лучшие результаты из всех доступных веб-документов без последующего анализа на достоверность источника[10], а также на релевантность и корректность предоставленных данных. Это приводит к тому, что LLM Google рассматривает веб-документы, содержащие недостоверную информацию, в том числе, написанную пользователями сети Интернет для умышленного введения других людей в заблуждение, как заслуживающую использования при генерации ответов. В идеальном случае алгоритм PageRank [8] должен гарантировать, что подобных ситуаций происходить не будет, однако существуют примеры обратного[11].

Учитывая вышеизложенное, планируется первоначально протестировать описываемую RAG-систему на базе знаний, состоящей, как минимум, из 40 тысяч документов, разделенных между двумя узлами источников данных. На данный момент такое количество документов нельзя назвать достаточным для конечной системы, однако использование небольшого количества документов представляется разумным в целях тестирования как системы, так и проверки документов на достоверность. Впоследствии предполагается расширение базы знаний и такой разработки архитектуры системы и ее реализации, чтобы она могла потенциально обрабатывать как минимум такое же количество документов, которое содержится в наборе данных CORD-19.

Подходы к решению проблемы

Следует подробнее рассмотреть некоторые вопросы, возникающие при описанных выше проблемах вопросно-ответной системы, обрабатывающей климатическую информацию: в частности, проблему повышения качества генерации ответов на вопросы и задачи, возникающие при разработке подхода к извлечению информации.

На данный момент не существует единого общепризнанного проверенного подхода повышения качества генерации ответов с помощью больших языковых моделей. Тем не менее представляется возможным выстроить систему таким образом, чтобы можно было эмпирически найти необходимые параметры, при которых система показала бы хорошие результаты, поскольку существуют некоторые отдельные решения повышения качества генерируемых ответов, и комбинация этих подходов могла бы дать более весомый результат, чем использование только одного.

Рассмотрим эти подходы детально.

Правильное извлечение информации

Перед генерацией ответа необходимо правильным образом извлекать информацию, подаваемую в качестве контекста LLM. Настраиваться может в том числе и сам процесс информационного поиска, например, через выбор библиотек и мер близости, которые используются для поиска документов в векторном пространстве.

Помимо этого, необходимо учитывать, что база знаний разрабатываемой вопросно-ответной системы хранения и обработки климатической информации является принципиально мультимодальной, поскольку она может содержать не только текстовые данные, но и карты, данные климатических наблюдений, изображения, видео и звуки. Примерами таких данных являются, помимо прочего, экологические словари, такие как [9], данные, полученные от Федеральной службы по гидрометеорологии и мониторингу окружающей среды (карты, новости,

[10] https://support.google.com/websearch/answer/14901683

[11] https://www.nytimes.com/2024/05/24/technology/google-ai-overview-search.html

таблицы)[12], данные повторного анализа системы климатического прогнозирования[13], научные статьи открытого доступа из таких проектов, как CyberLeninka[14] и др.

Определение «мультимодальные данные», которое дается данным разных форматов в ряде ИТ-задач, отражает совокупность разнородных документов, обрабатываемых и анализируемых современными информационными системами. Количество задач, требующих разработки подходов к обработке специфичных мультимодальных данных, увеличивается из года в год, и это связано, прежде всего, с преимуществом систем, способных анализировать данные разных форматов, в сравнении с системами, обрабатывающими только один тип данных, поскольку подобное характерно для самых разных областей современной науки и техники. В ряде задач схожие данные определяются как гетерогенные данные, что также отражает суть проблемы: в задачах, в которых должны использоваться различные данные, имеющие разную структуру, источники и методы обработки, принципиально важно определить подход, учитывающий их аутентичность, и выбрать соответствующие инструменты, позволяющие одновременно обрабатывать и анализировать данные разных форматов, а затем использовать полученные результаты в вопросно-ответных системах для поддержки принятия решений.

Для задач, связанных с обработкой климатической информации вопросно-ответными системами, способность системы обрабатывать мультимодальные данные является центральной, учитывая, что климатическая информация может поступать из различных источников, которые возможно сгруппировать по ряду признаков. В общем случае климатические данные могут быть следующих форматов: текстовые, числовые, графические, видео-, аудио-, картографические данные и данные мониторинга. Помимо этого, всю совокупность климатических данных можно разделить на две части. Во-первых, это статические данные, загружаемые из источников, находящихся в фиксированных и редко обновляемых базах данных, содержащих словари, тезаурусы, справочную литературу и т. д., т. е. источники, информация в которых слабо изменяется по своему содержанию, например, словарные статьи, содержащие текстовые и графические данные. Во-вторых, динамические данные, которые изменяются во временных интервалах, накапливаются или удаляются как неактуальные, – это данные мониторинга, а также картографические данные. Очевидно, что графические данные могут быть как частью статической информации, если они входят в словарную статью, так и изменяемыми в хронологическом аспекте, если они получены в результате обработки данных мониторинга. Таким образом, для корректной обработки информации необходимо дополнительно учитывать источники данных, их тип и характер.

Однако такая неоднородность данных, как по типу, так и по характеру, усложняет и выбор инструментов, и разработку архитектуры вопросно-ответной системы, работающей с мультимодальными данными, а также привносит другие проблемы, поскольку использование корректных метрик и алгоритмов для одних типов данных и некорректных для других может привести к превышению допустимого порога нерелевантных данных, что в свою очередь потенциально приводит к снижению качества ответа. Очевидно, что при проектировании способ генерации эмбеддингов для привязки к документам должен быть определен на основе используемого подхода к извлечению данных.

Один из таких подходов состоит в следующем. Системы обработки мультимодальных данных могут использовать различные алгоритмы извлечения и ранжирования для каждого из используемых типов данных, с учетом того, что различные типы данных в различных хранилищах могут храниться с использованием неидентичных векторных пространств. Противоположным подходом к извлечению данных могло бы быть объединение данных из различных

[12] ЕИП Росгидромета, https://eip.meteo.ru/opendata

[13] Climate Data Guide: Climate Forecast System Reanalysis (CFSR), https://climatedataguide.ucar.edu/climate-data/climate-forecast-system-reanalysis-cfsr

[14] Научная электронная библиотека «КиберЛенинка», https://cyberleninka.ru

модальностей в некоторую однородную форму – либо с использованием общего векторного пространства, либо путем координирования различных пространств, и урезания либо расширения мерности. Последний подход представляется многообещающим при работе с неоднородными данными сильно отличных между собой типов.

Определение уровня шума в выдаче

Документ, возвращаемый системой извлечения данных, может быть релевантным запросу, но при этом устаревшим, связанным с обозначенной тематикой, но не содержащим ответ на вопрос, либо вообще случайным. Последние рассматриваются как шум, от которого необходимо очистить данные для повышения качества сгенерированного ответа.

Однако некоторые исследования показывают, что это может быть не так [10], и некоторые нерелевантные документы, рассматриваемые как шум, имеют возможность повысить качество генерации, если они правильным образом размещены в контекст, подаваемый на вход LLM, в то время как семантически связанные документы, не содержащие ответа, значительно снижают качество. Важно отметить, что добавление шума ухудшает некоторые метрики оценки поиска, такие как fall-out rate; из этого можно сделать вывод, что в разрабатываемую систему необходимо добавить механизм, позволяющий при генерации ответа выбрать уровень «шума» в выдаче контекста в целях поиска границы нерелевантности, после которой качество ответов начинает падать.

Использование ансамблей

Еще одним подходом к улучшению результатов поиска является одновременное использование нескольких различных методов поиска и объединение результатов с помощью голосования или других методов комбинирования, этот подход называется ансамблем [11]. При таком подходе можно упорядочить части контекста при подаче на вход LLM таким образом, который может быть недостижимым при использовании одного метода, что, в свою очередь, может повлиять на качество сгенерированных ответов. Для проверки этого утверждения представляется целесообразным добавить возможность настройки включения пользователем использования ансамблей и выбора алгоритмов голосования при составлении контекста для генерации ответа.

Методы генерации эмбеддингов

Учитывая, что некоторые документы могут содержать большой объем данных, необходимо определить ответы на следующие вопросы: какую часть документа следует преобразовать в эмбеддинг; следует ли каждому документу составлять некоторое краткое описание или необходимо разбивать документы на фрагменты; насколько большим должно быть описание или часть документа; можно ли использовать два этих подхода совместно, несмотря на потерю данных при составлении обобщения; следует ли ранжировать или классифицировать фрагменты перед генерацией по ним эмбеддингов.

Выбор правильного для обработки размера фрагмента представляется целесообразным по следующим причинам. Небольшой размер фрагмента может обеспечить более высокую детализацию, точно указав необходимый контекст в документах. С другой стороны, это может означать, что некоторый контекст может быть упущен. Использование же фрагментов крупных размеров может определять большие затраты по времени при генерации ответа, поскольку больше информации передается LLM для обработки. В целях тестирования полезно иметь несколько векторных индексов в векторной базе данных или непересекающиеся векторные пространства, где каждый индекс либо векторное пространство соответствует разному раз-

меру фрагмента, и позволить пользователю выбирать, с каким размером фрагмента он хочет работать; однако это потенциально может привести к большой нагрузке на базу данных, поскольку размер используемых данных растет при увеличении количества индексов и векторных пространств.

Описанные выше подходы по фрагментированию данных перед их переводом в эмбеддинги могут быть актуальны не только для текстовых данных, но и для карт, данных мониторинга, изображений и видео; например, часть карты, находящейся в процессе преобразования во встраивания, может быть связана конкретно с Алтайским краем, другая часть может быть связана с Москвой, и эти части могут быть расположены в разных частях векторного пространства.

Плановый процесс загрузки данных и создания эмбеддингов в системе представлен на рис. 3.

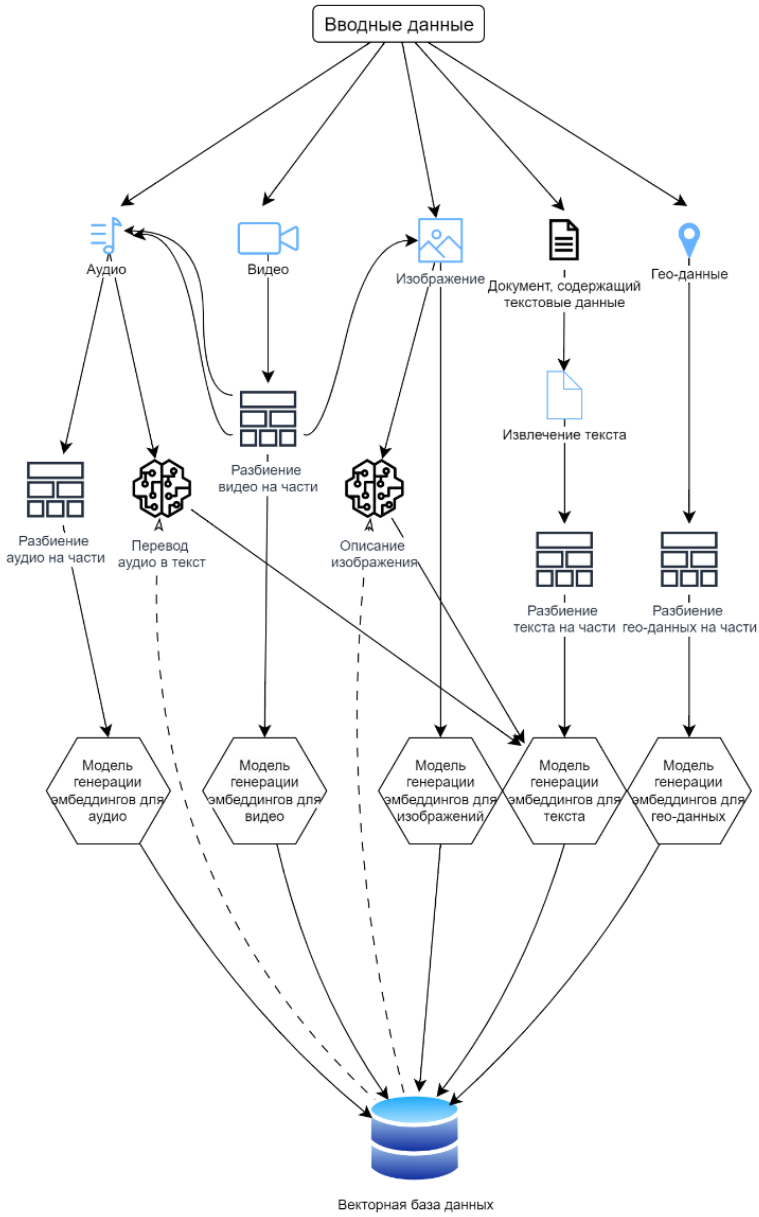


Рис. 3. Целевой вариант процесса загрузки данных и генерации эмбеддингов для дальнейшего извлечения и поиска информации

Fig. 3. The proposed process of data uploading and embeddings generation for further extraction and information retrieval

Заключение

Таким образом, учитывая вышеизложенное, можно сделать вывод, что при проектировании механизма поиска информации для вопросно-ответной системы, обрабатывающей климатические данные, необходимо предоставить конечному пользователю возможность устанавливать уровень шума, определять, как именно извлекать данные, использовать ли ансамблирование, какой тип данных использовать, в том числе выбирать векторное расстояние. Это даст возможность всесторонне протестировать модуль генерации ответов и определить, какие параметры извлечения данных обеспечивают наибольшую корректность сгенерированных ответов.

Список литературы

1. **Hirschman L., Gaizauskas R.** Natural language question answering: the view from here // Natural Language Engineering Journal. 2001. Vol. 7, no. 4. P. 275–300. DOI: 10.1017/S1351324901002807
2. **Keen P. G. W, Michael S. S. M.** Decision support systems: an organizational perspective. Michigan, Addison-Wesley, 1978.
3. **Woods W. A.** Progress in natural language understanding: an application to lunar geology // Proceedings of the national computer conference and exposition (AFIPS ‘73), 1974, Association for Computing Machinery, New York, NY, USA, p. 441–450. DOI: https://doi.org/10.1145/1499586.1499695
4. **Lewis P., Perez E., et al.** Retrieval-augmented generation for knowledge-intensive NLP tasks // Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS ‘20), 2020, Curran Associates Inc., Red Hook, NY, USA, Article 793, p. 9459–9474. DOI: 10.48550/arXiv.2005.11401
5. **Wang L., Lo K. et al.** CORD-19: The COVID-19 Open Research Dataset. ArXiv, abs/2004.10706, 2020. DOI: 10.48550/arXiv.2004.10706
6. **Rajpurkar P., Zhang J., Lopyrev K., Liang P.** Squad: 100,000+ questions for machine comprehension of text // Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, 2016, Association for Computational Linguistics, Austin, Texas, USA, p. 2383–2392. DOI: 10.18653/v1/D16-1264
7. **Magesh V., Surani F., Dahl M., Suzgun M. et al.** Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. ArXiv, abs/2405.20362, 2024. DOI: 10.48550/arXiv.2405.20362
8. **Page L., Brin S., Motwani R., Winograd T.** The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.
9. **Фадеев С. В.** Экологический словарь. СПб., 2011.
10. **Florin C., Giovanni T. et al:** The Power of Noise: Redefining Retrieval for RAG Systems // Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2024, Association for Computing Machinery, New York, NY, USA, p. 719–729. DOI: 10.1145/3626772.3657834
11. **Cormack G. V., Clarke C. L., Büttcher S.** Reciprocal rank fusion outperforms condorcet and individual rank learning methods // Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval, 2009, Association for Computing Machinery, New York, NY, USA, p. 758–759. DOI: 10.1145/1571941.1572114

References

1. **Hirschman L., Gaizauskas R.** Natural language question answering: the view from here. Natural Language Engineering Journal, 2001, vol. 7, no. 4, pp. 275–300. DOI: 10.1017/S1351324901002807
2. **Keen P. G. W, Michael S. S. M.** Decision support systems: an organizational perspective. Michigan, Addison-Wesley, 1978.
3. **Woods W. A.** Progress in natural language understanding: an application to lunar geology. *Proceedings of the national computer conference and exposition (AFIPS '73)*, 1974, Association for Computing Machinery, New York, NY, USA, pp. 441–450. DOI: https://doi.org/10.1145/1499586.1499695
4. **Lewis P., Perez E., et al.** Retrieval-augmented generation for knowledge-intensive NLP tasks. *Proceedings of the 34th International Conference on Neural Information Processing Systems (NIPS '20)*, 2020, Curran Associates Inc., Red Hook, NY, USA, Article 793, pp. 9459–9474. DOI: 10.48550/arXiv.2005.11401
5. **Wang L., Lo K. et al.** CORD-19: The Covid-19 Open Research Dataset. ArXiv, abs/2004.10706, 2020. DOI: 10.48550/arXiv.2004.10706
6. **Rajpurkar P., Zhang J., Lopyrev K., Liang P.** Squad: 100,000+ questions for machine comprehension of text. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, 2016, Association for Computational Linguistics, Austin, Texas, USA, pp. 2383–2392. doi: 10.18653/v1/D16-1264
7. **Magesh V., Surani F., Dahl M., Suzgun M. et al.** Hallucination-Free? Assessing the Reliability of Leading AI Legal Research Tools. ArXiv, abs/2405.20362, 2024. DOI: 10.48550/arXiv.2405.20362
8. **Page L., Brin S., Motwani R., Winograd T.** The PageRank Citation Ranking: Bringing Order to the Web. Technical Report, Stanford InfoLab, 1999.
9. **Fadeev S. V.** Ekologicheskij slovar'. Saint Petersburg, 2011 (in Russ.)
10. **Florin C., Giovanni T., et al:** The Power of Noise: Redefining Retrieval for RAG Systems. *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2024, Association for Computing Machinery, New York, NY, USA pp. 719-729. DOI: 10.1145/3626772.3657834
11. **Cormack G. V., Clarke C. ., Büttcher S.** Reciprocal rank fusion outperforms condorcet and individual rank learning methods. *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, 2009, Association for Computing Machinery, New York, NY, USA, pp. 758–759. DOI: 10.1145/1571941.1572114

Сведения об авторах

Гавенко Ольга Юрьевна, доктор технических наук, кандидат филологических наук, ведущий научный сотрудник Федерального исследовательского центра информационных и вычислительных технологий; старший преподаватель кафедры математического моделирования Новосибирского государственного университета

Шашок Наталья Александровна, аспирант Федерального исследовательского центра информационных и вычислительных технологий

Information about the Authors

Olga Yu. Gavenko, Doctor of Sciences (Technical Sciences), Candidate of Sciences (Philology), Leading Researcher, Federal Research Center for Information and Computational Technologies. Senior lecturer of the Department of Mathematical Modeling, Novosibirsk State University

Natalia A. Shashok, Ph. D Student. Federal Research Center for Information and Computational Technologies

Статья поступила в редакцию 07.12.2024;
одобрена после рецензирования 26.12.2024; принята к публикации 26.12.2024

The article was submitted 07.12.2024;
approved after reviewing 26.12.2024; accepted for publication 26.12.2024.

Научная статья

УДК 004.853
DOI 10.25205/1818-7900-2024-22-4-17-32

Мультимедиалекция с интеллектуальным педагогическим агентом – новый подход к организации лекционной работы в вузе

Виталий Геннадьевич Казаков

Новосибирский государственный университет
Новосибирск, Россия

Сочинский государственный университет
Сочи, Россия

vit.kazakov60@yandex.ru, https://orcid.org/0000-0002-0910-9315

Аннотация

Статья посвящена вопросам цифровизации лекционной формы работы в вузе. Показано, что ряд проблем, связанных с современными социальными и личностными трансформациями, ведет к снижению эффективности классической лекции. Предлагается новый подход к лекционному процессу, основанный на применении в учебном процессе мультимедиалекций с интеллектуальным педагогическим агентом. Показано, что применение таких лекций может способствовать повышению эффективности учебных занятий. Для практической проверки эффекта применения подхода осуществляется проектирование и реализация системы обучения, обеспечивающей создание и применение мультимедиалекций. Приводятся сведения об инфологической модели, положенной в основу системы, и ее облачной архитектуре. Описывается работа демонстрационного прототипа проигрывателя мультимедиалекций, включая функциональность педагогического агента. Делается вывод о возможности построения проигрывателя, включая функциональность педагогического агента, работающего на компьютерных устройствах бытового уровня. Приводятся планы дальнейших исследований.

Ключевые слова

учебная лекция, электронное обучение, мультимедиатехнологии, педагогический агент, интеллектуальный агент, искусственный интеллект, распознавание лиц, анализ и синтез речи

Финансирование

Исследование выполнено за счет финансовой поддержки (гранта) исследовательских центров, предоставленной Автономной некоммерческой организацией «Аналитический центр при Правительстве Российской Федерации», идентификатор соглашения о предоставлении субсидии 000000D730324P540002, договор о предоставлении гранта с Новосибирским государственным университетом от 27.12.2023 № 70-2023-001318.

Для цитирования

Казаков В. Г. Мультимедиалекция с интеллектуальным педагогическим агентом – новый подход к организации лекционной работы в вузе // Вестник НГУ. Серия: Информационные технологии. 2024. Т. 22, № 4. С. 17–32. DOI 10.25205/1818-7900-2024-22-4-17-32

Multimedia Lecture with an Intelligent Pedagogical Agent – a New Approach to Organizing Lectures at a University

Vitaliy G. Kazakov

Novosibirsk State University,
Novosibirsk, Russian Federation

Sochi State University,
Sochi, Russian Federation

vit.kazakov60@yandex.ru, https://orcid.org/0000-0002-0910-9315

Abstract

The article is devoted to the issues of digitalization of the lecture form of work at a university. It is shown that a number of problems associated with modern social and personal transformations lead to a decrease in the effectiveness of the classical lecture. A new approach to the lecture process is proposed, based on the use of multimedia lectures with an intelligent pedagogical agent in the educational process. It has been shown that the use of such lectures can help improve the effectiveness of training sessions. To practically test the effect of applying the approach, a training system is designed and implemented to ensure the creation and use of multimedia lectures. Information is provided about the information model underlying the system and its cloud architecture. The operation of a demonstration prototype of a multimedia lecture player, including the functionality of a pedagogical agent, is described. It is concluded that it is possible to build a player, including the functionality of a pedagogical agent running on consumer-level computer devices. Plans for further research are presented.

Keywords

Educational lecture, e-learning, multimedia technologies, pedagogical agent, intelligent agent, artificial intelligence, face recognition, speech analysis and synthesis

Acknowledgements

This work was supported by a grant for research centers, provided by the Analytical Center for the Government of the Russian Federation in accordance with the subsidy agreement (agreement identifier 000000D730324P540002) and the agreement with the Novosibirsk State University dated December 27, 2023 No. 70-2023-001318.

For citation

Kazakov V. G. On increasing the quality of the climate observations question-answering system's output data. *Vestnik NSU. Series: Information Technologies*, 2024, vol. 22, no. 4, pp. 17–32 (in Russ.) DOI 10.25205/1818-7900-2024-22-4-17-32

Введение

Родившись в Средние века, лекция и сегодня остается одной из основных форм учебной работы в университетах. За многие столетия она не претерпела существенных изменений. В то же время очевидно, что многие из основных задач, первоначально возлагаемых на лекцию, потеряли актуальность, поскольку могут в настоящее время эффективно выполняться другими способами.

Так, до массового внедрения книгопечатания конспектирование было способом фиксации знания. Конспект лекций потом мог использоваться своим создателем в течение всей его профессиональной деятельности, поскольку другие формы были труднодоступны и дороги. Распространение книгопечатания ликвидировало эту потребность. Сейчас результат конспектирования редко сохраняется после сдачи студентом экзамена по дисциплине, так как любые информационные материалы − учебные, научные, справочные и т. д. − доступны, например, в библиотечной системе, книготорговой сети, киберпространстве. В частности, практически по любой из дисциплин издаются десятки монографий, называемых «Лекции…», «Цикл лекций…» и т. п., созданные на основе реальных лекционных курсов.

С эпохи промышленной революции и до развития всемирной паутины лекция была средством передачи актуального научного знания, особенно в бурно развивающихся областях. Цик-

лы подготовки и издания научных журналов могли достигать многих месяцев, а сами журналы были труднодоступны. Публикация нового знания в виде научных монографий требовала, помимо прочего, включения в план издательства и занимала обычно не менее года. Учебники, которые издавались, как правило, основывались на опыте прочитанных курсов лекций, и от получения нового знания до его появления в печатной форме проходило уже несколько лет. Важным способом донесения нового знания до выпускников вузов были лекции, на которых профессора, владеющие последними научными результатами, изустно доносили их до студентов. Веб-сервис Интернет, запущенный Бернерсом Ли в 1989−1990 гг., обеспечил прямой и немедленный доступ к самым новым научным результатам, для публикации которых авторам требуется не более нескольких минут.

В связи с этим лекционная форма многим уже не кажется безусловно необходимой и подвергается критике за ряд недостатков, таких как отсутствие обратной связи, усредненность уровня сложности, возможность разной степени включенности в процесс. Хотя мнение, что лекционная форма занятий потеряла свою актуальность, не является сегодня исключительным, вузы еще не нашли альтернативных форм учебных занятий, которые могли бы заместить лекционную форму на практике [1]. Такое положение дел определяется тем, что эффективность лекции во многом зависит от способа передачи информации в ходе межличностного общения. В процессе лекции лектор может контролировать вовлеченность учащихся в учебный процесс, поддерживает мотивацию и внимательность. Таким образом, удобный и быстрый доступ к информации не может заменить классическую учебную лекцию, а любой формат учебных занятий, претендующий на такую роль, должен обеспечить личному участию лектора адекватную замену.

Помимо того, что появляются новые формы распространения информации, которые расширяют возможности обучения, сама лекционная форма работы в значительной мере потеряла свою эффективность. Такая потеря связана не собственно с самой методикой обучения, а с социальными и личностными трансформациями. Такими трансформациями выступают, например, изменение классической образовательной парадигмы на парадигму непрерывного обучения; изменения в ментальной сфере современного человека, такие как клиповое мышление и цифровая амнезия; повышение доступности высшего образования и процентный рост его носителей в обществе.

Цель данной работы − комплексно рассмотреть проблемы учебной лекции, вызванные современной социодинамикой, и возможные подходы к их преодолению. Для этого выделяются три наиболее важные проблемы лекционной формы учебной работы на современном этапе, формулируется новый подход к организации лекционной формы учебной работы в вузе, основанный на применении мультимедиалекций с интеллектуальным педагогическим агентом, и обосновывается его потенциальная эффективность; описываются проектные решения программной системы для создания и применения мультимедиалекций, соответствующие предложенному подходу. Наконец, описывается опыт построения программной системы, которая соответствует подходу, предложенному в статье, и проектным решениям, выполненным в соответствии с данным подходом. Обсуждается стек технологий и дизайн интерфейса, в том числе относящийся к функциональности педагогического агента – ассистента лектора, описываются наиболее принципиальные решения, приводятся сведения о построенном прототипе и его функциональности.

Современные проблемы лекционной формы работы

Хотя указанные выше трансформации – непрерывное образование, клиповое мышление и цифровая амнезия, повышение доли людей, получающих высшее образование, – на первый взгляд весьма различаются, на практике влияние всех этих трансформаций на лекционную

форму ведет к тому, что значительное количество учащихся не посещают лекционные занятия, не прилагают или не могут приложить достаточное количество усилий для усвоения учебного материала.

Поскольку в лекционном курсе, предполагающем систематическое изложение некоторой дисциплины, возможность понимания читаемого материала существенно зависит от уже пройденного, пробелы, получаемые из-за пропусков лекций или потери внимания, критичны. Они затрудняют понимание следующих лекций, что влияет на мотивацию к дальнейшим занятиям. Такая обратная связь приводит к крайне фрагментарному и поверхностному пониманию содержания курса.

Встает вопрос, в чем причина этих явлений? Как мы полагаем, существует ряд объективных проблем, связанных с указанными трансформациями, которые и приводят к такому результату. Три наиболее существенные мы здесь опишем.

Первую проблему возможно назвать «коллизией интересов». Она заключается в том, что представление об обучении в вузе, как главном профессиональном занятии студента, оказалось размыто. Распространенным стало мнение, что хорошо начинать трудовую деятельность как можно раньше, даже на первых курсах вуза, а получаемый на такой работе опыт будет только способствовать успешной учебе. Кроме того, с классическим вузовским обучением сегодня соседствует множество способов дополнительного образования и смежных мероприятий: всяческих инновационных школ, хакатонов, творческих семинаров и т. д. В таких условиях часто возникает коллизия: пойти в вуз на лекцию или на работу (на курсы ДПО, хакатон, конкурс…), которая далеко не всегда решается в пользу вуза. Заметим, что проблема «коллизии представлений» часто возникает у весьма мотивированных студентов, нацеленных на профессиональный рост.

Вторую проблему мы назовем «потерей мотивации». Она выражается в снижении готовности студентов к обучению, как в плане базовой грамотности, так и в части мотивированности. Хотя такое положение дел часто объясняют снижением общего уровня школьного образования, для объяснения потери мотивации таких допущений не требуется. В 1960-х годах в России в вузе студентом вуза становилось около 20 % выпускников, в настоящее время – более 60 %. Таким образом, ранее в вуз поступали студенты, прошедшие серьезный отбор и по уровню полученных знаний, и по мотивированности, и по психологической готовности к напряженному умственному труду и самодисциплине. Так, отбор 20 % всех выпускников соответствует нижней границе IQ приблизительно в 110–115 единиц, 60 % – 95–100 единиц, что является очень существенной разницей. Такую же значительную разницу можно предполагать в мотивированности и способности к умственному труду студентов этих двух временных периодов.

Наконец, третья проблема, назовем ее проблемой «снижения вовлеченности», заключается в том, что современный человек в своей массе значительно менее, по сравнению с людьми предыдущих эпох, готов к восприятию лекционной формы по ряду причин. Во-первых, привычка работы с современной информационной средой, наполненной более аудиовизуальными эффектами, чем рациональными аргументами, влечет за собой так называемое «клиповое сознание», опирающееся на яркие картинки с короткими комментариями и слабо подготовленное к восприятию теоретических конструкций [2]. Во-вторых, в условиях, когда информация легкодоступна по запросу в поисковых системах, ценность лекционного материала представляется учащемуся не слишком высокой, а ведущей стратегией выбирается отказ от запоминания сведений, которые можно найти в Сети. Такое поведение современного человека часто называют «гугл-эффектом», или «цифровой амнезией». Наконец, в-третьих, отсутствие привычки к систематической работе с информацией ведет к тому, что «… у людей ухудшилась память и способность к концентрации», и для них лекционный формат является тяжелым и утомительным [3].

Наиболее популярным ответом на современные проблемы учебной лекции в классическом формате стало использование видеолекций, которые получили достаточно высокую популярность в последние 10–15 лет, особенно в связи с новым видом обучения, получившим название массовых открытых онлайн-курсов (МООК) и развитием Интернета, обеспечившим свободный доступ широких категорий пользователей к видеолекциям [4].

На самом деле интерес к форме видеолекций существует в университетах со времени появления в них аппаратуры записи, воспроизведения и доставки видеосигнала. При наличии такой аппаратуры идея о возможности записи лекции с последующей организацией ее трансляции на целевую аудиторию лежит на поверхности. Одни из первых массовых применений видеолекций относятся к 60-м годам прошлого века и связаны с использованием инфраструктуры системы телевещания. Телевизионные лекции и основанные на базе ТВ-технологий телеуниверситеты решали важную проблему территориальной разделенности учащихся и обеспечивали возможность участия всем лицам, попадающим в зону телевещания, хотя проблема, связанная с определенным временным расписанием, не была решена кардинально. В то время появилось большое количество образовательных программ, курсов, проектов различных типов. Так, до настоящего времени действует телеканал English Club, специализирующийся на образовательных лекциях и передачах в области изучения английского языка.

Развитие телеуниверситетов во многом подготовило почву для использования видеолекций в системах дистанционного обучения, где они распространялись в аналоговом режиме на видеокассетах, цифровом – на CD-ROM и других цифровых носителях, а, с развитием технологий, и через Интернет. При этом в последнем варианте полностью исчезли какие-либо проблемы, связанные с временной разделенностью записи лекции и возможностью ее просмотра.

Рассмотрим, насколько видеолекция может решить все три описанные выше проблемы лекционной формы занятий. Мы уже обсудили, что при современном развитии ИКТ видеолекция доступна учащемуся без ограничений во времени и пространстве, что в значительной мере решает проблему «коллизии интересов». Однако заметим, что видеолекция отличается от очной лекции тем, что взаимодействие осуществляется по видеоканалу. Это накладывает определенные ограничения. Например, и лектор, и демонстрационный ряд (доска, где производятся записи, или экран, на котором отображается презентация) должны быть размещены в одном кадре. Как правило, качественная запись требует профессиональных усилий по съемке и монтажу, что обычно находится за рамками возможностей учебного заведения и тем более отдельного преподавателя.

Что касается проблемы «потери мотивации», то она, по всей вероятности, в видеолекции никак не решается. В самом деле, если учащийся не имеет достаточной мотивации, чтобы посещать очные лекции, непонятно, что сможет мотивировать его прослушивать видеолекцию. В то же время теряется возможность лектора влиять на мотивацию, например, через ведение журнала посещений.

Для проблемы «снижения вовлеченности» видеолекция также не является решением. У видеолекции отсутствуют видимые механизмы удержания внимания по сравнению с очной лекцией. В то же время теряется возможность лектора удерживать внимание аудитории, например, через диалоги или опросы.

Подход к лекционной форме занятий на основе мультимедиалекции с педагогическим агентом

Для решения сформулированных выше проблем лекционной работы нами предлагается подход к модификации лекционной работы в вузе, позволяющей преодолеть рассмотренные выше проблемы. Подход является развитием формата видеолекции и основан на использовании новых информационных технологий, в том числе мультимедиатехнологий и технологий

искусственного интеллекта. В подходе применяются два принципиальных решения, связанные между собой.

Первым таким решением является замена видеолекции мультимедиалекцией. Одним из важнейших преимуществ видеолекции является то, что в ее основание положен видеоформат данных, что обеспечивает широкую интероперабельность. Такая лекция с учетом возможностей перекодирования между видеоформатами может быть подготовлена с помощью широкого круга аппаратных и программных средств, доставлена с помощью разнообразных носителей или Интернета и проиграна без установки дополнительного ПО на любом компьютерном устройстве, включая мобильные.

Однако в настоящее время возможности практически всех устройств, которые используются для просмотра видеоконтента, пригодны и для проигрывания мультимедиа, под которым мы традиционно понимаем использование разнородного медиаконтента (видео, аудио, графика и др.) под управлением интерактивного программного обеспечения. Универсальным клиентом как для проигрывания, так и для производства такой лекции вполне может быть стандартный веб-браузер с базовым набором возможностей, который сегодня имеется практически на любом устройстве, используемом для просмотра видео.

Переход от видео к мультимедиа дает два направления развития. Во-первых, можно существенно улучшить эффективность демонстрационного ряда лекции, применяя наиболее подходящий формат данных [5]. Так, форматы видео плохо приспособлены для отображения графики. Использование соответствующих графических форматов, особенно в сочетании с соответствующими инструментами (например, масштабирование), во многих случаях будет полезно. Причем дальнейшее разделение графики, скажем, на растровую для полутоновых изображений и векторную для различных диаграмм, чертежей и т. д., также способно усиливать эффективность демонстрации. Кроме этого, 3D-графика, демонстрируемая с помощью соответствующих вьюверов, а не через запись видеофайла, существенно качественнее.

Во-вторых, поскольку мультимедиаформат предполагает внедрение управляющего программного обеспечения, то на это ПО могут быть возложены различные сервисы, помогающие работе с мультимедиалекцией. Например, это может быть предоставление навигационных средств работы с лекцией, таких как содержание, индексы, контекстный поиск и др.

Другим ключевым решением нашего подхода является использование интеллектуальных педагогических агентов. Термин «педагогический агент» мы будем использовать как метафору персонифицированного человекоподобного интерфейса между учащимся и учебным материалом образовательной среды [6; 7]. Принципиальная возможность использования педагогических агентов также основывается на наличии в мультимедиалекции программного компонента. Мы здесь не будем касаться обширного дискурса и многочисленных исследований, связанных с педагогическим агентом. Упомянем только, что эта концепция является проекцией программного ассистента на область учебной деятельности. Педагогический агент предназначен для моделирования типа взаимодействия между учеником и другим человеком. Такого агента можно определить как «персонажа, разыгрываемого компьютером, который взаимодействует с пользователем в социально привлекательной манере». Основная задача педагогического агента в рамках нашего подхода – выполнять педагогические задачи, которые в классической лекции выполнялись лектором, а в видеолекции были потеряны.

Характеристика «интеллектуальный» применена нами к педагогическому агенту, поскольку предполагается, что в технологии, на которых этот интерфейс основан, включен и ряд возможностей, которые обычно относят к технологиям искусственного интеллекта (ИИ). Вообще, реализация антропоморфного агента (в частности, педагогического) не требует обязательного обращения к технологиям ИИ. Примером этого могут служить многочисленные реализации чат-ботов, многие из которых представляют весьма несложные логические программы, успешно при этом имитируя диалоги. Однако представляется естественным для построения педагогических агентов использовать технологии искусственного интеллекта. При этом возникает

искушение в построении педагогического агента, который бы мог заместить преподавателя, в том числе и через участие в содержательных диалогах с учащимся на базе технологий обработки и понимания естественного языка (NLP, NLU). В этом направлении имеется множество теоретических исследований и некоторое количество интересных «лабораторных» реализаций (см., например, обзор О. И. Долгой [8]). В то же время сколько-нибудь массового использования педагогических агентов, выполняющих содержательные педагогические задачи на основе понимания естественного языка в учебном процессе образовательных организаций, не наблюдается, а перспективы такого использования остаются за рамками наших исследований.

Мы ставим перед педагогическим агентом, включаемым в состав мультимедиалекции, вполне конкретную и ограниченную задачу, не предполагающую от агента знания учебного материала хоть в каком-нибудь смысле. Эта задача полностью возлагается на материалы мультимедиалекции – видео лектора, читающего лекцию и демонстрационный ряд, используемый им при ее чтении. В задачи же педагогического агента входит мониторинг процесса изучения лекции: темп изучения, внимательность контроль за усвоением учебного материала, поддержка мотивации и внимания учащегося.

Разумеется, чтобы педагогический агент был способен заместить взаимодействие лектора со слушателями, имеющее место при очном формате лекции, он должен обладать достаточно развитыми возможностями взаимодействия с обучающимися, должен быть достаточно развитым интеллектуальным агентом в том смысле, как это понимается в исследованиях в области искусственного интеллекта. Такие возможностями могут основываться на способности видеть окружающую действительность, например, с помощью веб-камеры, и распознавать получаемые с камеры изображения, в том числе идентифицировать по изображению лица учащихся, определять их внимательность, эмоциональный настрой и другие визуальные параметры. Также для организации взаимодействия, хоть сколько-нибудь приближенного к реальному, интеллектуальному агенту необходимо обладать способностью двунаправленного речевого взаимодействия с учащимися, основанной на распознавании и синтезе речи. Это требует от нашего агента владения технологиями распознавания лиц, а также анализом и синтезем речи. Необходимо отметить, что современный уровень развития компьютерной техники и информационных технологий вполне позволяет реализацию таких возможностей для учебных систем, в том числе организованных, как веб-сервисы. При этом его действия напоминают работу простого чат-бота, действующего не по анализу содержания беседы, а по некоторому логическому алгоритму, зависящему не от смысла речи, а от совпадения распознанных слов со словарем ключевых.

Покажем, что предлагаемый подход потенциально способен решить рассмотренные выше проблемы лекционной формы, снижающие ее эффективность. Проблема «коллизии интересов» решается в рамках подхода так же хорошо, как и в видеолекции. Коллизия снимается тем, что мультимедиалекция может быть просмотрена (прослушана) в любое удобное для учащегося время и вне зависимости от его территориальной удаленности. Более того, интерактивное ПО делает просмотр мультимедиалекции существенно удобнее по сравнению с видеолекцией в тех случаях, когда слушателю требуется разбить ее изучение на несколько приемов или повторить какие-либо части. В то время как видеолекция лишена внутренней структуры, в мультимедиалекции возможно организовать содержание лекции с указанием длительности каждой части. Такая информация позволит слушателю планировать свою деятельность. Кроме того, мультимедиалекция может запоминать и подсказывать слушателю, какие из частей уже были им освоены, а какие еще предстоит изучить.

Как мы показали выше, проблема «потери мотивации» не решается удовлетворительно ни в классической лекции, ни в видеолекции. Лектор не может повторно читать лекции для отсутствовавших студентов. Студенты, не имеющие мотивации к учебе, не прослушают видеолекцию без контроля. Что касается мультимедиалекции, то педагогический агент мог бы такой контроль организовать. Например, агент мог бы выдавать студенту некоторый серти-

фикат о прослушивании мультимедиалекции. При выдаче такого сертификата агент мог бы основываться на том, прослушал ли студент лекцию, всю ли, лично ли, с должной ли степенью внимательности. Заключение об этом агент мог бы составлять, прежде всего, на основе распознавания лица (лиц) на сигнале с веб-камеры.

Кроме того, эффективным средством, страхующим от того, что студент присутствует при воспроизведении мультимедиалекции перед экраном веб-камеры, но занимается посторонними делами, могут быть экспресс-вопросы, задаваемые агентом слушателю лекции. Такие вопросы могут быть рассчитаны не столько на проверку усвоения лекционного материала, сколько на контроль внимательности слушателя, например, через вопросы, требующие простого воспроизведения сказанного. Алгоритм задавания вопросов может быть различным: в конце каждой части лекции или при обнаружении агентом потери внимательности. Правильный ответ на вопрос мог бы открывать возможность дальнейшего прослушивание мультимедиалекции, а неправильный – необходимость вновь прослушать предыдущий фрагмент. После прослушивания всех таких частей лекции педагогический агент мог бы отмечать ее как прослушанную. Мы полагаем, что для преодоления проблемы «потери мотивации» достаточно в качестве допуска к аттестации по курсу требовать, чтобы студент прослушал все лекции курса.

Наконец, мы полагаем, педагогический агент может эффективно бороться и с проблемой «снижения внимания». Собственно, уже упомянутый экспресс-опрос является первым инструментом удержания внимания: студент знает, что при прослушивании лекции он в каждый момент времени может получить вопрос о прослушанном, что заставляет его быть сосредоточенным. Но возможности педагогического агента для удержания внимания существенно шире. Он может удерживать внимание студента через управление ходом лекции в сочетании с речевыми диалогами. Так, если агент диагностирует усталость, апатию студента, он может инициировать диалог, спросив о состоянии студента и подбодрив его или предложив сделать короткий перерыв. Кроме того, фиксация хода освоения мультимедиалекции, выраженная в баллах или других измеряемых единицах, также мобилизует студента, поскольку наглядно представляет его прогресс и является, в некотором смысле, элементом геймификации.

Опыт проектирования и реализации системы, выполненной в рамках подхода

Формулировка подхода, данная нами выше, указывает общие принципы и функциональные возможности системы, но не определяет сколько-нибудь конкретно структуру данных и алгоритмы работы. Однако практическое подтверждение эффективности подхода может быть осуществлено только через его реализацию, которая требует определения конкретных структур данных и алгоритмов работы программного обеспечения, В этом разделе приводим основные решения ЛЕкционной МультиМедиа Аудитории (ЛЕММА) – нашей реализации подхода мультимедиа лекций с педагогическим агентом.

В ЛЕММА нами была определена инфологическая модель мультимедиалекции, сочетающая, на наш взгляд, необходимую функциональность с относительной простотой реализации. Выбранная конструкция имеет следующие основные решения в части организации учебного материала.

1. Структурной основой мультимедиалекции (далее Лекция), помимо названия и других метаданных, является упорядоченный список содержательных частей – секций. Программное обеспечение Лекции собирает названия секций и ее метаданные в интерактивное оглавление, обеспечивающее произвольный доступ к материалам Лекции с возможностью выбора секции для просмотра на основе информации о названии секции, ее кратком описании, длительности. Секция является неделимой содержательной единицей Лекции, в частности, посекционно фиксируется прогресс в освоении Лекции: секция может быть отмечена просмотренной (зачтенной) только полностью.

2. Основу каждой секции составляют фрагмент видео лектора и демонстрационный слайд, связанный с данным фрагментом. При просмотре секции на рабочем экране проигрывается в одном фрейме видеофрагмент лектора, а во втором демонстрируется слайд. При этом слайды Лекции не являются статичными, они обладают динамикой, зависящей от типа слайда (на программном уровне слайды являются объектами различных полиморфных классов с общими базовыми свойствами абстрактного слайда) и задаваемой лектором при записи секции. При записи секции лектор управляет динамикой слайда, которая, как и видео лектора, записывается, а при просмотре Лекции воспроизводится совместно с видео лектора. Видео лектора и динамика слайда при проигрывании секции синхронизированы так, как это было задано лектором при записи Лекции.

3. Хотя экран лектора и экран слайда могут казаться двумя синхронизированными видео, для реализации динамики слайдов в Лекции используется другой механизм: на экране слайда представляется сам слайд-объект, которому подаются команды управления в том же порядке и с такими же задержками, как это произошло при манипуляциях лектора в момент записи Лекции. Работа с такими слайдами-объектами вместо видео имеет множество преимуществ, одним из которых является снижение трафика для представления слайда, что уменьшает задержки и в ряде случаев (например, для векторной графики) улучшает качество слайда.

4. Полиморфизм слайдов включает в себя ряд общих методов. Во-первых, это методы записи команд управления (манипуляции) слайдом со стороны лектора. Хотя сам список команд и соответствующих им действий индивидуален для каждого класса слайдов (например, play/pause для видео или зумм для графики), механизм записи команд и их тайминга (т. е. задержки относительно начала записи секции), а также их чтения и подачи слайду-объекту в необходимый момент времени, являются общими, т. е. наследуемыми от прототипа. Также общими являются методы приостановки и возобновления показа слайда, позволяющие приостанавливать и возобновлять показ секции учащимся при ее просмотре, и некоторые другие методы.

5. Работа со слайдами как объектами вместо видео дает возможность использовать слайд-объект как интерактивное средство обучения. В процессе просмотра секции со слайдом процесс просмотра может быть приостановлен и быть построен клон слайда-объекта. Этот клон может быть размещен в отдельном окне/фрейме экрана, снабжен элементами управления, связанными с данным типом слайда, и передан учащемуся для манипуляций. Через этот механизм демонстрационный ряд Лекции становится самостоятельным интерактивным обучающим элементом и поддерживает активную работу ученика с предоставляемым материалом. В самом простом случае ученик может копировать со слайда информацию, например, тексты, формулы, программные коды для дальнейшего использования в курсе. Кроме того, он может проводить учебные манипуляции с объектом, например, рассматривать объект на 3D-слайде, вращая и масштабируя его, что может быть актуально при анализе археологических артефактов или деталей механизмов.

6. С каждой секцией может быть связано одно или несколько заданий, которые могут быть заданы студенту после успешного просмотра секции. Задания могут быть вопросами с целью проконтролировать внимательность студента в ходе прослушивания секции, или упражнениями, являющимися элементами обучения, способом закрепления материала. Методика предъявления вопросов может быть определена в настройках Лекции и, в некоторых случаях и моментах, делегирована интеллектуальному агенту, о котором пойдет речь (если, например, агент видит, что студент часто отвлекается от прослушивания, ему задается больше вопросов).

7. Мультимедиалекция всегда прослушивается в контексте конкретного учащегося и состоит из просмотров (прослушиваний) секций и ассоциированных с ними заданий. В соответствии с заданными преподавателем параметрами просмотра Лекции учащемуся могут быть доступны различные режимы просмотра.

8. При этом результаты фиксируются и запоминаются, так что изучение Лекции учащимся может осуществляться в несколько этапов, до получения необходимого результата. В насто-

ящее время для демонстрационного прототипа реализован простейший алгоритм: просмотр секции оценивается в один балл, выполненное задание – также в один балл. Для того чтобы Лекция была зачтена как прослушанная, необходимо набрать определенное количество баллов.

Алгоритмы работы мультимедиалекции реализуются интеллектуальным педагогическим агентом, для которого в системе ЛЕММА используется метафора тьютора. Здесь этот термин трактуется так, как это принято в дистанционном обучении (в том числе его онлайн-варианте). Вообще, термин тьютор известен по крайней мере с XIII века, когда в Кембридже и Оксфорде зафиксированы упоминания об особой педагогической роли, в которой преподавательская деятельность сочетается с выполнением задач наставничества и опекунства. Поскольку при дистанционном обучении собственно преподавательские задачи передаются электронным средствам обучения – видеолекциям, системам педагогического тестирования и т. д., в дистанционном обучении за тьютором остаются задачи мониторинга процесса обучения, мотивирование подопечных, контроль освоения учебного материала и коррекция индивидуальной образовательной траектории. В ЛЕММА педагогический агент (далее Тьютор) реализует следующие основные алгоритмы.

1. Тьютор осуществляет биометрическую аутентификацию пользователя, используя технологию распознавания лиц. При начале работы с Лекцией, а также с каждой секцией Лекции осуществляется стартовая аутентификация, в процессе просмотра секции или выполнения задания – периодическая аутентификация пользователя. Аутентификация считается пройденной, если в изображении с веб-камеры распознано только одно лицо, которое соответствует лицу на фотографии, сохраненной в профиле учащегося.

2. В соответствии с результатами аутентификации Тьютор осуществляет допуск к материалам Лекции и ведет учет пройденного материала. Так, допуск к прослушиванию каждой секции требует успешной стартовой аутентификации. Секция засчитывается, только если учащийся прослушал секцию, а затем выполнил задание секции. При этом для того чтобы получить допуск к заданию, необходимо, чтобы аутентификация во время прослушивания была успешной не менее чем в 90 % случаев. В свою очередь, задание считается выполненным, если учащийся уложился в отведенное для задания время, ответ верный, а периодическая аутентификация за время выполнения задания была успешной в 90 % случаев. В том случае, если задание не выполнено, секция должна быть прослушана повторно.

3. Тьютор ведет учет прослушанных секций Лекции, запоминает результаты сеанса пользователя, загружает результаты в следующем сеансе того же пользователя, помечая прослушанные секции в содержании лекции. При прослушивании всех секций выставляет для данного учащегося в задании на прослушивание лекции статус «выполнено».

4. Тьютор поддерживает с учащимся постоянную речевую коммуникацию посредством голосового и/или текстового чата (по выбору учащегося). Тьютор информирует учащегося о стадиях прослушивания Лекции, зачете/незачете прослушивания секции или выполнения задания, причинах незачета, выдает другую существенную для учебного процесса информацию. Тьютор способен принимать голосовые команды управления интерфейсом и выполнять их.

5. Тьютор также способен применять возможность речевого анализа для проверки верности выполнения задания, ответ на который должен быть дан голосом (например, повторить строки стихотворения). В этом случае правильность ответа определяется по норме близости двух строк – строки ответа учащегося и эталонной.

Разумеется, так определенные модели мультимедиалекции и педагогического агента могут быть существенно расширены, однако, как мы полагаем, они вполне могут продемонстрировать эффективность развиваемого подхода.

В настоящее время автором описываемого проекта осуществляется разработка системы мультимедиалектория, выполняемая в рамках рассматриваемого подхода и в соответствии с представленной в предыдущем разделе моделью ЛЕММА.

В настоящее время нами получен конкретный результат: построен демонстрационный прототип ключевой компоненты программного обеспечения системы – проигрыватель (плеер) мультимедиалекции. Важность этой компоненты заключается в том, что она реализует основную функцию системы, а также что именно для этой компоненты действуют наиболее жесткие условия функционирования: для конечных пользователей системы – учащихся – необходимо обеспечить минимально возможные требования к аппаратной части, и в то же время именно она будет требовать наибольшее количество как вычислительных, так и сетевых ресурсов. Действительно, компьютерное устройство пользователя должно обеспечивать без задержек работу с мультимедиаконтентом в сочетании с работой модулей распознавания лиц, анализа и синтеза речи, а также передачу всех необходимых данных по сети. При этом мы должны ориентироваться на такие вычислительные и сетевые ресурсы клиентской стороны, которые де-факто имеются у подавляющего числа потенциальных пользователей.

Для того чтобы с помощью данного плеера можно было демонстрировать лекции, нам пришлось разработать и ряд скриптов, позволяющих такие лекции создавать, однако описание их реализации остается за рамками данной работы.

На современном уровне развития информационных технологий представляется наиболее оправданным создавать системы электронного обучения в виде интернет-сервисов. Большинство пользователей таких систем имеют доступ к каналам с достаточной коннективностью, а стандартные веб-браузеры сегодня обладают широкой функциональностью, включая работу с мультимедиаконтентом, современные императивные и декларативные средства программирования.

На первом этапе ПО строится в модели ПО как сервис (SaaS, Software as a Service) и реализуется как насыщенное интернет-приложение (RIA, Rich Internet Application). В качестве принципиальной архитектуры системы используется классическая трехзвенная клиент-серверная модель с SQL-сервером в качестве сервера баз данных, веб-сервером и комплектом PHP-скриптов в качестве сервера приложений и веб-браузером с DHTML в качестве универсального клиента.

Серверная часть ПО системы выполнена на инструментальном стеке LAMP. Код клиента – чистый HTML + JavaScript + CSS. Никаких PHP и JS-фреймворков, сторонних библиотек, за исключением библиотеки распознавания лиц[1], не использовалось. Клиентская часть предполагает использование браузера Google Chrome, который поддерживает все необходимые возможности Java Script (в том числе интерфейсы File API и Web Speech API). Поскольку используются только стандартные возможности[2], ПО должно работать и на других браузерах, в которых эти стандарты поддерживаются в необходимом объеме, однако тестирование для них не проводилось.

Проигрыватель вызывается из личного кабинета учащегося, а его входными параметрами являются логин слушателя и идентификатор выбранной для прослушивания Лекции. Демонстрационный прототип проигрывателя позволяет авторизованному пользователю просматривать мультимедиалекции и фиксирует полученные в ходе обучения результаты. В состав проигрывателя включена реализация тьютора, демонстрирующего работу с технологиями распознавания лиц, анализа и синтеза речи и основной функционал педагогического агента модели ЛЕММА: биометрическую авторизацию, голосовой и текстовый чат с учащимся.

Интерфейс программы состоит из главного окна и нескольких вспомогательных окон. Главное окно (рис. 1) состоит из трех областей: заголовка, рабочей области, области инструментов. В заголовке отображается название Лекции, имя и фамилия учащегося, авторизованного для работы с ней. Справа расположены три информационные области. В первой отображается результат текущей биометрической аутентификации, осуществляемой с периодом 1 секун-

[1] https://justadudewhohacks.github.io/face-api.js/docs

[2] Resources for Developers, by Developers. URL: https://developer.mozilla.org

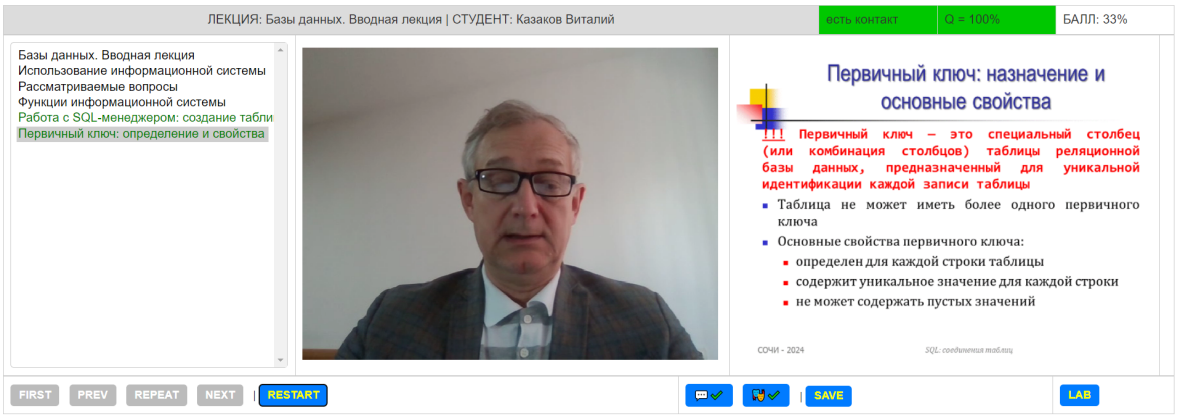


Рис. 1. Главное окно проигрывателя мультимедиалекций
Fig. 1. Main window of the player of multimedia lectures

да. Если аутентификация пройдена, то фон области имеет зеленый цвет, если нет – красный. Цветовое решение позволяет в процессе просмотра Лекции увидеть проблему (может быть вызвана жестом, закрывающим лицо, плохим освещением, выходом из кадра и т. д.) и скорректировать работу. Вторая область работает в режиме просмотра секции или выполнения задания и отображает интегральный процент успешных аутентификаций за период действия режима. Помимо отображения числового значения также используется цветовая градиентная индикация, где 0 % отражается красным, а 100 % – зеленым цветом области. Наконец, в третьей области отображается количество набранных слушателем баллов, приведенное к 100 %, максимально возможному в процентах.

В средней, рабочей области расположены три фрейма. В первом отображается содержание Лекции в виде списка секций, причем, просмотренные, но не засчитанные секции отображаются красным цветом текста, а засчитанные – зеленым. Выбор секции кликом мышки приводит к началу ее просмотра. Во втором фрейме отображается видео лектора, в третьем – демонстрируется слайд.

В подвале окна располагаются различные кнопки, активирующие инструменты работы с Лекцией. Первая группа – FIRST, PREV (previous), REPEATE, NEXT – позволяет выбрать секцию для прослушивания. Эти кнопки (как и содержание Лекции) не активны во время работы с секцией Лекции. Во время работы с секцией Лекции доступны кнопки PAUSE / RESTART, вызывающие приостановку Лекции, а также кнопку LAB, создающую клон слайд-объекта с инструментами управления и передающую его учащемуся так, как описано выше. Наконец, имеется кнопка SAVE, позволяющая сохранить промежуточные результаты работы. Также присутствуют две кнопки с иконками разговора и театральных масок, вызывающие вспомогательные окна «Чат» и «Аутентификация».

На рис. 2. представлены два вспомогательных окна. Окно слева внизу отображает сигнал с веб-камеры, используемый для биометрической аутентификации. В окне рамками отображаются области, распознанные как человеческие лица. Если лицо определено, как соответствующее профилю учащегося, авторизованного для просмотра Лекции, рамка помечается подписью с текстом логина учащегося. В противном случае – подписью undefined.

Окно справа отображает чат Тьютора с учащимся. В настоящее время чат используется Тьютором для передачи учащемуся информации, связанной с просмотром лекционного материала, например, с каким результатом учащийся выполнил задание. Информация сообщается Тьютором голосом и дублируется в чате. Учащийся в текущей реализации может управлять просмотром Лекции, отдавая команды голосом. Команды распознаются и также дублируются в чате в текстовом виде.

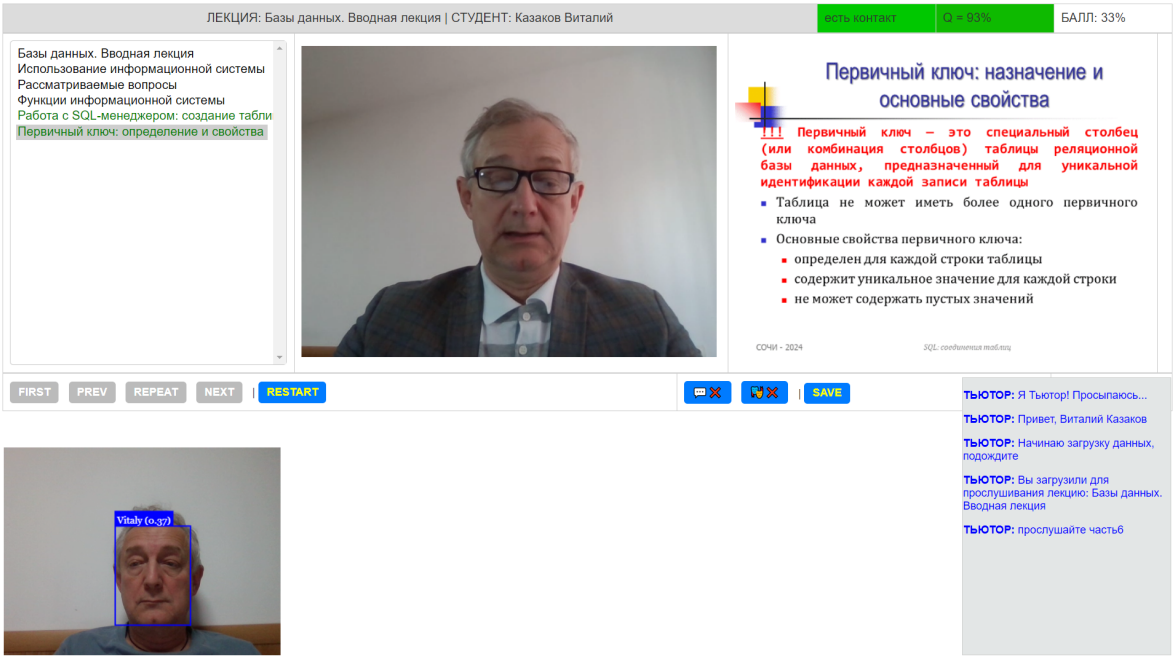


Рис. 2. Вспомогательные окна чата Тьютор – учащийся (справа) и биометрической авторизации (слева)
Fig. 2. Auxiliary windows: chat Tutor – student (right) and biometric authorization (left)

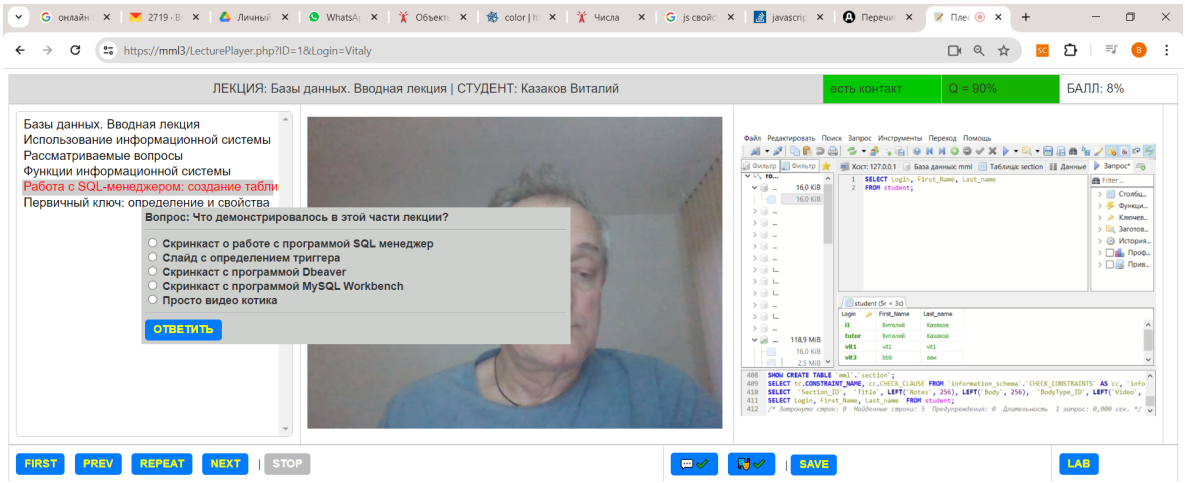


Рис. 3. Окно выполнения задания
Fig. 3. Job execution window

На рис. 3. представлено еще одно окно – окно задания. Оно предназначено для отображения задания, связанного с прослушанной секцией Лекции и его выполнения учащимся. Окно вызывается Тьютором после того, как он зафиксировал, что секция прослушана учащимся с соблюдением правил. Окно завершает работу после выполнения задания по нажатию студентом кнопки ОТВЕТИТЬ либо по истечению времени, отведенного на выполнение (10 секунд по умолчанию).

Наиболее важные результаты, полученные в ходе тестирования демонстрационного прототипа.

1. Продемонстрирована принципиальная возможность перманентной биометрической авторизации на основе технологии распознавания лиц по изображению с веб-камеры с работой алгоритма на стороне клиента без передачи видеопотока на серверную сторону.

2. Показана возможность использования стандартной JavaScript библиотеки Web Speech API для анализа и синтеза речи в организации речевой коммуникации педагогического агента с учащимся.

3. Определено, что при просмотре мультимедиалекции на компьютерных устройствах бытового назначения (персональных компьютерах, ноутбуках, планшетах, смартфонах) работа во всех режимах, включая одновременный просмотр видео в высоком разрешении, биометрическую авторизацию, работу модулей распознавания и синтеза речи, протекает без явных задержек.

Таким образом, показана возможность создания программной системы в формате RIA, обеспечивающей на компьютерных устройствах бытового уровня просмотр мультимедиалекций с интеллектуальным педагогическим агентом, выполненных в соответствии с описанным подходом.

Заключение

Таким образом, в работе получены следующие результаты.

1. Проведен анализ проблем лекционной формы работы в вузе, порожденных социальными трансформациями последних десятилетий. Показано, что популярный формат видеолекций снимает проблемы классической лекции, связанные с территориальной и временной удаленностью, но не обеспечивает необходимый контроль лекционного процесса, поддержку мотивации и внимания учащихся.

2. Предложен новый подход, основанный на применении в дистанционной лекции мультимедиатехнологий в сочетании с интеллектуальным педагогическим агентом.

3. Показано, что в рамках предложенного подхода могут быть эффективно преодолены проблемы лекционной формы работы, в том числе по контролю и поддержке мотивации учащихся.

4. Для практической проверки эффективности мультимедиалекций с педагогическим агентом на практике построена инфологическая модель системы ЛЕММА – создания и использования мультимедиалекций.

5. Начата практическая реализация системы, определены архитектура системы и стек технологий реализации. Построен демонстрационный прототип ключевой части системы – проигрывателя (плеера) мультимедиалекций, демонстрирующий возможность реализации подхода и его основные преимущества.

На основании проведенных работ и полученных результатов можно констатировать, что нами сформулирован новый подход к организации лекционной работы в вузе, основанный на использовании мультимедиатехнологий в сочетании с технологиями искусственного интеллекта, а также предложена и обоснована гипотеза, что применение данного подхода в вузе будет способствовать эффективности лекционной формы работы.

В дальнейшем планируется построение действующего полнофункционального прототипа системы, реализующей все средства, необходимые для производства и применения мультимедиалекций. Действующий прототип предполагается применить в режиме апробации на отдельных занятиях и курсах в учебном процессе вузов и других образовательных задачах, например, курсах повышения квалификации, а также оценить эффективность такого применения и сформулировать методические рекомендации по использованию мультимедиалекций в образовательном процессе высшей школы.

Список литературы

1. **Ибрагимов Г. И., Гайнутдинов Р. Г.** Лекция в вузе: теория, история, практика: Монография. Казань: Школа, 2017. 196 с.
2. **Маклюэн М.** Галактика Гутенберга. Становление человека печатающего. М.: Академический проект, 2005. 496 с.
3. **Carr N.** Is Google making us stupid? – The Atlantic. July/August 2008. URL: https://web.lib.unb.ca/instruction/bcull/ARTICLES/Reading/GoggleCBCA.pdf (дата обращения: 06.05.2024).
4. **Kaplan A. M., Haenlein M.** Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster // Business Horizons. 2016. Vol. 59 (4). P. 441–450. doi:10.1016/j.bushor.2016.03.008
5. **Носков И. В., Казаков В. Г., Казаков В. В., Щеглов Ю. А.** Веб-студия для создания и применения учебных мультимедиа лекций // Технологии информационного общества в науке, образовании и культуре: Тр. XVII Всерос. объединенной конф. «Интернет и современное общество» (IMS-2014). Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики; Библиотека Российской Академии наук. СПб., 2014. С. 343–346.
6. **Chin D. B., Dohmen I. M., Cheng B. H., Oppezzo M. A., Chase C. C., Schwartz D. L.** Preparing students for future learning with Teachable Agents // Education Tech Research Dev. 2010. Vol. 58 (6). P. 649–669. doi:10.1007/s11423-010-9154-5
7. **Mabanza N., de Wet L.** Determining the Usability Effect of Pedagogical Interface Agents on Adult Computer Literacy Training. E-Learning Paradigms and Applications // Studies in Computational Intelligence. 2014. Vol. 528. P. 145–183. doi:10.1007/978-3-642-41965-2
8. **Долгая О. И.** Искусственный интеллект и обучение в школе: ответ на современные вызовы // Школьные технологии. 2020. Вып. 4. С. 29–38.

References

1. **Ibragimov G. I., Gainutdinov R. G.** University lecture: theory, history, practice. Kazan', Shkola publ., 2017. 196 p. (In Russ.)
2. **McLuhan M.** The Gutenberg Galaxy : the making of typographic man. Toronto, Canada: University of Toronto Pressб 1962, p. 293.
3. **Carr N.** Is Google making us stupid? – The Atlantic. July/August 2008.
4. **Kaplan A. M., Haenlein M.** Higher education and the digital revolution: About MOOCs, SPOCs, social media, and the Cookie Monster. *Business Horizons*, 2016, vol. 59 (4), pp. 441–450. DOI:10.1016/j.bushor.2016.03.008
5. **Noskov I. V., Kazakov V. G., Kazakov V. V., Shcheglov Yu. A.** Veb studiya dlya sozdaniya i primeneniya uchebnykh mul'timedia lektsii. In: *Tekhnologii informatsionnogo obshchestva v nauke, obrazovanii i kul'ture. sbornik nauchnykh statei. Trudy XVII Vserossiiskoi ob"edinennoi konferentsii «Internet i sovremennoe obshchestvo»* (IMS-2014). Saint Petersburg,. 2014, pp. 343–346. (In Russ.)
6. **Chin D. B., Dohmen I. M., Cheng B. H., Oppezzo M. A., Chase C. C., Schwartz D. L.** Preparing students for future learning with Teachable Agents. *Education Tech Research Dev*., 2010, vol. 58 (6), pp. 649–669. DOI:10.1007/s11423-010-9154-5
7. **Mabanza N., de Wet L.** Determining the Usability Effect of Pedagogical Interface Agents on Adult Computer Literacy Training. *E-Learning Paradigms and Applications. Studies in Computational Intelligence*, 2014, vol. 528, pp. 145–183. DOI:10.1007/978-3-642-41965-2
8. **Dolgaya O. I.** Iskusstvennyi intellekt i obuchenie v shkole: otvet na sovremennye vyzovy. *Shkol'nye tekhnologii*, 2020, iss. 4, pp. 29−38. (in Russ.)

Сведения об авторе

Казаков Виталий Геннадьевич, кандидат физико-математических наук, доцент кафедры информационных технологий и математики Сочинского государственного университета, ведущий научный сотрудник Исследовательского центра в сфере искусственного интеллекта Новосибирского государственного университета
Scopus ID: 8378777200
Researcher ID: T-6050-2017

Information about the Author

Vitaly G. Kazakov, Ph.D., Associate Professor, Department of Information Technologies and Mathematics, Sochi State University (Sochi), Leading Researcher at the Research Center in the Field of Artificial Intelligence, Novosibirsk National Research State University (Novosibirsk)
Scopus ID: 8378777200
Researcher ID: T-6050-2017

Статья поступила в редакцию 15.08.2024;
одобрена после рецензирования 04.12.2024; принята к публикации 04.12.2024

The article was submitted 15.08.2024;
approved after reviewing 04.12.2024; accepted for publication 04.12.2024

Научная статья

УДК 004.032.26:378.146:159.9
DOI 10.25205/1818-7900-2024-22-4-33-48

Поддержка принятия решений в учебном процессе вуза на основе когнитивной модели обучения с использованием нейронной сети

Андрей Петрович Клишин [1], Екатерина Сергеевна Шталина[2]
Фаррух Джамшедович Пираков [3], Людмила Владимировна Ахметова [4]
Наталия Леонидовна Ерёмина [5]

[1,2,4]Томский государственный педагогический университет
Томск, Россия

[3]Томский государственный университет систем и радиоэлектроники
Томск, Россия

[5]Томский государственный университет
Томск, Россия

[1]klishin@tspu.edu.ru
[2]shtalina@tspu.edu.ru
[3]farrukh.9559@gmail.com, https://orcid.org/0000-0003-4105-3179
[4]axme-lv@yandex.ru, https://orcid.org/0000-0002-2079-7710
[5]26051971@mail.ru, https://orcid.org/0000-0001-9508-3256

Аннотация

В работе рассматривается применение технологии нейронных сетей для поддержки принятия решений в учебном процессе вуза с использованием когнитивной модели обучения. Разработано программное решение, цифровой профиль обучающегося на базе электронного портфолио студентов с привлечением алгоритмов искусственного интеллекта, современных веб-технологий, а также когнитивных моделей обучения. Обучение нейронной сети проводилось на подготовленных данных студентов, которые были получены с использованием специально разработанного психодиагностического комплекса. Использование цифрового профиля позволяет студентам отслеживать свой процесс обучения на основе рекомендаций, предлагаемых нейронной сетью, принимать оптимальные решения, строить персонализированные образовательные траектории, а также корректировать образовательные траектории обучения.

Ключевые слова

цифровой профиль, нейронная сеть, когнитивный подход, электронное портфолио, поддержка принятия решений, Keras

Для цитирования

Клишин А. П., Шталина Е. С., Пираков Ф. Дж., Ахметова Л. В., Ерёмина Н. Л. Поддержка принятия решений в учебном процессе вуза на основе когнитивной модели обучения с использованием нейронной сети // Вестник НГУ. Серия: Информационные технологии. 2024. Т. 22, № 4. С. 33–48. DOI 10.25205/1818-7900-2024-22-4-33-48

Decision Support in the Educational Process of the University based on a Cognitive Learning Model using a Neural Network

Andrey P. Klishin[1], Ekaterina S. Shtalina[2], Farrukh D. Pirakov[3], Lyudmila V. Akhmetova[4], Natalia L. Eryomina[5]

[1,2,4]Tomsk State Pedagogical University,
Tomsk, Russian Federation

[3]Tomsk State University of Systems and Radioelectronics,
Tomsk, Russian Federation

[5]Tomsk State University,
Tomsk, Russian Federation

[1]klishin@tspu.edu.ru
[2]shtalina@tspu.edu.ru
[3]farrukh.9559@gmail.com, https://orcid.org/0000-0003-4105-3179
[4]axme-lv@yandex.ru, https://orcid.org/0000-0002-2079-7710
[5]26051971@mail.ru, https://orcid.org/0000-0001-9508-3256

Annotation

The paper discusses the use of neural network technology to support decision-making in the educational process of a university using a cognitive learning model. A software solution has been developed for a digital profile of a student based on an electronic portfolio of students, using artificial intelligence algorithms, modern web technologies, as well as cognitive learning models. The neural network was trained on prepared student data, which was obtained using a specially developed psychodiagnostic complex. Using a digital profile allows students to track their learning process based on recommendations offered by a neural network, make optimal decisions, build personalized educational trajectories, and also adjust educational learning trajectories.

Keywords

digital profile, neural network, cognitive approach, electronic portfolio, decision support, Keras

For citation

Klishin A. P., Shtalina E. S., Pirakov F. Dzh., Akhmetova L. V., Eryomina N. L. Decision support in the educational process of the University based on a cognitive learning model using a neural network. *Vestnik NSU. Series: Information Technologies*, 2024, vol. 22, no. 4, pp. 33–48 (in Russ.) DOI 10.25205/1818-7900-2024-22-4-33-48

Введение

В настоящее время в сфере высшего образования важную роль играют методы поддержки принятия управленческих решений, которые на различных уровнях управления эффективно используют интегрированные инструменты автоматизации и управления. Цифровая среда становится все более сложной и динамичной, в связи с чем возникает потребность в быстрой ориентации и реагировании на различные ситуации, требующие разработки инновационных решений в сфере управления [1]. Разработка и внедрение новых методов поддержки принятия управленческих решений позволяет адаптивно управлять обучением за счет использования различных комплексных программных систем. В связи с этим одним из способов организации эффективной подсистемы управления образовательным процессом в вузе является когнитивный подход. Для таких случаев в процессах управления возникает необходимость принятия решений в слабоструктурированных динамических ситуациях, в случае если закономерности развития ситуации частично описываются качественными значениями [2].

В быстро меняющейся цифровой среде представляется целесообразным использовать адаптивное управление образовательным процессом с применением когнитивного подхода, что ведет к разработке новых интеллектуальных систем для поддержки управленческих решений и адаптации их к самому широкому спектру возможных условий. Наиболее перспективным направлением здесь является использование искусственных нейронных сетей [3;

4]. В Российской Федерации уделяется большое внимание развитию систем искусственного интеллекта, так, в 2019 г. утверждена Национальная стратегия развития искусственного интеллекта на период до 2030 г., в рамках которой начал реализоваться федеральный проект «Искусственный интеллект», где предусмотрено в 2021–2024 гг. бюджетное финансирование в размере 27,4 млрд руб., а также дополнительное привлечение ресурсов из внебюджетных источников – 4,1 млрд руб. [5].

Применение технологии нейронных сетей позволит значительно оптимизировать образовательный процесс, а также процесс принятия управленческих решений, благодаря тому, что такая технология способна анализировать большие объемы данных, распознавать сложные информационные структуры, скрытые закономерности, образы и обрабатывать информацию быстрее, чем традиционные методы. Использование учебной аналитики (Learning Analytics, LA) одновременно с интеллектуальным анализом и обработкой образовательных данных (Educational Data Mining, EDM) открывает перспективы разработки новых системных моделей, характеризующих свойства, поведение обучающихся, а также их динамические параметры [6–9]. В образовательных системах нейронные сети можно использовать для задач прогнозирования, классификаций, а также для построения рекомендательных систем [10, 11]. Для задач классификации (в данном случае формирования рекомендаций) было решено использовать алгоритмы глубокого обучения, которые основаны на применении нейронных сетей.

В связи с этим цель данной статьи заключается в разработке элементов поддержки принятия решений в учебном процессе вуза на основе когнитивной модели обучения с использованием нейронной сети и соответствующего программного обеспечения.

1. Система электронного портфолио и цифровой профиль обучающегося

В настоящее время в Томском государственном педагогическом университете (ТГПУ) разработана и активно используется система электронного портфолио (е-портфолио) обучающегося, выполняющая функцию хранения достижений студентов, академических результатов, а также являющаяся одним из элементов единой электронной образовательной среды вуза [12]. Студенты университета (в количестве $n = 6530$) ежедневно взаимодействуют с электронным портфолио непосредственно либо посредством использования единой электронной образовательной среды, а также благодаря использованию системы электронного онлайн-обучения.

Дальнейшее совершенствование системы е-портфолио проводится в направлении персонализации электронного обучения и создания условий для максимально полного удовлетворения информационных потребностей пользователей на основе анализа загруженных материалов (достижений студентов) LA/EDM, создания и использования открытых моделей обучающихся (OLM) [11; 12]. Для решения этих задач было разработано веб-приложение «Цифровой профиль обучающегося» (ЦПО), набор образовательных веб-сервисов, а также подсистема проведения электронных конкурсов для получения стипендий различных уровней на основе загруженных материалов.

Цифровой профиль обучающегося – программная система для визуализации набора параметров обучающегося/выпускника (когнитивные параметры, учебные и общественные достижения, успеваемость и др., LA), а также веб-отображение образовательной траектории (OLM), сформированной в процессе обучения в вузе [12; 13]. ЦПО разрабатывался с учетом образовательной модели обучающегося и представляет собой составной программный компонент, составленный из различных подсистем управления учебным процессом (рис. 1).

Образовательная модель обучающегося/выпускника вуза рассматривается как совокупность усвоенных знаний и приобретенных общих и профессиональных компетенций и относится к классу OLM. Параметрами модели являются: требования, предъявляемые к выпускнику в соответствии с необходимыми стандартами; требования работодателей, формируемые исхо-

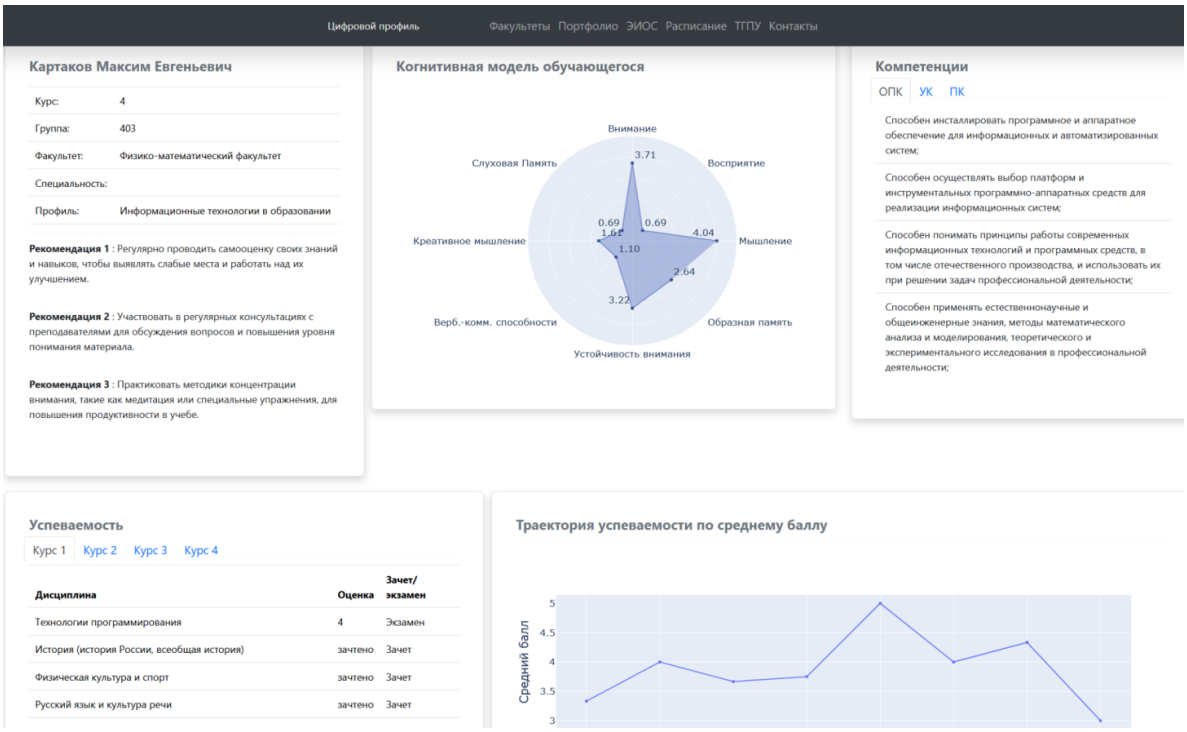


Рис. 1. Фрагмент главной формы цифрового профиля обучающегося
Fig. 1. Fragment of the main form of a student's digital profile

дя из текущей социально-экономической ситуации как в регионе, так и стране; набор личностных характеристик выпускника, способствующих эффективному взаимодействию с коллегами и в обществе в целом; набор характеристик, раскрывающих уровень знаний в профессиональной сфере [12; 13]. Модель также содержит набор учебных компетенций, список дисциплин, успеваемость (результаты зачетов и экзаменов), достижения в образовательной сфере (дипломы, грамоты, стипендии различного уровня и т. д.), а также набор личностных характеристик.

При разработке ЦПО использовался фреймворк Django, реализующий паттерн Model-View-Template (MVT), который позволяет гибко управлять веб-приложениями и работать с базами данных. В качестве СУБД для работы с данными ЦПО использовалась MySql. Для реализации построения когнитивных моделей, графиков и траекторий в работе использовалась библиотека для визуализации данных plotly.

2. Когнитивная модель обучения

В связи с высоким уровнем развития информационных технологий возникли новые перспективные возможности в реализации педагогических концепций обучения, ориентированные на индивидуально-психологические особенности обучающихся. Вместе с этим возникла потребность в теоретическом обосновании разработки и применения цифрового инструментария.

Системный подход к пониманию психологии личности обучающегося позволяет рассматривать обучение и развитие как единый, взаимосвязанный процесс, в основе которого структурно-функциональная организация психического аппарата имеет основополагающее значение [14]. Исследования психического аппарата человека на основе теоретико-методологических принципов концепции структурно-функционального подхода показали, что пси-

хический аппарат представляет собой сложную интегральную структурно-функциональную систему когнитивных функций, обеспечивающих познавательную деятельность личности на информационном и операциональном уровнях. Эти уровни имеют разные подструктуры, на которых решаются специфические когнитивные задачи [15]. Когнитивная деятельность личности на символико-концептуальном уровне осуществляется посредством механизмов знаковой коммуникации различных уровней, которые обеспечивают построение разных моделей мира. Применение технологии нейронных сетей с этой точки зрения раскрывает широкие перспективы для поддержки принятия решений и выполнения коррекционных задач в учебно-образовательном процессе.

Таким образом, когнитивная модель, используемая в ЦПО, разработана с применением психолого-дидактического подхода к обучению. Согласно положениям психолого-дидактического подхода к обучению (Д. С. Брунер), использование психологических параметров деятельности в процессе моделирования и проектирования образовательной среды является приоритетным. Деятельность обучающегося в указанных условиях характеризуется целенаправленностью, рефлексивностью, регуляцией своего поведения (В. Д. Шадриков, К. Х. Прибрам, А. Бандура, Дж. А. Келли, В. А. Лефевр).

На рис. 2 приведена схема структуры когнитивной модели ЦПО, которая отображает взаимосвязь основных ее компонентов. Особую роль в структуре когнитивной модели обучающегося имеют индивидуально-психологические и социально-психологические характеристики, представленные двумя блоками.

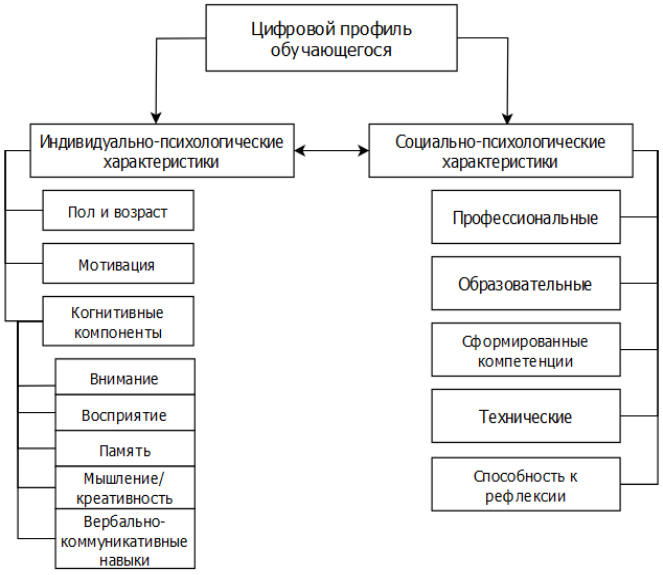


Рис. 2. Когнитивная модель, используемая в цифровом профиле обучающегося
Fig. 2. Cognitive model used in the student's digital profile

Когнитивная модель обучающегося для цифрового профиля включает базовые индивидуально-психологические и социально-психологические интегральные характеристики личности. Основным содержанием индивидуально-психологических характеристик являются половозрастные характеристики обучающихся, индивидуально-психологические свойства личности, особенности мотивации профессиональной и учебной деятельности, когнитивные компоненты (параметры когнитивного развития), которые определяются системой когнитивных признаков (восприятие, внимание, память, творческие способности, система мыслительных способностей).

Структурными компонентами социально-психологических характеристик являются учебно-профессиональные, в том числе: сформированность профессиональных компетенций, способность к рефлексии, учебные и технические достижения, включающие возможность и способность использовать информационно-технологические и иные инструментальные образовательные средства.

Представленные в когнитивной модели студента компоненты взаимосвязаны между собой, образуют динамическую систему характеристик, формирующихся в процессе обучения в образовательной среде высшего учебного заведения.

С целью определения основных параметров когнитивной деятельности обучающихся был разработан и реализован электронный психодиагностический комплекс M-CSP-test (табл. 1). Время, затрачиваемое обучающимся на тестирование, составляло в среднем 45 минут. Данные, полученные в ходе тестирования, направлялись для дальнейшего анализа в базу данных и использовались в соответствии с поставленными задачами на определенных этапах исследования.

Таблица 1

Психодиагностический комплекс для формирования когнитивной модели обучения

Table 1

Psychodiagnostic complex for the formation of a cognitive model of learning

№	Когнитивные признаки	Параметр	Шкала измерений, балл	Время (t), мин
1	Внимание	V	1–43	4
2	Мышление	IQ	1–60	25
3	Восприятие	T	1–9	2
4	Память	A	1–9	2
5	Креативное мышление	B	1–20	2
6	Вербально-коммуникативные способности	G	1–20	4
7	Устойчивость внимания	U	1–36	4
8	Образная память	M	1–16	2

3. Когнитивный подход к управлению учебным процессом в условиях цифровой трансформации образования

Когнитивный подход к управлению слабоструктурированных систем направлен на разработку формальных моделей и методов, поддерживающих интеллектуальный процесс решения проблем благодаря учету в них когнитивных возможностей (восприятие, представление, познание, понимание, объяснение) субъектов управления при решении управленческих задач [10; 13]. Подход был опробован на различных социально-экономических, экологических, образовательных и других сложных системах [2].

В основе современных изменений образовательных подходов лежит необходимость развития различных видов интеллектуальных способностей студентов с опорой на достижения в области информационных технологий и хорошо развитый интеллект с цифровыми компетенциями [11]. Когнитивные технологии направлены на активизацию студентов в образовательном процессе, а также стимулируют увеличение результативности процесса обучения, так как преподаватель при таком подходе более ориентирован на обучаемого, а не на группу учащихся [12]. При использовании механизмов анализа когнитивных процессов становится воз-

можным формировать персонализированные траектории и стратегии обучения, что оптимизирует применение ресурсов и ведет к повышению эффективности образовательного процесса.

Схема взаимодействия когнитивной подсистемы управления с образовательным процессом в вузе представлена на рис. 3. В когнитивной подсистеме (2) формируется каждые полгода набор когнитивных данных (параметров) обучающихся на основе экспресс-тестирования с использованием психодиагностического комплекса M-CSP-test, которые далее передаются в блок для построения когнитивных моделей обучения (4). На основе сформированных моделей вычисляются несколько вариантов образовательных траекторий (5) с привлечением данных из ЦПО (1) и е-портфолио. Метакогнитивная регуляция (3), в свою очередь, позволяет учащимся отслеживать и осознавать свой прогресс в обучении, а также выбирать наиболее эффективные стратегии для преодоления трудностей, что поможет им успешно выполнять когнитивные задачи [7].

Рис. 3. Схема взаимодействия когнитивной подсистемы управления с образовательным процессом в вузе
Fig. 3. Scheme of interaction of the cognitive management subsystem with the educational process at the University

Построенная система управления позволяет обучающимся проводить анализ достижений (академической успеваемости), выбирать и формировать траектории обучения в вузе, участвовать в конкурсах по научной и общественной деятельности и в целом управлять своим образовательным процессом (6), а также взаимодействовать с будущими работодателями на основе рейтинговых позиций в рамках образовательных и когнитивных моделей.

4. Исследование свойств нейронной сети для поддержки принятия решений

В качестве архитектуры нейронной сети была выбрана полносвязная нейронная сеть (англ. FCNN), которая состоит из нескольких слоев нейронов, где каждый нейрон представляет собой узел, соединенный с каждым нейроном из следующего и предыдущего слоев. Для реализации нейронной сети был выбран язык Python и библиотека глубокого обучения Keras. При работе с сетью также использовались библиотеки: для построения графиков matplotlib, pandas и numpy для работы с массивами данных.

Для обучения нейронной сети использовались данные, которые были получены в результате целенаправленного тестирования студентов на двух факультетах ТГПУ, в количестве $n = 256$. Полученные данные были предварительно нормализованы с помощью метода логарифмической нормализации. На основании полученных и обработанных в результате тестирования данных был сформирован датасет 9×512, в котором 8 столбцов отведено для значений когнитивных параметров студента, а 9-й столбец содержит соответствующую к ним рекомендацию.

В ходе исследования свойств нейронной сети было проведено сравнение оптимизаторов, варьирование количества нейронов в скрытых слоях, а также влияние параметра скорости обучения (lr). В работе рассматривались следующие оптимизаторы: Adam, RMSprop, Adamax, SGD, поскольку в задачах классификации они демонстрируют лучшие результаты [10]. На рис. 4 представлены расчетные значения параметра точности для вышеуказанных оптимизаторов.

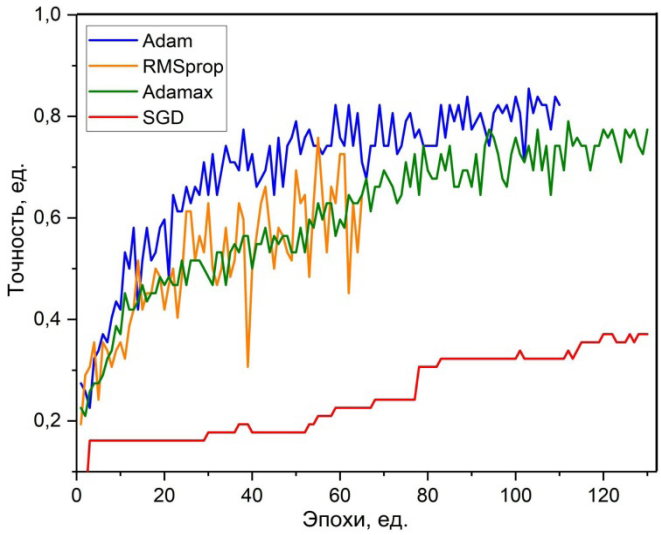
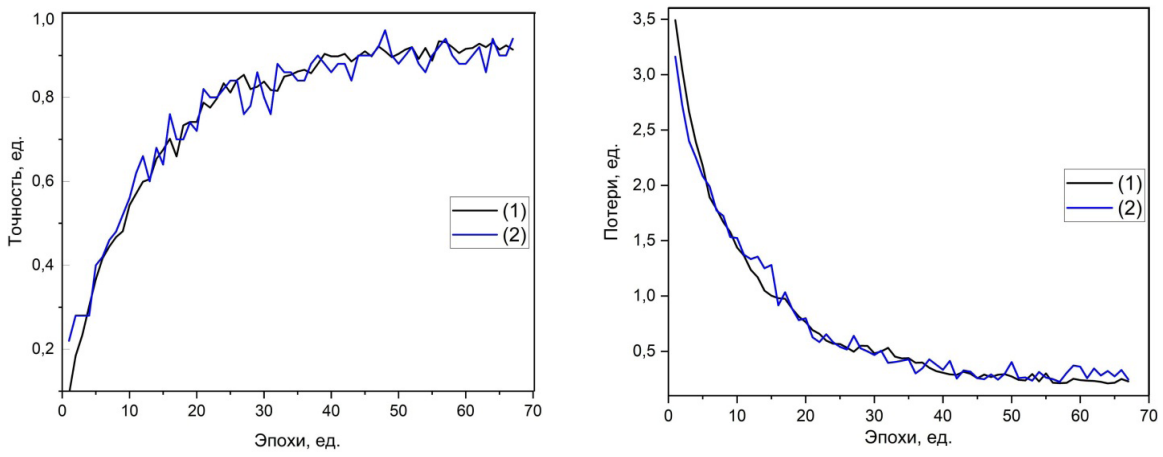


Рис. 4. Расчетные значения параметра точности при сравнении четырех оптимизаторов: Adam, RMSprop, Adamax, SGD
Fig. 4. Calculated values of the accuracy parameter when comparing four optimizers: Adam, RMSprop, Adamax, SGD

Из представленных на рис. 4 данных можно видеть, что оптимизатор Adam за наименьшее количество временных циклов (эпох) приводит к максимальной точности на построенном датасете по сравнению с другими оптимизаторами, поэтому в дальнейшем все расчеты проводились с использованием данного оптимизатора.

На рис. 5 показаны графики параметров точности обучения и валидации (*а*), а также графики параметров потерь обучения и валидации (*б*) при режиме обучения нейронной сети со скоростью $lr = 0{,}01$. Топология сети содержит два скрытых слоя, которые обозначены как $n_1 = 256$ и $n_2 = 128$ нейронов соответственно. В целом построенные кривые обладают

хорошей сходимостью, небольшой волатильностью, что говорит об устойчивых показателях нейронной сети.



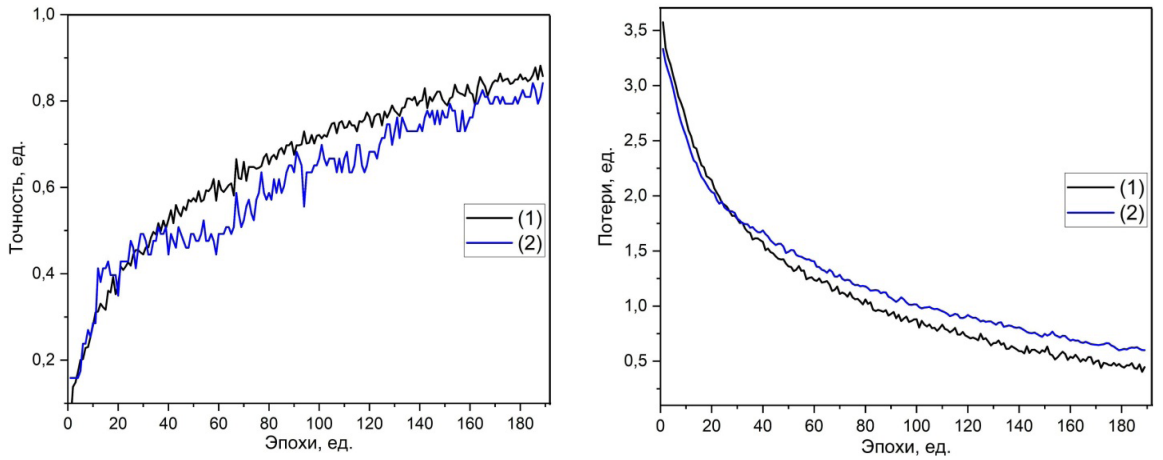
а *б*

Рис. 5. Расчетные значения параметров: точности обучения нейронной сети (1) и валидации (2) – *а*; график параметров: потерь обучения (1) и валидации (2) – *б* ($n_1 = 256$, $n_2 = 128$, $lr = 0{,}01$)

Fig. 5. Calculated values of the parameters: neural network training accuracy (1) and validation (2) – *a*; graph of parameters: training losses (1) and validation – *b*. ($n_1 = 256$, $n_2 = 128$, lr = 0,01)

Можно отметить закономерность на рис. 5, которая показывает, что по мере увеличения числа временных циклов (эпох) параметры точности обучения и валидации возрастают умеренно, с логарифмическим ростом функции, и стремятся к 1 (100 %), а потери в свою очередь – к 0.

Далее проводилось исследование изменения поведения параметров нейронной сети при скорости обучения $lr = 0{,}001$, что отражено на рис. 6, где приведен график точности обучения (*а*) и график потерь обучения (*б*) нейронной сети и валидации тестовых данных соответственно.

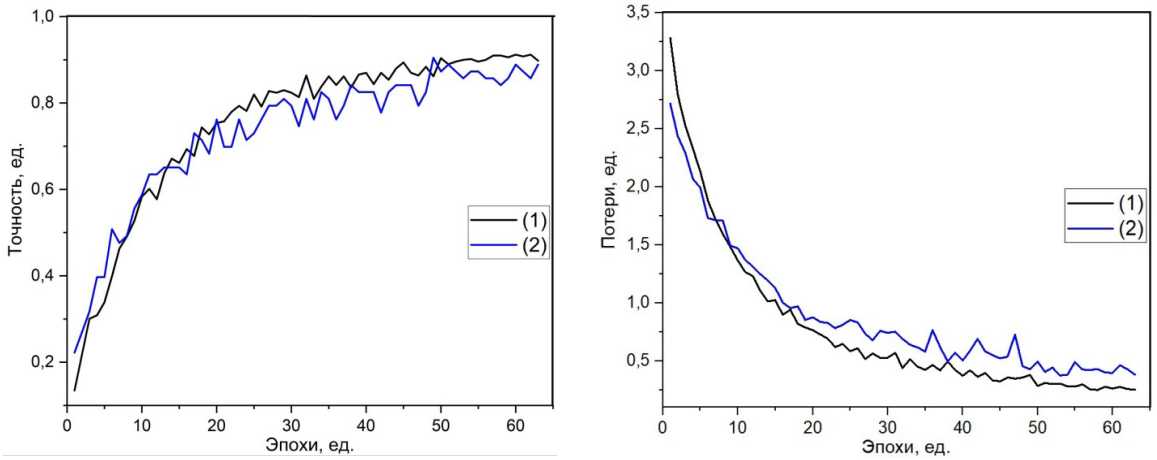


а *б*

Рис. 6. Расчетные значения параметров: точности обучения (1) нейронной сети и валидаци (2) тестовых данных – *а*, график параметров: потерь обучения (1) нейронной сети и валидации (2) – *б* ($n_1 = 256$, $n_2 = 128$, $lr = 0{,}001$)

Fig. 6. Calculated values of the parameters: training accuracy (1) of the neural network and validation (2) of test data – *a*, graph of parameters: training losses (1) of the neural network and validation (2) – *б* ($n_1 = 256$, $n_2 = 128$, $lr = 0{,}001$)

При уменьшении параметра скорости обучения *lr* в 10 раз, для случая когда *lr* = 0,001, количество эпох возрастает в ~2,6 раза, причем точность обучения и валидации (рис. 6, *а*) едва доходит до 90 %, а потери уменьшаются заметно медленнее (рис. 6, *б*), чем при *lr* = 0,01 (рис. 5, *б*). Далее исследовалась работа нейронной сети при различных параметрах n_1 и n_2 в скрытом слое. Пример рассчитанных значений параметров точности и потерь для нейронной сети при $n_1 = n_2 = 128$ представлены на рис. 7.



а *б*

Рис. 7. Расчетные значения параметров: точности обучения (1) нейронной сети и валидация (2) тестовых данных – *а*, график параметров: потерь обучения (1) нейронной сети и валидации (2) – *б* ($n_1 = 128$, $n_2 = 128$, *lr* = 0,01)

Fig. 7. Calculated values of the parameters: training accuracy (1) of the neural network and validation (2) of test data – *a*, graph of the parameters: training losses (1) of the neural network and validation (2) – *b* ($n_1 = 128$, $n_2 = 128$, *lr* = 0,01)

Уменьшив количество нейронов в первом слое до $n_1 = 128$, получим увеличение расхождения значений параметров между точностью обучения и валидацией, а также для потерь на обучение и валидацию (рис. 7), в сравнении с рис. 5. Данный факт указывает на то, что при такой архитектуре нейронная сеть плохо обобщает данные.

Проведенное исследование позволило реализовать формирование рекомендаций по улучшению образовательного процесса в веб-приложении ЦПО, была определена структура нейронной сети и ее следующие параметры: оптимизатор Adam, 2 скрытых слоя $n_1 = 128$ и $n_2 = 128$, *lr* = 0,01, функция активации в скрытых слоях ReLU, в выходном – softmax и для функции потерь выбрана разреженная категориальная кросс-энтропия. Для определения параметра точности предсказаний модели была выбрана метрика *accuracy*. Чтобы избежать переобучения нейронной сети, использовался метод регуляризации, остановка обучения (EarlyStopping), где обучение прекращалось, если происходило 10 эпох без улучшений.

5. Поддержка принятия решений с использованием рекомендаций нейронной сети

Методологической основой при структурном анализе ситуации, связанной с принятием решения, служит системный подход, в основе которого лежит рассмотрение объекта или ситуации как системы. В работе при структурном анализе ситуации использовался SWOT-анализ, а для изучения внешнего воздействия на систему применялся PEST-анализ. При проектировании основных модулей системы поддержки принятия решений применялись методология

структурного проектирования (SADT) и технология объектно-ориентированного анализа (ООА).

Система поддержки принятия управленческих решений с использованием когнитивной модели обучения объединяет несколько различных программных инструментов, методик и моделей управления (рис. 8). Взаимодействуя между собой, они способствуют оптимизации образовательного процесса с учетом индивидуальных когнитивных особенностей обучающихся.

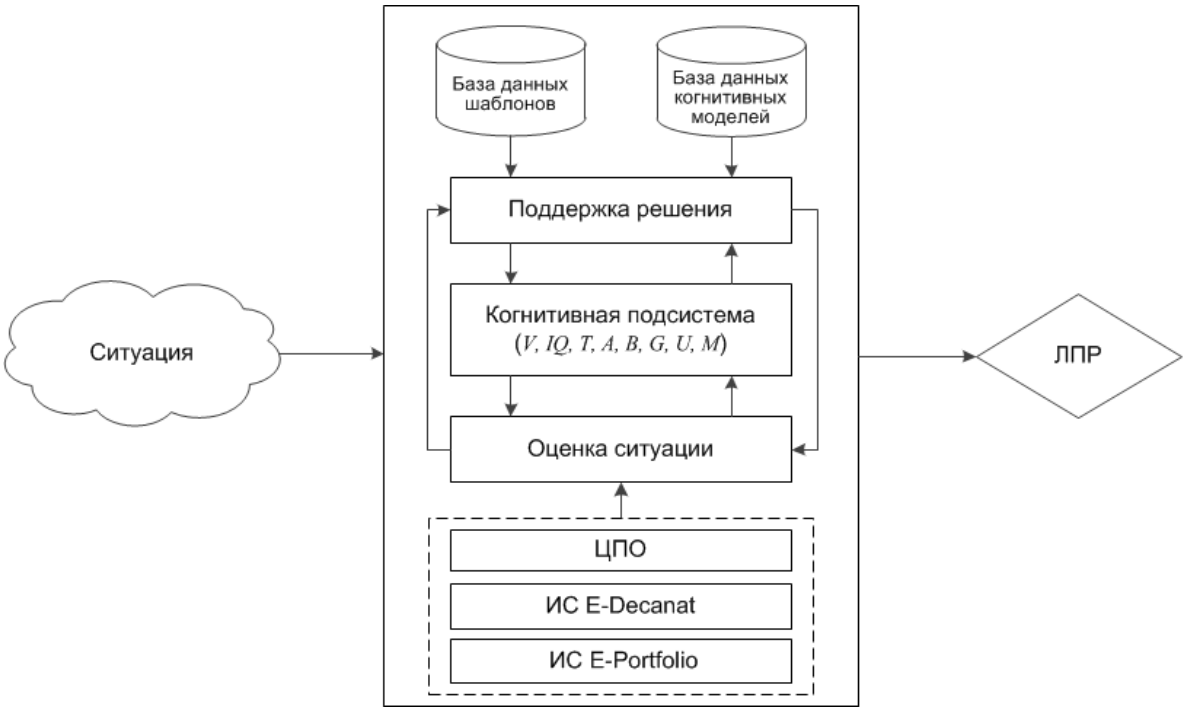


Рис. 8. Схема поддержки принятия решений с использованием когнитивной модели
Fig. 8. Decision support scheme using a cognitive model

Данная схема позволит персонализировать образовательные программы, адаптировать методики преподавания, а также создать оценочные инструменты, соответствующие требованиям работодателей и способностям учащихся.

Цифровой профиль обучающегося реализован как встроенное веб-приложение системы е-портфолио ТГПУ. Являясь компонентом информационной среды управления учебным процессом, ЦПО тесно связан и взаимодействует с внутренними информационными подсистемами университета (табл. 2). При формировании управленческого решения интеллектуальная система учитывает индивидуальные стремления студента в профессиональной самореализации, стратегию формирования основ его творческого развития, а также перспективные направления проектной деятельности вуза с целью привлечения студентов к участию в научных и образовательных конкурсах, выполнению инновационных проектов и научных исследований, тем самым проводя опережающую подготовку кадров и формируя высококлассных специалистов.

Таким образом, обработка учебной аналитики (LA/EDM) с использованием полносвязной нейронной сети позволяет упростить процесс формирования индивидуальных образовательных траекторий и дает возможность повысить удовлетворенность обучающегося учебным процессом в вузе.

Таблица 2

Поддержка принятия решений с использованием ЦПО

Table 2

Support for decision-making using SDP

№	Наименование подразделения вуза	Программные компоненты, модули	Форма взаимодействия и поддержка принятия решений
1	Учебное управление	ЦПО, е-портфолио, ЭИОС	Данные из интеллектуальной подсистемы передаются в подсистему проведения конкурсов по научной/общественной деятельности. Улучшение качества проводимых конкурсов. Решения о назначении стипендии. Выдвижение кандидатов и формирование научного и общественного актива
2	Деканаты (учебные офисы)	ЦПО, e-Decanat, е-портфолио	Построение персонализированных образовательных траекторий. Выдвижение кандидатов на научную и общественную деятельность. Решение о выдвижении кандидатов на научную и общественную деятельность. Формирование временных творческих коллективов обучающихся для реализации университетских проектов
3	Центр карьеры	ЦПО, online-сервис для передачи данных е-портфолио, веб-сайт	Передача данных для работодателей с согласия обучающихся на веб-сайт центра. Улучшение анализа потребностей работодателей. Сокращение времени, затрачиваемого при найме на работу обучающихся/выпускников
4.	Учебные отделы и другие подразделения	ЦПО, e-Decanat, е-портфолио, ЭИОС	Анализ когнитивных возможностей позволяет производить корректировку образовательных/научных программ, улучшить конкурентоспособные характеристики вуза

В табл. 3 представлен результат работы нейронной сети, где приведен вывод семи рекомендаций для улучшения образовательного процесса. Вероятности полученных рекомендаций нормируются в программе с использованием *softmax*.

Нейронная сеть выбирает рекомендации из предварительно подготовленного словаря. Рекомендации словаря формируются на основе интерпретации данных психодиагностического комплекса тестов и являются обобщением числовых результатов тестирования, так как выбираются на основе когнитивных параметров студента. Выбранный подход к формированию рекомендаций направлен на развитие объективной самооценки студента и поможет ему осознать свои сильные и слабые стороны (выбрать образовательные траектории), а также развить навыки в различных областях деятельности и улучшить соответствующие когнитивные показатели.

Таблица 3

Пример рекомендаций ЦПО, сформированных с использованием нейронной сети

Table 3

Example of SDP recommendations generated using a neural network

№	Рекомендация	Вероятность
1	Регулярно проводить самооценку своих знаний и навыков, чтобы выявлять слабые места и работать над их улучшением	0,86818
2	Участвовать в регулярных консультациях с преподавателями для обсуждения вопросов и повышения уровня понимания материала	0,08678
3	Практиковать методики концентрации внимания, такие как медитация или специальные упражнения, для повышения продуктивности в учебе	0,02296
4	Применять различные стратегии запоминания информации, основанные на индивидуальных особенностях памяти	0,01505
5	Внедрить техники «мозгового штурма» для развития креативного мышления и генерации новых идей	0,00274
6	Внедрить в учебный процесс приложения для формирования графических конспектов и ментальных карт	0,00268
7	Искать дополнительные образовательные ресурсы, такие как онлайн-курсы или литературу, для более глубокого изучения интересующих тем	0,00118

Заключение

Представлена модель системы поддержки принятия решений в образовательном процессе на основе когнитивной модели обучения как эффективный элемент принятия управленческих решений. Разработано программное решение, цифровой профиль обучающегося на базе электронного портфолио студентов с привлечением алгоритмов искусственного интеллекта, современных веб-технологий, а также когнитивных моделей обучения, что позволит сформировать блок рекомендаций для обучающегося. Используя представленный в работе когнитивный подход и программное обеспечение, становится возможным построение когнитивных моделей обучающихся, а также использование более широкого класса открытых моделей обучения (OLM).

В ходе исследования свойств нейронной сети было проведено сравнение оптимизаторов (Adam, RMSprop, Adamax, SGD), варьирование количества нейронов в скрытых слоях, а также оценка влияния параметра скорости обучения. Проведенное исследование позволило реализовать формирование рекомендаций по улучшению образовательного процесса, была определена структура нейронной сети и конфигурационные параметры: оптимизатор Adam, топология слоев, режим обучения сети. Разработанное программное решение (ЦПО) с использованием нейронной сети позволит сократить время принятия управленческих решений в части организации различных конкурсов (по научной, общественной и другой деятельностей), ускорить организацию временных творческих коллективов обучающихся при выполнении проектов, а также повысить уровень персонализированного обучения.

В дальнейшем планируется совершенствование и более глубокое обучение нейронной сети, расширение спектра ее возможностей. Анализ моделей поведения обучающихся с ис-

пользованием ЦПО позволит преподавателям и административным работникам университета своевременно принимать решения о помощи и выделении дополнительных ресурсов обучающимся, а также позволит проводить коррекцию индивидуальных траекторий и тем самым повысит вовлеченность обучающихся в учебный процесс.

Список литературы

1. **Малкова Т. В.** Цифровая трансформация как средство модернизации образовательного процесса и достижения качественных изменений в различных аспектах образовательной деятельности // Современная наука. 2023. № 1. С. 46–48.
2. **Авдеева З. К., Коврига С. В., Макаренко Д. И., Максимов В. И.** Когнитивный подход в управлении // Проблемы управления. 2007. № 3. С. 2–8.
3. **Лисовский А. Л.** Применение нейросетевых технологий для разработки систем управления // Стратегические решения и риск-менеджмент. 2020. Т. 11, №4. С. 378–389. DOI: 10.17747/2618-947X-2020-4-378-389
4. **Шамсутдинова Т. М.** Проблемы и перспективы применения нейронных сетей в сфере образования // Открытое образование. 2022. Т. 26, № 6. С. 4–10. DOI: 10.21686/1818-4243-2022-6-4-10
5. Федеральный проект «Искусственный интеллект». Министерство экономического развития РФ. URL: https://ai.gov.ru/strategy/federalnyy-proekt-ii/?ysclid=ly6szfvxwp523566810 (дата обращения: 04.07.2024).
6. **Hao G., Kenneth K, Brian J.** Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement // Lecture Notes in Computer Science. 2006. Vol. 4053. P. 164–175. DOI: 10.1007/11774303_17
7. **Binali T., Tsai C. C., Chang H. Y.** University students' profiles of online learning and their relation to online metacognitive regulation and internet-specific epistemic justification // Computers and Education. 2021. Vol. 175. P. 104315. DOI: 10.1016/j.compedu.2021.104315
8. **Zhao Y., Lorente A.P., Gómez M. C.** Digital competence in higher education research: A systematic literature review // Computers and Education. 2021. Vol. 168. P. 104212. DOI: 10.1016/j.compedu.2021.104212.
9. **Matzavela, V., Alepis E.** Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments // Computers and Education: Artificial Intelligence. 2021. Vol. 2. P. 100035. DOI: 10.1016/j.caeai.2021.100035.
10. **Luo Q., Yang J.** The Artificial Intelligence and Neural Network in Teaching // Computational intelligence and neuroscience. 2022. 1778562. DOI: 10.1155/2022/1778562.
11. **Морозевич Е. С., Коротких В. С., Кузнецова Е. А.** Разработка модели формирования индивидуальных образовательных траекторий с использованием методов машинного обучения // Бизнес-информатика. 2022. Т. 16, № 2. С. 21–35. DOI: 10.17323/2587-814X.2022.2.21.35
12. **Пираков Ф. Д., Клишин А. П., Ерёмина Н. Л., Клыжко Е. Н.** Разработка и применение системы электронного портфолио обучающегося в вузе // Вестник НГУ. Серия: Информационные технологии. 2019. Т. 17, № 4. С. 87–100. DOI 10.25205/1818-7900-2019-17-4-5-87-100
13. **Пираков Ф. Д., Шталина Е. С., Клишин А. П.** Управление учебным процессом в вузе с использованием цифрового профиля выпускника и электронного портфолио // Прикладная математика и информатика: современные исследования в области естественных и технических наук: сб. материалов IX Междунар. науч.-практ. конф. (школы-семинара) молодых ученых. Тольятти: ТГУ, 2023. С. 489–494.

14. **Ахметова Л. В.** Когнитивная сфера личности – психологическая основа обучения // Вестник Том. гос. пед. ун-та. 2009. Т. 87, № 9. С. 108–115.
15. **Ахметова Л. В., Иванкина Л. И., Языков К. Г.** Нейропсихологические и философские основы понятия «когнитивная сфера личности». Томск: НТЛ, 2021. 235 с.

References

1. **Malkova T. V.** Digital transformation as a means of modernizing the educational process and achieving qualitative changes in various aspects of educational activity. *Modern science*, 2023, no. 1, pp. 46–48. (in Russ.)
2. **Avdeeva Z. K., Kovriga S. V., Makarenko D. I., Maksimov V. I.** Kognitivnyy podkhod v upravlenii [Cognitive approach to management]. Management problems, 2007, no. 3, p. 2–8. (in Russ.)
3. **Lisovsky A. L.** Application of neural network technologies for management development of systems. *Strategic decisions and risk management*, 2020, vol. 4, no. 11, p. 378–389. DOI: 10.17747/2618-947X-2020-4-378-389 (in Russ.)
4. **Shamsutdinova T. M.** Problems and Prospects for the Application of Neural Networks for the Sphere of Education. *Open education*, 2022. vol. 26, no. 6, p. 4–10. DOI: 10.21686/1818-4243-2022-6-4-10 (in Russ.)
5. Federal project “Artificial Intelligence”. Ministry of Economic Development of the Russian Federation. URL: https://ai.gov.ru/strategy/federalnyy-proekt-ii/?ysclid=ly6szfvxwp523566810 (date of access: 04.07.2024) (in Russ.)
6. **Hao G., Kenneth K, Brian J.** Learning Factors Analysis – A General Method for Cognitive Model Evaluation and Improvement. *Lecture Notes in Computer Science*, 2006, vol. 4053, pp. 164–175. DOI: 10.1007/11774303_17.
7. **Binali T., Tsai C. C., Chang H. Y.** University students’ profiles of online learning and their relation to online metacognitive regulation and internet-specific epistemic justification. *Computers and Education*, 2021, vol. 175, pp. 104315. DOI: 10.1016/j.compedu.2021.104315.
8. **Zhao Y., Lorente A.P., Gómez M. C.** Digital competence in higher education research: A systematic literature review. *Computers and Education*, 2021, vol. 168, pp. 104212. DOI: 10.1016/j.compedu.2021.104212.
9. **Matzavela V., Alepis E.** Decision tree learning through a Predictive Model for Student Academic Performance in Intelligent M-Learning environments. *Computers and Education: Artificial Intelligence*, 2021, vol. 2, p. 100035. DOI: 10.1016/j.caeai.2021.100035.
10. **Luo Q., Yang J.** The Artificial Intelligence and Neural Network in Teaching. *Computational intelligence and neuroscience*, 2022, 1778562. DOI: 10.1155/2022/1778562.
11. **Morozevich E. S., Korotkikh V. S., Kuznetsova E. A.** Razrabotka modeli formirovaniya individualnykh obrazovatel’nykh trayektoriy s ispolzovaniyem metodov mashinnogo obucheniya [Development of a model for the formation of individual educational trajectories using machine learning methods]. Business Informatics, 2022, vol. 16, no. 2, pp. 21–35. (In Russ.) DOI: 10.17323/2587-814X.2022.2.21.35
12. **Pirakov F. D., Klishin A. P., Eremina N. L, Klyzhko E. N.** Development and application of an electronic portfolio system for a student at a university. *Vestnik NSU. Series: Information technologies*, 2019, vol. 17, no. 4, pp. 87–100. (In Russ.) DOI 10.25205/1818-7900-2019-17-4-5-87-100
13. **Pirakov F. D., Shtalina E. S., Klishin A. P.** Upravleniye uchebnym protsessom v vuze s ispolzovaniyem tsifrovogo profilya vypusknika i elektronnogo portfolio [Managing the educational process at a university using a digital graduate profile and electronic portfolio]. *Proceedings of the IX International Scientific and Practical Conference (school-seminar) of*

young scientists: Applied mathematics and computer science: modern research in the field of natural and technical sciences. Tolyatti, TSU, 2023, pp. 489–494. (in Russ.)
14. **Akhmetova L. V.** Kognitivnaya sfera lichnosti – psikhologicheskaya osnova obucheniya [The cognitive sphere of personality is the psychological basis of learning]. *Bulletin of the Tomsk State Pedagogical University*, 2009, vol. 87, no. 9, pp. 108–115. (in Russ.)
15. **Akhmetova L. V., Ivankina L. I., Yazykov K. G.** Neyropsikhologicheskiye i filosofskiye osnovy ponyatiya «kognitivnaya sfera lichnosti» [Neuropsychological and philosophical foundations of the concept Cognitive sphere of personality]. Tomsk, NTL Publ. House, 2021. 235 p. (in Russ.)

Сведения об авторах

Клишин Андрей Петрович, кандидат физико-математических наук, заведующий студенческой научно-исследовательской лабораторией информационных технологий УИТ Томского государственного педагогического университета

Шталина Екатерина Сергеевна, бакалавр Томского государственного педагогического университета

Пираков Фаррух Джамшедович, аспирант кафедры автоматизации обработки информации Томского университета систем и радиоэлектроники

Ахметова Людмила Владимировна, кандидат психологических наук, доцент кафедры психологии и развития личности Томского государственного педагогического университета

Ерёмина Наталия Леонидовна, кандидат технических наук, доцент кафедры системного анализа и математического моделирования Томского государственного университета

Information about the Authors

Andrey P. Klishin, Candidate of Physical and Mathematical Sciences, Head of the Lab Student Research Laboratory of Information Technologies UIT Tomsk State Pedagogical University

Ekaterina S. Shtalina, Bachelor of Tomsk State Pedagogical University

Farrukh D. Pirakov, Graduate Student of the Department of automation of information processing, Tomsk University of Systems and Radioelectronics

Lyudmila V. Akhmetova, Candidate of Psychological Sciences, Associate Professor of the Department of Psychology and Personality Development, Tomsk State Pedagogical University

Natalia L. Eryomina, Candidate of Technical Sciences, Associate Professor of the Department of System Analysis and Math Modeling, Tomsk State University

Статья поступила в редакцию 31.10.2024;
одобрена после рецензирования 23.01.2025; принята к публикации 23.01.2025

The article was submitted 31.10.2024;
approved after reviewing 23.01.2025; accepted for publication 23.01.2025

Научная статья

УДК 004
DOI 10.25205/1818-7900-2024-22-4-49-61

Эффективность нейросетевых алгоритмов в автоматическом реферировании и суммаризации текста

Кирилл Вячеславович Ребенок

Московский финансово-юридический университет МФЮА,
Москва, Россия

rebenokkv@gmail.com, https://orcid.org/0009-0003-2015-033X

Аннотация

Статья посвящена анализу роли и эффективности нейросетевых алгоритмов в задачах автоматического реферирования и суммаризации текстов, которые являются ключевыми в области обработки естественного языка (NLP). Основная цель автоматического реферирования — извлечение и генерация важнейшей информации из текстов для обеспечения быстрого доступа к основному содержанию без необходимости читать весь документ. В статье рассматриваются основные проблемы, с которыми сталкиваются разработчики при реализации алгоритмов реферирования, включая понимание контекста, иронии, сохранение связности текста, адаптацию к разным языкам и стилям. Особое внимание уделяется нейросетевым моделям, таким как Transformer, BERT и GPT, которые благодаря своей способности обучаться на больших объемах данных показали выдающуюся эффективность в автоматическом реферировании текстов. Статья также освещает вклад ведущих ученых в области глубокого обучения и анализирует методы, лежащие в основе современных алгоритмов NLP, подчеркивая значимость непрерывного технологического прогресса в улучшении качества реферирования и доступности информации. Статья будет интересна широкому кругу читателей, включая исследователей в области искусственного интеллекта и NLP, разработчиков программного обеспечения, занимающихся автоматизацией обработки текстов, а также специалистов в областях, где требуется быстрая обработка и анализ больших объемов текстовой информации, таких как юридическая практика, медицинская диагностика и научные исследования. Кроме того, материал статьи будет полезен преподавателям и студентам, изучающим технологии обработки данных и искусственного интеллекта, предоставляя им актуальные примеры применения теоретических знаний в практических проектах.

Ключевые слова

естественный язык, NLP, нейросетевые алгоритмы, метрики, автоматическое реферирование, суммаризация текста

Для цитирования

Ребенок К. В. Эффективность нейросетевых алгоритмов в автоматическом реферировании и суммаризации текста // Вестник НГУ. Серия: Информационные технологии. 2024. Т. 22, № 4. С. 49–61. DOI 10.25205/1818-7900-2024-22-4-49-61

Efficiency of Neural Network Algorithms in Automatic Abstracting and Summarization Text

Kirill V. Rebenok

Moscow University of Finance and Law MFUA,
Moscow, Russian Federation

rebenokkv@gmail.com, https://orcid.org/0009-0003-2015-033X

Abstract

The article is devoted to the analysis of the role and efficiency of neural network algorithms in the tasks of automatic abstracting and summarization of texts, which are key in the field of natural language processing (NLP). The main goal of automatic abstracting is to extract and generate essential information from texts to provide quick access to the main content without having to read the whole document. The paper discusses the main challenges faced by developers in implementing abstracting algorithms, including understanding context, irony, maintaining text cohesion, and adapting to different languages and styles. Special attention is given to neural network models such as Transformer, BERT, and GPT, which have shown outstanding performance in automatic text abstracting due to their ability to learn on large amounts of data. The article also highlights the contributions of leading researchers in the field of deep learning and analyzes the methods underlying state-of-the-art NLP algorithms, highlighting the importance of continuous technological progress in improving abstracting quality and information accessibility. The article will be of interest to a wide range of readers, including researchers in the field of artificial intelligence and NLP, software developers engaged in automation of text processing, as well as specialists in areas where fast processing and analysis of large amounts of textual information is required, such as legal practice, medical diagnostics and scientific research. In addition, the material of the article will be useful for teachers and students studying data processing and artificial intelligence technologies, providing them with actual examples of applying theoretical knowledge in practical projects.

Keywords

natural language, NLP, neural network algorithms, metrics, automatic abstracting, text summarization

For citation

Rebenok K. V. Efficiency of neural network algorithms in automatic abstracting and summarization text. *Vestnik NSU. Series: Information Technologies*, 2024, vol. 22, no. 4, pp. 49–61 (in Russ.) DOI 10.25205/1818-7900-2024-22-4-49-61

Введение

Автоматическое реферирование и суммирование текстов – важнейшая из задач в области обработки естественного языка (NLP), призванная упростить доступ к информации и улучшить ее восприятие. В условиях развития Интернета и цифровых технологий объем доступной текстовой информации стремительно растет, и ее полноценное изучение и анализ в ручном режиме становится невозможным. В этой связи автоматическое реферирование текста – это процесс создания краткого и сжатого резюме длинного документа с сохранением его основного содержания и ключевых идей.

Цель автоматического реферирования заключается в том, чтобы извлекать или генерировать наиболее важную информацию из текста таким образом, чтобы конечный пользователь мог быстро получить представление о его содержании без необходимости читать весь документ. В особенности это актуально для новостных статей, научных публикаций, юридических документов и любых других областей, где требуется быстрый доступ к сжатой форме информации.

Многие разработчики и пользователи технологий NLP нередко сталкиваются с проблемами при автоматическом реферировании и резюмировании текстов, включая понимание контекста и иронии, сохранение связности и логичности изложения, адаптацию к различным языкам и стилям письма.

Нейросетевые модели играют одну из ключевых ролей в решении этих задач благодаря своей способности обучаться на больших объемах данных и выявлять сложные зависимости

в тексте. Модели глубокого обучения, такие как Transformer, BERT, GPT и их производные, продемонстрировали потрясающую производительность в задачах автоматического реферирования, научившись генерировать краткие и осмысленные резюме текстов на основе обучающих примеров. Технологический прогресс непрерывно стремится повысить качество реферирования и сделать автоматическое реферирование максимально доступным и эффективным инструментом для обработки информации в различных областях.

Проблема автоматического реферирования и резюмирования текстов активно изучается множеством исследователей, так как она имеет ключевое значение для развития области обработки естественного языка (NLP). В их числе следует выделить таких ученых, как Дж. Хинтон, Я. Лекун и Й. Бенджио, внесших значительный вклад в развитие глубокого обучения и нейронных сетей. Исследования И. Суцкевера, А. Крисевски и Д. Сильвера позволяют лучше понять методы нейросетевого обучения, лежащие в основе современных алгоритмов NLP. Важное место в области машинного обучения занимают работы таких авторов, как С. Рудер и Т. Вольф, развивающих и адаптирующих такие алгоритмы, как Transformer и BERT, с целью эффективного извлечения и создания кратких изложений текста. Примечателен также вклад Л. Куна и его работа над моделями GPT, показавшая отличные результаты в генеративных задачах NLP, включая резюмирование.

В отечественной научной среде проблема автоматического реферирования и резюмирования текстов не получила широкого распространения, несмотря на мировой интерес к этой области. Недостаток ресурсов и доступа к большим данным, а также трудности, связанные с особенностями русского языка, тормозят прогресс в разработке соответствующих алгоритмов. В то же время потенциал дальнейшего развития все же существует благодаря наличию квалифицированных специалистов и растущему интересу к технологиям искусственного интеллекта. Упор на развитие специализированных образовательных программ и укрепление связей между академическими, исследовательскими и коммерческими организациями может стать ключом к преодолению существующих барьеров и стимулированию развития этой области.

Материалы и методы

В рамках данного исследования были использованы актуальные научные публикации, освещающие вопросы нейросетевых алгоритмов, больших данных, машинного обучения, реферировании и суммаризации текстов, методов интеллектуального анализа данных. В этой статье активно применялись различные методологические подходы: монографический анализ для детального исследования темы, оценочный подход для анализа и интерпретации данных, а также метод рефлексии, позволяющий глубоко осмыслить и критически оценить полученные результаты.

Результаты исследования

Анализ моделей обработки естественного языка (NLP) имеет решающее значение для оценки их эффективности, удобства использования и надежности в реальных приложениях. Метрики, включая BLEU, ROUGE, METEOR и BERTScore, играют ключевую роль в этом процессе, предоставляя количественные показатели эффективности модели. Метрики помогают оценить, насколько эффективно модель справляется с задачами перевода, обобщения, генерации или понимания текста в сравнении с человеческими оценками или эталонными данными. Такая оценка жизненно важна не только для точной настройки и улучшения моделей, но и для обеспечения их соответствия необходимым стандартам, необходимым для развертывания в чувствительных приложениях, таких как медицинская диагностика, юридический анализ или автоматизация обслуживания клиентов.

Обсуждение результатов

За последние годы текстовые генеративные модели ИИ добились значительных успехов в решении задач обработки естественного языка, таких как перевод, обобщение текста и создание диалогов. Они способны генерировать текст, зачастую неотличимого от человеческого, что делает их все более популярными в различных отраслях, включая обслуживание клиентов, создание контента и анализ данных. И хотя эти модели могут быть невероятно мощными и полезными, они также могут выдавать неожиданные или даже пагубные результаты, в связи с чем за ними необходимо внимательно следить.

При автоматическом реферировании и суммировании текстов используются различные нейросетевые алгоритмы, каждый из которых обладает уникальными особенностями и принципами работы. Традиционно эти алгоритмы можно разделить на генеративные и извлекающие методы резюмирования.

Извлекающий метод резюмирования работает путем выделения и копирования ключевых фраз или предложений из исходного текста для формирования сжатого содержания. При таком подходе исходный текст не изменяется, а фильтруется и изымается наиболее значимая его часть. При этом методы извлечения основываются на анализе важности слов и предложений в документе с помощью таких метрик, как частота слов, положение предложения в тексте и связность предложений [1].

Генеративные методы обобщения, напротив, создают новые предложения, не обязательно присутствующие в исходном тексте, чтобы отразить суть содержания. При использовании этих методов глубокое обучение позволяет генерировать связные и последовательные сводки, которые могут включать перефразирование или обобщение информации. Модели генеративного типа требуют более сложных алгоритмов и значительных вычислительных ресурсов, но способны создавать гораздо качественные и естественные обобщения [2].

Определение эффективности нейросетевых алгоритмов в задаче автоматического реферирования и резюмирования текстов включает в себя оценку различных аспектов, таких как точность, качество генерируемых обобщений, их релевантность и связность, а также способность алгоритмов адаптироваться к различным текстовым данным.

Рассмотрим более подробно каждую из метрик, которые применяются для автоматического реферирования и суммирования текста.

Это набор показателей, используемых для оценки автоматического обобщения и машинного перевода. Он сравнивает автоматически созданное резюме или перевод с набором справочных резюме (обычно написанных человеком). ROUGE [3] измеряет качество резюме путем подсчета количества перекрывающихся единиц, таких как n-граммы[1], последовательности слов и пары слов между текстом, созданным моделью, и справочными текстами.

Наиболее распространенные варианты ROUGE:

- ROUGE-N, фокусирующийся на n-граммах (фразах из N слов);
- ROUGE-1 и ROUGE-2 наиболее распространенные, которые фокусируются на униграммах[2] и биграммах[3] соответственно;

[1] N-граммы – это последовательности из *n* слов, извлеченных из текста. Например, в предложении «Я иду домой» биграммы (2-граммы) будут «Я иду» и «иду домой». N-граммы используются для оценки степени сходства между двумя текстами на уровне словесных последовательностей.

[2] Униграммы – это одиночные слова, извлечённые из текста. В контексте обработки текста и анализа данных униграммы представляют собой самые простые элементы, используемые для анализа и сравнения текстов. Они могут быть использованы для статистического анализа частотности слов, оценки сходства текстов и других задач, связанных с языковыми моделями.

[3] Биграммы – это последовательности из двух слов, следующих друг за другом в тексте. Они являются основным инструментом в текстовом анализе и помогают уловить связи между словами, что важно для понимания структуры предложений и для создания статистических моделей языка.

• ROUGE-L, основанный на самой длинной общей подпоследовательности (LCS), учитывающий сходство структуры на уровне предложений и автоматически определяющий самые долгие последовательные n-граммы[4].

Рассмотрим пример применения метрики ROUGE (табл. 1).

Таблица 1

ROUGE: практическая реализация (составлено автором)

Table 1

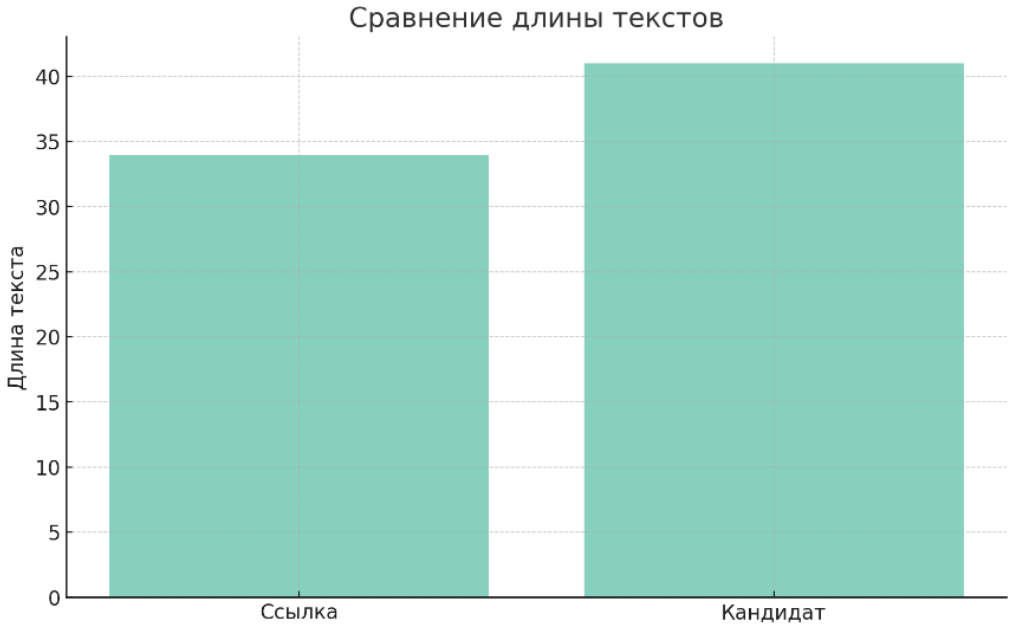
ROUGE: practical implementation (compiled by the author)

Тип текста	Текст	Ключевые аспекты	Сравнение с эталонным суммарием
Исходный текст	«Глобальное потепление вызывает серьёзные изменения в климате Земли, приводя к увеличению частоты и интенсивности экстремальных погодных явлений. Эти изменения угрожают сельскому хозяйству, водным ресурсам и здоровью человека, требуя незамедлительных действий по сокращению выбросов парниковых газов и адаптации к уже неизбежным последствиям»	Глобальное потепление, климатические изменения, угрозы экосистемам, сельскому хозяйству, водным ресурсам, здоровью, требование действий	Охватывает весь спектр проблем, связанных с глобальным потеплением
Эталонное резюме	«Глобальное потепление угрожает экосистемам, увеличивая экстремальные погодные явления и требует срочных мер по сокращению выбросов и адаптации»	Угроза экосистемам, увеличение экстремальных явлений, необходимость срочных мер по сокращению выбросов и адаптации	Суммирует основные точки исходного текста, упрощая детали
Сгенерированное суммарие	«Увеличение частоты экстремальных погодных явлений из-за глобального потепления требует действий для снижения выбросов парниковых газов»	Увеличение экстремальных погодных явлений, необходимость снижения выбросов парниковых газов	Сфокусировано на погодных явлениях и снижении выбросов

Реализация данного примера на языке Python будет выглядеть, как изображено на рисунке.

BLEU − метрика, разработанная исследователями IBM для оценки текста, которая позволяет оценить точность перевода путем измерения совпадения n-грамм между машинносгенерированным текстом и набором высококачественных эталонных переводов. Основное внимание уделяется точности. BLEU славится своей простотой и эффективностью, что делает его эталоном в области машинного перевода. Вместе с тем он оценивает лексическое сходство на поверхностном уровне, часто упуская из виду более глубокие семантические и контекстуальные нюансы языка [4−6].

[4] ROUGE and BLEU scores for NLP model evaluation. URL: https://clementbm.github.io/theory/2021/12/23/rouge-bleu-scores.html (дата обращения: 12.03.2024).



Сравнение длины текстов для примера использования метрики ROUGE
Comparison of text lengths for an example using the ROUGE metric

Среди основных недостатков использования таких метрик, как BLEU или ROUGE, можно назвать тот факт, что эффективность работы моделей генерации текстов зависит от точных совпадений. Возможно, точные совпадения важны для таких сценариев использования, как машинный перевод, но для генеративных моделей ИИ, которые пытаются генерировать осмысленные и похожие тексты на основе имеющегося массива данных, точные совпадения могут быть не очень верными.

Метрика METEOR, разработанная для более глубокой оценки машинного перевода, призвана устранить такие недостатки BLEU, как недооценка семантической связности текста. В отличие от BLEU, METEOR учитывает не только точные совпадения слов, но также включает основы и синонимы для оценки переводов, что позволяет охватить более широкий диапазон лингвистических сходств [7]. METEOR позволяет точно и эффективно оценивать качество переводов текстов. Он учитывает не только точность перевода, но и то, насколько легко запомнить переведённый текст, добавляя штрафы за изменения в порядке слов. METEOR отличается высокой согласованностью с оценками, которые дают люди, особенно при анализе отдельных предложений. Это делает его подходящим для тщательной оценки качества переводов. Однако стоит отметить, что METEOR более сложен и требует больше ресурсов для расчётов по сравнению с более простыми методами, такими как BLEU [8; 9].

Для реализации метрики METEOR в Python можно использовать библиотеку NLTK, которая, помимо прочего, предоставляет инструменты для работы с метрикой BLEU. В отличие от BLEU, стандартная поддержка METEOR в NLTK отсутствует, что требует более сложной реализации с использованием внешних инструментов или создания собственной функции расчета.

METEOR, как и BLEU, предназначена для оценки качества машинного перевода, сравнивая сгенерированный текст с одним или несколькими эталонными переводами. Она учитывает точность и полноту, а также синонимичность и порядок слов. Тем не менее можно адаптировать подходы к расчету METEOR для оценки суммаризации текстов. Прямая реализация METEOR на Python может быть достаточно сложной из-за необходимости учитывать синони-

мы, морфологический анализ и порядок слов. В качестве альтернативы можно использовать готовые реализации или обращаться к инструментам, таким как Meteor Universal Tool, который предоставляется в виде Java-приложения.

BERTScore − это новая метрика для оценки качества созданных текстов, которая опирается на передовые технологии в области искусственного интеллекта. Она использует модель глубокого обучения под названием BERT для анализа текстов. Модель способна учитывать контекст каждого слова в тексте, что помогает оценить, насколько хорошо сгенерированный текст соответствует оригинальному или эталонному тексту. BERTScore не просто измеряет поверхностное совпадение слов, но и анализирует глубокое семантическое сходство, используя метод подсчёта косинусного сходства, что позволяет более точно оценить качество текста, учитывая его смысловое содержание [10]. Такой подход позволяет ИТ-специалистам оценивать качество создания текста с акцентом на семантическое содержание и контекст, что делает его более чувствительным к смыслу, передаваемому в тексте. Несмотря на то что BERTScore предлагает более детальную оценку, чем традиционные метрики, основанные на перекрытии, ее использование требует больших вычислительных затрат и ресурсов, поскольку она опирается на большие, предварительно обученные языковые модели.

Выбор наиболее эффективной метрики зависит от специфических целей оценки суммаризации. Если важно оценить точность воспроизведения конкретных фактов и данных, ROUGE может быть наиболее подходящей. Для более общего анализа качества перефразирования и семантической близости текстов лучше подойдут METEOR и BERTScore.

Преимущества и недостатки рассмотренных метрик представлены в табл. 2. Чтобы контролировать работу генеративных моделей, важно применять комплексный подход. Например, когда есть эталонный текст для сравнения, можно использовать BLEU для оценки точности перевода или ROUGE для измерения полноты.

Meteor, с другой стороны, учитывает и точность, и полноту, показывает хорошую корреляцию с результатами человеческих оценок как на уровне предложений, так и на уровне сегментов. BERTScore полезен для оценки семантической близости между генерируемым и эталонным текстом с помощью контекстуализированных вкраплений слов[5].

Недавнее исследование метрик BLEU, METEOR и BERTScore показало, что они не всегда эффективно различают критические и некритические ошибки перевода, особенно когда ошибка перевода изменяет сентимент[6] сообщения[7] [11; 12]. Это показывает, что при оценке качества перевода важно использовать комплексный подход и сочетать различные метрики для получения наиболее полной картины. С учетом этих факторов, важно подходить к выбору метрик для оценки качества суммаризации и машинного перевода с учетом специфики задачи и ограничений каждой метрики. При обработке естественного языка (NLP) ключевую роль играют методы, основанные на вычислении векторных представлений (embeddings). С их помощью слова, предложения или документы преобразуются в векторы чисел, что делает их пригодными для компьютерной обработки.

Нейросетевые алгоритмы суммаризации текстов находят широкое применение в различных отраслях, значительно увеличивая эффективность работы специалистов и делая информацию более доступной для общественности. Так, в юридической практике суммаризация текстов облегчает анализ и обработку большого количества юридических документов, таких

[5] Evaluating NLP Models: A Comprehensive Guide to ROUGE, BLEU, METEOR, and BERTScore Metrics. URL: https://plainenglish.io/community/evaluating-nlp-models-a-comprehensive-guide-to-rouge-bleu-meteor-and-bertscore-metrics-d0f1b1 (дата обращения: 12.03.2024); Tekgul H. Monitoring Text-Based Generative AI Models Using Metrics Like Bleu Score. URL: https://arize.com/blog-course/generative-ai-metrics-bleu-score/ (дата обращения: 12.03.2024).

[6] Сентимент сообщения, или анализ тональности текста, − это процесс определения эмоциональной окраски текста.

[7] BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text. URL: https://ar5iv.labs.arxiv.org/html/2109.14250 (дата обращения: 12.03.2024).

Таблица 2

Преимущества и недостатки метрик для суммаризации и реферирования текста (составлено автором)

Table 2

Advantages and disadvantages of metrics for text summarization and abstracting (compiled by author)

Метрика	Преимущества	Недостатки
BLEU	Простая и широко используемая. Хорошо подходит для оценки точности на уровне слов и фраз	Не учитывает семантическую связность и может приводить к искаженной оценке качества при слишком буквальном переводе
ROUGE	Учитывает как точность, так и полноту, предоставляя более сбалансированную оценку	Может быть менее эффективной для языков с сложной структурой из-за фокуса на совпадении слов и фраз
METEOR	Более сбалансированная метрика, включающая как точность, так и полноту, с учетом синонимов и парадигм	Сложнее в реализации и вычислении, чем BLEU или ROUGE
BERTScore	Использует контекстные вложения для оценки семантической близости, захватывая более тонкие нюансы языка	Требует больших вычислительных ресурсов и может быть зависимой от качества предварительно обученной модели BERT
Self-BLEU	Помогает оценить разнообразие в сгенерированном тексте, предотвращая избыточное повторение	Не учитывает качество содержания; фокусируется только на разнообразии
WMD	Эффективно оценивает семантическую близость на основе расстояния между словами в векторном пространстве	Может быть вычислительно затратной для длинных текстов и менее точной для очень коротких фраз

как законы, судебные решения и договоры, что позволяет юристам экономить время на подготовку к делам, улучшая понимание существующих прецедентов и законодательных требований. А в академическом мире алгоритмы суммаризации текстов значительно упрощают литературный обзор и анализ научных публикаций. Исследователи могут быстрее ознакомиться с последними достижениями в своей области, выявляя ключевые идеи и результаты из больших объемов научной литературы.

Техники резюмирования текстов делятся на две категории: экстрактивную и абстрактивную. Каждая из этих техник используется для создания краткого изложения длинного текста, но они делают это разными способами.

Экстрактивное резюмирование — это процесс выбора ключевых предложений или фраз непосредственно из исходного текста и их компиляции для создания резюме. Эта техника не вносит изменений в текст: она просто извлекает наиболее значимые части. Основные характеристики:

- система определяет и извлекает наиболее информативные предложения или фразы из текста на основе различных метрик, таких как частотность слов, важность темы и так далее;
- извлеченные предложения остаются неизменными, сохраняя оригинальный стиль и структуру автора;
- так как не требуется генерация нового текста, экстрактивное резюмирование может быть более простым и быстрым в реализации.

Таблица 3

Оценка эффективности различных языковых моделей в задачах резюмирования и суммаризацции текста*

Table 3

Evaluating the effectiveness of different language models in text summarization and summarization tasks

Название модели	**Разработчик**	**Год выпуска**	**Количество параметров**	**Архитектура**	**Преимущества**	**Недостатки**
BERT	Google AI	2018	110 миллионов	Transformer	Хорошо справляется с контекстом	Требует тонкой настройки
T5	Google AI	2019	11 миллиардов	Transformer	Гибкость в задачах	Ограничен в длине текста
RoBERTa	Facebook AI	2019	355 миллионов	Transformer	Улучшенная обработка текста	Требует больших вычислительных ресурсов
XLNet	Google/CMU	2019	340 миллионов	Transformer	Отличные результаты на различных задачах	Сложность в использовании
GPT-2	OpenAI	2019	1.5 миллиарда	Transformer	Хорошая генерация текста	Ограничения по контексту
GPT-3	OpenAI	2020	175 миллиардов	Transformer	Мощная генерация текста	Высокая стоимость
GPT-4	OpenAI	2023	более 100 миллиардов	Transformer	Улучшенное понимание и генерация текста	Требует значительных вычислительных ресурсов

* Составлена автором на основе [13–15].

Абстрактивное резюмирование переформулирует исходный текст, создавая новые предложения, которые могут не встречаться напрямую в исходном материале. Этот метод часто считается более сложным и продвинутым, так как требует глубокого понимания текста и способности к его творческой переработке. Основные характеристики:

- модель создает новые предложения, которые резюмируют оригинальный контент, используя передовые NLP-модели;
- лучше справляются с передачей основных идей текста в сжатой форме, поскольку они не ограничены только тем, что написано в исходнике;
- необходимы развитые алгоритмы понимания языка и генерации текста, такие как трансформеры и модели на основе искусственного интеллекта.

В табл. 3 представлены основные характеристики и возможности нескольких передовых языковых моделей, используемых для резюмирования текстов.

Изучение сравнительной таблицы языковых моделей для резюмирования и суммаризации текстов выявляет ключевые различия и потенциальные сферы применения каждой модели:

- BERT идеально подходит для задач, где требуется глубокое понимание контекста и точность в экстрактивном резюмировании;
- GPT-4 выделяется в абстрактивном резюмировании и суммаризации, предлагая высококачественную генерацию текста, хотя и с высокими требованиями к ресурсам;
- T5 обеспечивает выдающуюся гибкость и адаптируемость, что делает её подходящей для широкого спектра задач резюмирования/суммаризации и других задач NLP;
- BART эффективно справляется с абстрактивным резюмированием/суммаризацией благодаря своей способности к восстановлению и переформулировке текста.

Выбор подходящей модели зависит от специфических требований проекта, включая язык, на котором представлен контент, требуемую скорость обработки, доступные вычислительные ресурсы и предпочтения в стилях резюмирования/суммаризации.

Заключение

Использование нейросетевых алгоритмов в автоматическом реферировании и резюмировании текстов является перспективным направлением развития систем обработки информации. Нейросети открывают новые возможности для работы с большими объемами данных, повышают эффективность поиска и анализа информации, а также имеют широкий спектр практического применения. Существенно продвинуться в этой области позволяет использование таких технологий, как векторные представления слов (Word2Vec, GloVe), трансформаторы (BERT, GPT) и различные нейросетевые архитектуры. Рассмотренные модели способны обрабатывать и обобщать информацию, сохраняя при этом семантическую целостность и релевантность содержания, что особенно важно в областях, требующих быстрой обработки больших объемов данных, таких как новостные публикации или анализ научных текстов.

Из-за сложности языка, в том числе особенностей контекста, идиоматических выражений и культурных аллюзий, возникают проблемы с точной интерпретацией моделей и определением качественных показателей оценки. Традиционные метрики фокусируются на поверхностных характеристиках текста, таких как совпадение слов, которые могут не полностью отражать способность модели понимать или генерировать семантически и синтаксически корректный язык. Использование эталонных наборов данных для оценки может привести к предвзятости или ограничить область оценки, поскольку эти наборы данных могут не отражать всего многообразия употребления языка в реальном мире.

Перспективы нейросетевых алгоритмов в области автоматического реферирования весьма многообещающи. Непрерывный рост вычислительных мощностей, создание новых алгоритмов обучения и оптимизации моделей, а также улучшение предварительной обработки

и последующего анализа текстовых данных могут значительно повысить точность и адаптивность систем реферирования. Кроме того, применение технологий искусственного интеллекта для анализа эмоциональной окраски, контекста и стиля текста может открыть новые горизонты для разработки более совершенных и ориентированных на человека систем обработки естественного языка.

Список литературы

1. **Divakar Y., Jalpa D., Arun K. Y.** Automatic Text Summarization Methods: A Comprehensive Review. 2020. https://doi.org/10.48550/arXiv.2204.01849
2. **Salchner M. F., Adam A. A** Survey of Automatic Text Summarization Using Graph Neural Networks // In Proceedings of the 29th International Conference on Computational Linguistics. Gyeongju, Republic of Korea. International Committee on Computational Linguistics. 2022. P. 6139−6150.
3. **Vamvas J., Domhan T., Trenous S., Sennrich R., Hasler E.** Trained MT Metrics Learn to Cope with Machine-translated References // Conference: Proceedings of the Eighth Conference on Machine Translation. 2023. https://doi.org/10.18653/v1/2023.wmt-1.95.
4. **Mathur N., Baldwin T., Cohn T.** Tangled up in BLEU: Reevaluating the Evaluation of Automatic Machine Translation Evaluation Metrics // Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020. P 4984−4997.
5. **Reiter E.** A Structured Review of the Validity of BLEU // Computational Linguis-tics. 2018. № 44 (3). P. 393−401.
6. **Tianyi Z., Kishore V., Wu F., Weinberger K. Q., Artzi Y.** BERTScore: Evaluating Text Generation with BERT. ArXiv abs/1904.09675. 2019.
7. **Guo Y., Hu J.** Meteor++ 2.0: Adopt syntactic level paraphrase knowledge intomachine translation evaluation // In Proceedings of the Fourth Conference on Ma-chine Translation. 2019. Vol. 2. P. 501−506. https://doi.org/10.18653/v1/W19-5357
8. **Ayub S.A., Gaol F.L., Matsuo T.** A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. IEEE Access. 2021. № 9. P. 13248−13265. https://doi.org/10.1109/ACCESS.2021.3052783
9. **Al E. W., Awajan A. A.** SemG-TS: Abstractive Arabic Text Summarization Using Semantic Graph Embedding // Mathematics. 2022. № 10 (18). P. 3225. https://doi.org/10.3390/math10183225
10. **Tianyi Zhang T., Kishore V., Wu F., Weinberger K.Q., Artzi Y.** BERTSCORE: Evaluating Text Generation with BERT. Department of Computer Science and Cornell Tech, Cornell University. 2019.
11. **Saadany H., Orasan C.** BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text // Conference: TRITON (TRanslation and Interpreting Technology Online). 2021. https://doi.org/10.26615/978-954-452-071-7_006
12. **Sudoh K., Takahashi K., Nakamura S.** Is this translation error critical?: Classification-based human and automatic machine translation evaluation focusingon critical errors // Proceedings of the Workshop on Human Evaluation of NLPSystems (HumEval). 2021. P. 46−55.
13. **Siddhant A., Johnson M., Tsai T., Ari N.** Evaluating the Cross-Lingual Effectiveness of Massively Multilingual Neural Machine Translation // Proceedings of the AAAI Conference on Artificial Intelligence. 2020. № 34(05). P. 8854−8861. https://doi.org/10.1609/aaai.v34i05.6414.
14. **Lin J., Nogueira R., Yates A.** Pretrained Transformers for Text Ranking: BERT and Beyond // Synthesis Lectures on Human Language Technologies. 2021. № 14 (4). P. 1−325. https://doi.org/10.2200/S01123ED1V01Y202108HLT053.
15. **Chistyakova K., Kazakova T.** Grammar in Language Models: BERT Study // National research university higher school of economics. 2023. № 115.

References

1. **Divakar Y., Jalpa D., Arun K. Y.** Automatic Text Summarization Methods: A Comprehensive Review. 2020. https://doi.org/10.48550/arXiv.2204.01849
2. **Salchner M.F., Adam A. A** Survey of Automatic Text Summarization Using Graph Neural Networks. In: *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea. International Commit-tee on Computational Linguistics, 2022, pp. 6139−6150.
3. **Vamvas J., Domhan T., Trenous S., Sennrich R., Hasler E.** Trained MT Metrics Learn to Cope with Machine-translated References. C*onference: Proceedings of the Eighth Conference on Machine Translation*. 2023. https://doi.org/10.18653/v1/2023.wmt-1.95.
4. **Mathur N., Baldwin T., Cohn T.** Tangled up in BLEU: Reevaluating the Eval-uation of Automatic Machine Translation Evaluation Metrics. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 2020, pp. 4984−4997.
5. **Reiter E.** A Structured Review of the Validity of BLEU. *Computational Lin-guis-tics*, 2018, № 44 (3), pp. 393−401.
6. **Tianyi Z., Kishore V., Wu F., Weinberger K. Q., Artzi Y.** BERTScore: Evaluat-ing Text Generation with BERT. ArXiv abs/1904.09675. 2019.
7. **Guo Y., Hu J.** Meteor++ 2.0: Adopt syntactic level paraphrase knowledge intomachine translation evaluation. In: *Proceedings of the Fourth Conference on Ma-chine Translation*, 2019, vol. 2, pp. 501−506. https://doi.org/10.18653/v1/W19-5357
8. **Ayub S. A., Gaol F. L., Matsuo T.** A Survey of the State-of-the-Art Models in Neural Abstractive Text Summarization. *IEEE Access*, 2021, № 9, pp. 13248−13265. https://doi.org/10.1109/ACCESS.2021.3052783
9. **Al E. W., Awajan A. A.** SemG-TS: Abstractive Arabic Text Summarization Us-ing Semantic Graph Embedding. *Mathematics*, 2022, № 10 (18), p. 3225. https://doi.org/10.3390/math10183225
10. **Tianyi Zhang T., Kishore V., Wu F., Weinberger K. Q., Artzi Y.** BERTSCORE: Evaluating Text Generation with BERT. Department of Computer Science and Cornell Tech, Cornell University, 2019.
11. **Saadany H., Orasan C.** BLEU, METEOR, BERTScore: Evaluation of Metrics Performance in Assessing Critical Translation Errors in Sentiment-oriented Text. *Conference: TRITON (TRanslation and Interpreting Technology Online)*, 2021. https://doi.org/10.26615/978-954-452-071-7_006
12. **Sudoh K., Takahashi K., Nakamura S.** Is this translation error critical?: Classification-based human and automatic machine translation evaluation focus-ingon critical errors. In: *Proceedings of the Workshop on Human Evaluation of NLPSystems (HumEval)*, 2021, pp. 46−55.
13. **Siddhant A., Johnson M., Tsai T., Ari N.** Evaluating the Cross-Lingual Effectiveness of Massively Multilingual Neural Machine Translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, № 34 (5), pp. 8854−8861. https://doi.org/10.1609/aaai.v34i05.6414.
14. **Lin J., Nogueira R., Yates A.** Pretrained Transformers for Text Ranking: BERT and Beyond. *Synthesis Lectures on Human Language Technologies*, 2021, № 14 (4), pp. 1−325. https://doi org/10.2200/S01123ED1V01Y202108HLT053.
15. **Chistyakova K., Kazakova T.** Grammar in Language Models: BERT Study. *National research university higher school of economics*, 2023, № 115.

Сведения об авторах

Ребенок Кирилл Вячеславович, аспирант Московского финансово-юридического университет МФЮА

Information about the Author

Kirill V. Rebenok, Postgraduate Student of the Moscow University of Finance and Law MFUA, Moscow, Russian Federation

Статья поступила в редакцию 10.04.2024;
одобрена после рецензирования 07.11.2024; принята к публикации 07.11.2024

The article was submitted 10.04.2024;
approved after reviewing 07.11.2024; accepted for publication 07.11.2024

Научная статья

УДК 004.93
DOI 10.25205/1818-7900-2024-22-4-62-70

Выбор нейросетевой модели на основе метода анализа иерархий

Румиль Мухутдинович Хусаинов

Казанский национальный исследовательский технический университет им. А. Н. Туполева – КАИ,
Казань, Россия

rumil_husainov98@mail.ru, https://orcid.org/0000-0003-4985-7833

Аннотация

В статье рассматривается выбор оптимальной нейросетевой модели YOLOv8 (YOLOv8s, YOLOv8l, YOLOv8x, YOLOv8m, YOLOv8n) с использованием метода анализа иерархий, который позволяет структурировать и систематизировать сложные решения, основанные на многокритериальных оценках. Основное внимание уделяется выявлению и сравнительному анализу наиболее значимых критериев для оценки эффективности нейросетевых моделей, таких как время обучения, а также метрики Precision, Recall и F1-score. Эти метрики играют ключевую роль в задачах компьютерного зрения, особенно когда речь идет о детекции объектов. При проведении исследования построены матрицы попарных сравнений, которые позволяют не только визуально представить относительную важность каждого из выбранных параметров, но и количественно оценить их влияние на общую эффективность модели. Процесс формирования матриц попарных сравнений включает в себя мнение экспертов в области машинного обучения и компьютерного зрения, что обеспечивает высокую степень надежности полученных результатов. После обработки данных и выполнения расчетов, включающих взвешивание каждого критерия, выведены приоритеты для альтернативных моделей YOLOv8. В результате расчетов выявлено, что нейросетевая модель YOLOv8n обладает максимальным приоритетом среди всех оцененных альтернатив. Это подчеркивает ее превосходство по сравнению с другими модификациями данной модели.

Ключевые слова

нейросетевая модель, метод анализа иерархий, модификация, попарное сравнение, вектор матриц, метрика Precision, критерий подбора, вектор приоритетов

Для цитирования

Хусаинов Р. М. Выбор нейросетевой модели на основе метода анализа иерархий // Вестник НГУ. Серия: Информационные технологии. 2024. Т. 22, № 4. С. 62–70. DOI 10.25205/1818-7900-2024-22-4-62-70

Selection of a Neural Network Model Based on the Hierarchy Process Analysis Method

Rumil M. Khusainov

Kazan National Research Technical University named after A. N. Tupolev – KAI,
Kazan, Russian Federation

rumil_husainov98@mail.ru, https://orcid.org/0000-0003-4985-7833

Abstract

This article discusses the selection of the optimal neural network model YOLOv8 (YOLOv8s, YOLOv8l, YOLOv8x, YOLOv8m, YOLOv8n) using the hierarchy process analysis method, which allows structuring and systematizing com-

plex decisions based on multi-criteria assessments. The main focus is on identifying and comparative analysis of the most significant criteria for assessing the effectiveness of neural network models, such as training time, as well as the Precision, Recall and F1-score metrics. These metrics play a key role in computer vision tasks, especially when it comes to object detection. During the study, pairwise comparison matrices were constructed, which allow not only to visually represent the relative importance of each of the selected parameters, but also to quantitatively assess their impact on the overall effectiveness of the model. The process of forming pairwise comparison matrices includes the opinion of experts in the field of machine learning and computer vision, which ensures a high degree of reliability of the results. After processing the data and performing calculations, including weighting each criterion, priorities for alternative YOLOv8 models were derived. As a result of the calculations, it was revealed that the YOLOv8n neural network model has the highest priority among all the alternatives evaluated. This emphasizes its superiority compared to other modifications of this model.

Keywords

neural network model, hierarchy process analysis method, modification, pairwise comparison, matrix vector, Precision metric, selection criterion, priority vector

For citation

Khusainov R. M. Selection of a neural network model based on the hierarchy process analysis method. Vestnik NSU. Series: Information Technologies, 2024, vol. 22, no. 4, pp. 62–70 (in Russ.) DOI 10.25205/1818-7900-2024-22-4-62-70

Введение

Современные технологии компьютерного зрения становятся все более важными для решения широкого спектра задач в различных областях, таких как системы видеонаблюдения, медицинская диагностика и автомобильная промышленность. Одной из наиболее перспективных и широко используемых моделей для распознавания объектов является нейронная сеть YOLO (You Only Look Once). Последняя версия, YOLOv8, предлагает улучшенные алгоритмические подходы и архитектуру по сравнению с предыдущими версиями, что обеспечивает более высокую точность и скорость обработки.

Тем не менее выбор оптимальной нейросетевой модели для конкретной задачи может быть сложным и многофакторным процессом. Метод анализа иерархий представляет собой полезный инструмент для принятия решений, позволяющий структурировать задачу выбора на основе критических факторов и критериев. Это позволит не только учитывать качественные и количественные характеристики моделей, но и проводить многокритериальный анализ, что делает процесс выбора более обоснованным.

Расчет метода анализа иерархий

Для практического применения нейросетевой модели YOLOv8 необходимо выбрать наиболее оптимальную из пяти ее модификаций (YOLOv8n, YOLOv8s, YOLOv8m,

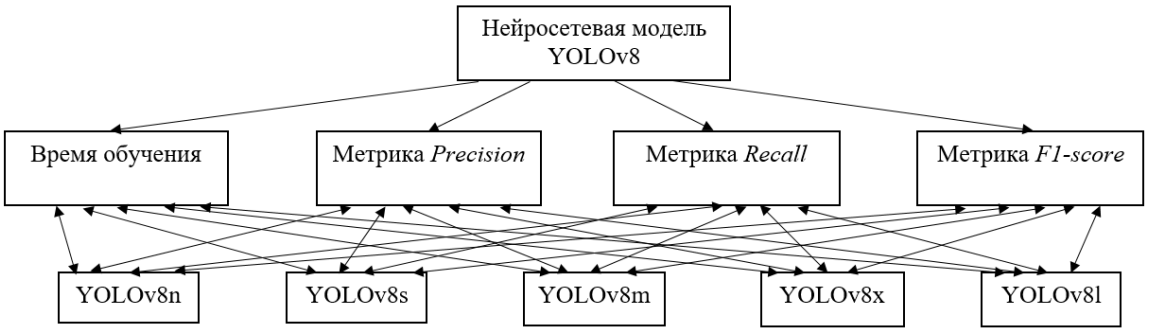


Рис. 1. Построение дерева альтернатив для выбора оптимальной модели YOLOv8
Fig. 1. Construction of a tree of alternatives for choosing the optimal YOLOv8 model

YOLOv8x, YOLOv8l). Поэтому для принятия решений предлагается использовать математический инструмент на основе метода анализа иерархий (метод попарных сравнений) [1–3]. Данный метод заключается в сравнении изучаемых факторов (критериев, альтернатив) между собой.

Нейросетевая модель YOLOv8 (рис. 1) включает 4 критерия (время обучения, метрика Precision, метрика Recall, метрика F1-score) и 5 модификаций модели (YOLOv8n, YOLOv8s, YOLOv8m, YOLOv8l, YOLOv8x) [4–5].

После обучения нейросетевой модели YOLOv8 получены следующие результаты:

- время обучения модели (YOLOv8n − 1 ч 21 мин, YOLOv8s – 1 ч 47 мин, YOLOv8m − 1 ч 15 мин, YOLOv8l – 1 ч 26 мин, YOLOv8x − 1 ч 32 мин);
- метрика Precision (YOLOv8n – 0,845; YOLOv8s – 0,703; YOLOv8m – 0,705; YOLOv8l – 0,845; YOLOv8x - 0,649);
- метрика Recall (YOLOv8n – 0,409; YOLOv8s – 0,449; YOLOv8m – 0,51; YOLOv8l – 0,426; YOLOv8x - 0,421);
- метрика F1-score (YOLOv8n – 0,52; YOLOv8s – 0,519; YOLOv8m – 0,58; YOLOv8l – 0,566; YOLOv8x − 0,491).

Для оценки критерия модели YOLOv8 построим матрицу попарных сравнений (A_{ji}). Пусть $a_{ii} = 1$, тогда отношение критерия *i* к критерию *j* определяется (1):

$$a_{ji} = \frac{1}{a_{ij}} \tag{1}$$

Попарное сравнение (табл. 1) оценивается по шкале интенсивности от 1 до 9 (1 – равно, 3 – немного лучше, 5 – лучше, 7 – значительно лучше, 9 – принципиально лучше).

При построении каждой из матриц важно обеспечить объективность оценок, используя согласованность мнений среди экспертов или автоматизированные методы. В результате получим наглядное представление о том, какие модели показывают лучшие результаты в разных аспектах, это позволит принять более обоснованное решение при выборе наиболее подходящей модели для использования в конкретных условиях.

Таблица 1

Результаты значений матрицы попарных сравнений

Table 1

Results of the values of the pairwise comparison matrix

Критерий подбора	Метрика Precision	Метрика Recall	Метрика F1-score	Время обучения
Метрика Precision	1	6	7	5
Метрика Recall	0,166666667	1	6	7
Метрика F1-score	0,142857143	0,166666667	1	9
Время обучения	0,2	0,142857143	0,111111111	1
Итого	1,50952381	7,30952381	14,11111111	22

Для вычисления вектора приоритетов выбран метод автора Т. Саати – «Метод анализа иерархий», разработанный как технология принятия решений на основе математических расчетов с применением метода попарных сравнений, используемый для определения вектора *w* [1]. Метод основан на одном из принципов линейной алгебры: искомый вектор является собственным вектором матрицы попарных сравнений, который соответствует максимальному числу λ_{max} [6–8].

Используем нормализованную оценку для *j*-го фактора $\hat{k}_j$, где *j* – обозначение фактора по строке в матрице попарных сравнений.

Собственные векторы матриц вычисляются на основе приближенных значений столбцов, используя метод среднегеометрического измерения расстояний между рассматриваемыми факторами (2):

$$\hat{k}_{геом} = \sqrt[n]{k_1 \cdot k_2 \cdot \ldots \cdot k_n}, \quad (2)$$

где *n* – количество оцениваемых факторов (критериев, альтернатив) [9–11].

Нормализованные оценки вектора приоритета вычисляется по формуле (3):

$$\hat{k}_j = \frac{\hat{k}_{геомj}}{\sum_{j=1}^{n} \hat{k}_{геомj}}. \quad (3)$$

Матрица формируется на основе парных сравнений критериев, где элементы матрицы представляют собой относительные оценки важности одного элемента относительно другого (табл. 2).

Таблица 2

Результаты значений матрицы по шкале Т. Саати

Table 2

Results of matrix values according to T. Saaty scale

Критерий подбора	Метрика Precision	Метрика Recall	Метрика F1-score	Время обучения	Вектор приоритетов $\hat{k}_{геом}$	Нормализованные оценки вектора приоритета $\hat{k}_j$
Метрика Precision	1	6	7	5	3,806754096	0,599387486
Метрика Recall	0,166666667	1	6	7	1,626576562	0,256110485
Метрика F1-score	0,142857143	0,166666667	1	9	0,680374933	0,107127545
Время обучения	0,2	0,142857143	0,111111111	1	0,237368104	0,037374484
Итого	1,50952381	7,30952381	14,11111111	22	6,351073695	1

Максимальное значение вектора матрицы попарных сравнений определяется по формуле (4):

$$\lambda_{\max} = \sum_{j=1}^{n} S_j \cdot \hat{k}_j, \quad (4)$$

где S_j – сумма j-го столбца [12].

Далее вычисляется индекс согласованности (*ИС*) по формуле (5):

$$ИС = \frac{\lambda_{\max} - n}{n-1} \quad (5)$$

Отношение согласованности (*ОС*) определяется по формуле (6):

$$ОС = \frac{ИС}{СИ \cdot 100}, \quad (6)$$

где *СИ* – среднее значение случайного индекса согласованности [13].

Проверка условия приемлемости *ОС* определяется по выражению (7):

$$ОС \leq 10\%. \quad (7)$$

Значение *ОС* считается для матрицы попарных сравнений приемлемым.

Построим матрицы попарных сравнений (табл. 3) по отдельным критериям (метрика Precision, метрика Recall, метрика F1-score, время обучения).

Таблица 3

Результаты значений матрицы попарных сравнений
для критерия «Метрика Precision»

Table 3

Results of the values of the matrix of pairwise comparisons
for the criterion «Precision Metric»

Метрика Precision	YOLOv8n	YOLOv8s	YOLOv8m	YOLOv8l	YOLOv8x
YOLOv8n	1	5	7	9	4
YOLOv8s	0,2	1	0,111111111	0,142857143	0,2
YOLOv8m	0,142857143	9	1	0,2	0,333333333
YOLOv8l	0,111111111	7	5	1	3
YOLOv8x	0,25	5	3	0,333333333	1
Итого	1,703968254	27	16,11111111	10,67619048	8,533333333

Далее вычисляются векторы приоритетов и нормализованные оценки вектора приоритетов по формуле 2, 3 (табл. 4). Вектор приоритетов отражает относительную важность каждой альтернативы по критерию Precision. Нормализованные оценки упрощают интерпретацию полученных результатов и помогают идентифицировать наиболее предпочтительные альтернативы.

Определим максимальное число вектора матрицы попарных сравнений λ_{max}:

$$\lambda_{max} = 1{,}70397 \cdot 0{,}54213 + 27 \cdot 0{,}02982 + 16{,}11111 \cdot 0{,}07955 + 10{,}72619 \cdot 0{,}21253 + 8{,}31111 \cdot 0{,}13596 = 6{,}42016773$$

$$ИС = \frac{6{,}42016773 - 5}{5 - 1} = 0{,}35504193$$

$$ОС = \frac{0{,}35504193}{1{,}12 \cdot 100} = 0{,}00317002$$

Матрица попарных сравнений для критерия «Метрика Precision» по ОС является согласованной. Аналогично проводятся расчеты для других критериев.

Последним этапом для выбора модификации нейросетевой модели YOLOv8 является синтез альтернатив. Векторы приоритетов определяются для всех построенных матриц попарных сравнений.

Приоритеты альтернатив q определяются по формуле (8):

$$q = \hat{k}_{jk} \cdot \hat{k}_j, \quad (8)$$

где $\hat{k}_{jk}$ – нормализованные оценки вектора приоритета по матрице критериев.

Таблица 4

Результаты значений матрицы попарных сравнений для критерия «Метрика Precision» по шкале Т. Саати

Table 4

Results of the values of the matrix of pairwise comparisons for the criterion «Precision Metric» according to the scale of T. Saaty

Метрика Precision	YOLOv8n	YOLOv8s	YOLOv8m	YOLOv8l	YOLOv8x	Вектор приоритетов $\hat{k}_{геом}$	Нормализованные оценки вектора приоритета $\hat{k}_j$
YOLOv8n	1	5	7	9	4	4,16941	0,54213
YOLOv8s	0,2	1	0,11111	0,14286	0,2	0,22937	0,02982
YOLOv8m	0,14286	9	1	0,2	0,33333	0,6118	0,07955
YOLOv8l	0,11111	7	5	1	3	1,63452	0,21253
YOLOv8x	0,25	5	3	0,33333	1	1,04564	0,13596
Итого	1,70397	27	16,11111	10,67619	8,53333	7,69074	1

Расчеты векторов приоритетов (табл. 5) содержат нормализованные значения оценки вектора приоритета $\hat{k}_j$ в каждом столбце. Каждый элемент вектора приоритета соответствует оценке конкретной альтернативы по определенному критерию, что в дальнейшем используется для принятия обоснованных решений. Важно отметить, что нормализованные значения векторов приоритетов могут изменяться в зависимости от выбранных критериев и методов оценивания.

Таблица 5

Результаты расчетов векторов приоритетов для альтернатив

Table 5

Results of calculations of priority vectors for alternatives

Альтернативы	Критерий				Приоритеты альтернатив
	Метрика Precision	Метрика Recall	Метрика F1-score	Время обучения	
	Численное значение приоритета				
	0,599387486	0,256110485	0,107127545	0,037374484	
YOLOv8n	0,542133382	0,556102472	0,551514047	0,510038725	0,545516419
YOLOv8s	0,029824689	0,027114539	0,024555448	0,032917772	0,028681713
YOLOv8m	0,079550458	0,06424326	0,066621028	0,063636045	0,073650233
YOLOv8l	0,212530566	0,251458918	0,223314595	0,263833779	0,225573223
YOLOv8x	0,135960904	0,101080811	0,133994883	0,129573679	0,126578412

Как видно из таблицы, максимальное значение соответствует альтернативе YOLOv8n и составляет 0,545516419.

После проведенных исследований строится диаграмма нейросетевой модели YOLOv8 с указанием приоритетов альтернатив.

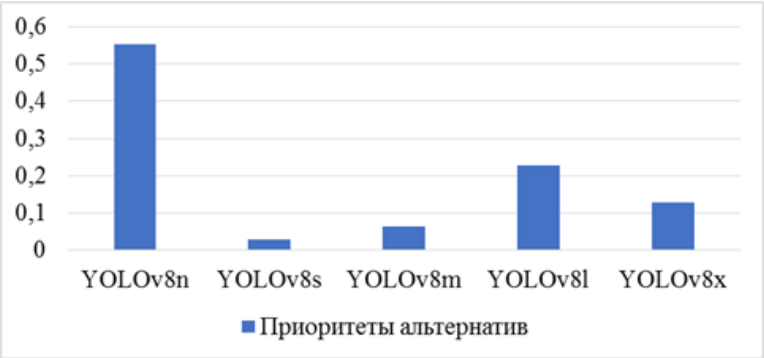


Рис. 2. Результаты расчета приоритета альтернатив для выбора оптимальной модели YOLOv8
Fig. 2. Results of calculating the priority of alternatives for choosing the optimal YOLOv8 model

Следовательно, нейросетевая модель YOLOv8n обеспечивает оптимальные показатели (рис. 2), и в дальнейшем может использоваться при разработке программного комплекса и при распознавании объектов дорожной инфраструктуры по классам.

Заключение

Таким образом, выбор оптимальной нейросетевой модели YOLOv8 с использованием метода анализа иерархий показывает важность многокритериального подхода в процессе принятия решений в области компьютерного зрения. Анализ ключевых критериев, таких как время обучения, а также метрики Precision, Recall и F1-score, позволяет более глубоко оценить эффективность различных модификаций модели YOLOv8 (YOLOv8s, YOLOv8l, YOLOv8x, YOLOv8m, YOLOv8n).

Методы, использованные для построения матриц попарных сравнений и вовлечение экспертов в процесс оценки, подтверждают надежность и обоснованность результатов.

Полученные данные показывают, что YOLOv8n обладает наибольшим приоритетом, что может быть связано с ее оптимизированной архитектурой или качеством распознавания объектов.

Список литературы

1. **Саати Т.** Принятие решений. Метод анализа иерархий. М.: Радио и связь, 1989. 316 с.
2. **Скрипина И. И.** Анализ и выбор математической модели с помощью метода анализа иерархий // Научный результат. Информационные технологии. 2021. Т. 6, № 2. С. 41–46. DOI: 10.18413/2518-1092-2021-6-2-0-6.
3. **Ибрагимова З. А.** Сравнительный анализ межсетевых экранов следующего поколения на основе метода анализа иерархий // Информационные технологии и математические методы в экономике и управлении. 2023. С. 82–89.
4. **Хусаинов Р. М., Талипов Н. Г., Катасев А. С, Шалаева Д. В.** Нейросетевая технология анализа транспортных потоков в автоматизированных системах управления дорожным движением // Программная инженерия. 2023. Т. 14, № 10. С. 513–519.
5. **Хусаинов Р. М., Талипов Н. Г., Катасев А. С.** Нейросетевая модель и программный комплекс распознавания объектов дорожной инфраструктуры // Информационные технологии. 2023. Т. 29, № 9. С. 484–491.
6. **Никул Е. М., Сидоров С. С.** Алгоритм анализа матриц попарных сравнений с помощью вычисления векторов приоритетов // Известия ЮФУ. Технические науки. 2012. № 2 (127). С. 241–247.
7. **Holovko S.** Analysis of non-rigid pavement design options using the hierarchy method // Dorogi i mosti. 2022. No. 25. P. 31–39.
8. **Мошенко И. Н., Пирогов Е. В.** Метод факторного анализа иерархий // Инженерный вестник Дона. 2017. № 4 (47). С. 144.
9. **Ибрагимова З. А.** Сравнительный анализ межсетевых экранов следующего поколения на основе метода анализа иерархий // Информационные технологии и математические методы в экономике и управлении. 2023. С. 82–89.
10. **Лубенцов А. В., Кобзистый С. Ю.** Системный анализ параметров сложной системы с применением каскадного метода анализа иерархий // Техника и безопасность объектов уголовно-исполнительной системы: сб. материалов Междунар. науч.-практ. конф. Воронеж, 2023. С. 187–191.
11. **Kyrylych T., Povstenko Yu.** Multi-criteria analysis of startup investment alternatives using the hierarchy method // Entropy. 2023. Vol. 25. No. 5. P. 723.
12. **Кротова А. В., Дрогалов Д. А., Солдатов Е. С.** Сравнительный анализ методологий управления IT-проектами при помощи метода анализа иерархий // Научный форум.: сб. ст. V Междунар. науч.-практ. конф. Пенза, 2023. С. 75–79.
13. **Романова П. С., Романова И. П.** Применение метода анализа иерархий для оптимизации выбора кровельного материала для скатной крыши // Инженерный вестник Дона. 2018. № 4 (51). С. 85.

References

1. **Saati T.** Decision Making. The Analytic Hierarchy Process. Moscow, Radio and Communications publ., 1989, 316 p.
2. **Skripina I. I.** Analysis and Selection of a Mathematical Model Using the Analytic Hierarchy Process. *Scientific Result. Information Technology*, 2021, vol. 6. no. 2, pp. 41−46. (in Russ.) DOI: 10.18413/2518-1092-2021-6-2-0-6.
3. **Ibragimova Z. A.** Comparative Analysis of Next-Generation Firewalls Based on the Analytic Hierarchy Process. In: *Information Technology and Mathematical Methods in Economics and Management*, 2023, pp. 82−89. (in Russ.)
4. **Khusainov R. M., Talipov N. G., Katasev A. S., Shalaeva D. V.** Neural Network Technology for Analyzing Traffic Flows in Automated Traffic Control Systems. *Software Engineering*, 2023, vol. 14, no. 10, pp. 513−519. (in Russ.)
5. **Khusainov R. M., Talipov N. G., Katasev A. S.** Neural network model and software package for recognizing road infrastructure objects. *Information technologies*, 2023, vol. 29, no. 9, pp. 484−491. (in Russ.)
6. **Nikul E. M., Sidorov S. S.** Algorithm for analyzing pairwise comparison matrices using priority vector calculation. *Bulletin of SFU. Technical sciences*, 2012, no. 2 (127), pp. 241−247. (in Russ.)
7. **Holovko S.** Analysis of non-rigid pavement design options using the hierarchy method. *Dorogi i mosti*, 2022, no. 25, pp. 31−39.
8. **Moshenko I. N., Pirogov E. V.** Method of factor analysis of hierarchies. *Engineering Bulletin of the Don*, 2017, no. 4 (47), pp. 144. (in Russ.)
9. **Ibragimova Z. A.** Comparative Analysis of Next-Generation Firewalls Based on the Analytic Hierarchy Process. In: *Information Technology and Mathematical Methods in Economics and Management*, 2023, pp. 82−89. (in Russ.)
10. **Lubentsov A. V., Kobzisty S. Yu.** Systems Analysis of Parameters of a Complex System Using the Cascade Analytic Hierarchy Process. In: *Technology and Safety of Penal System Facilities. Collection of Materials of the International Scientific and Practical Conference*. Voronezh, 2023, pp. 187−191. (in Russ.)
11. **Kyrylych T., Povstenko Yu.** Multi-criteria Analysis of Startup Investment Alternatives Using the Hierarchy Method. *Entropy*, 2023, vol. 25, no. 5, p. 723.
12. **Krotova A. V., Drogalov D. A., Soldatov E. S.** Comparative analysis of IT project management methodologies using the hierarchy analysis method. In: *Scientific forum. collection of articles of the V International scientific and practical conference*. Penza, 2023, pp. 75−79. (in Russ.)
13. **Romanova P. S., Romanova I. P.** Application of the hierarchy process analysis method to optimize the choice of roofing material for a pitched roof. *Engineering Bulletin of the Don*, 2018, no. 4 (51), p. 85. (in Russ.)

Сведения об авторах

Хусаинов Румиль Мухутдинович, аспирант
SPIN-код: 1247-6906
Author ID: 1160304

Information about the Author

Khusainov Rumil Mukhutdinovich, Postgraduate student
SPIN code: 1247-6906
Author ID: 1160304

Статья поступила в редакцию 27.09.2024;
одобрена после рецензирования 14.01.2025; принята к публикации 14.01.2024

The article was submitted 27.09.2024;
approved after reviewing 14.01.2025; accepted for publication 14.01.2025

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 язык**е. 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 язык**е, 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 «p» (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., Окольнишников 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.