ОГЛАВЛЕНИЕ

УДК 004.62

О ПОДХОДЕ К АВТОМАТИЗАЦИИ ОЦЕНКИ ЗНАНИЙ В ОБЛАСТИ РАЗРАБОТКИ ПРОГРАММНОГО ОБЕСПЕЧЕНИЯ НА ОСНОВЕ АНАЛИЗА ДАННЫХ ПРОЕКТНОЙ РАБОТЫ

А. А. Атнагулов[1] [0000-0001-9766-4804], М. М. Абрамский[2] [0000-0003-3063-8948]

[1, 2]*Институт информационных технологий и интеллектуальных систем, Казанский (Приволжский) федеральный университет.*

[1]i@atnartur.dev, [2]mabramsk@kpfu.ru

Аннотация

Проектный подход широко используется в организации подготовки ИТ-специалистов в вузах. Несмотря на то, что организуемые процессы разработки крайне близки к процессам, применяемым в ИТ-компаниях, анализ процесса разработки студенческих проектов практически не интегрирован в систему оценивания студентов, а также в большинстве случаев выполняется в ручном режиме.

В статье предложен подход к выстраиванию аналитики процесса разработки студенческого проекта, а также рассмотрены варианты использования результатов аналитики в оценке работы студентов.

Ключевые слова*: проектная работа, обучение ИТ-специалистов, разработка программного обеспечения, оценка образовательного результата, данные проекта*

ВВЕДЕНИЕ

Применение проектной деятельности в высшем учебном заведении является эффективным способом развития компетенций студентов [1]. Использование проектного подхода для обучения будущих специалистов IT-сферы позволяет приблизить процесс обучения к работе над реальным проектом [2, 3]. Также проектный подход помогает перейти от передачи готовых знаний к самостоятельному поиску знаний и достижению результата, что приближает образовательную среду к индустриальной [4, 5].

Для анализа процесса проектной работы существует ряд практик и инстру-

ментов. Аналитика осуществляется с помощью собственных разработок компаний, а также существующих систем управления процессами (Gitlab[1], Jira[2] и другие) [6–8]. В то же время стоит отметить, что названные инструменты не используются для анализа процесса разработки обучающего проекта с точки зрения качества кода и персональной эффективности обучающегося в рамках проекта.

В статье предложен подход к анализу студенческой командной проектной работы, а также рассмотрена реализация предложенного подхода и оценено влияние ее внедрения на процесс обучения.

АНАЛИЗ ПРОЕКТНОЙ РАБОТЫ В КОММЕРЧЕСКИХ КОМПАНИЯХ

Существует целый ряд метрик оценки процесса разработки. Выбор метрик зависит от компании и проекта, а их сбор осуществляется с помощью собственных разработок и существующих систем управления процессами (Gitlab, Jira и других). Эти инструменты строят различные отчеты, которые могут помочь понять, что происходит с процессом разработки [6–8]. Ниже приведен краткий перечень информации, которая может быть представлена в подобных инструментах.

- Аналитика в Gitlab:
    - Количество коммитов по дням с распределением по команде, дням недели, часам.
    - Количество успешных и проваленных сборок в непрерывной интеграции (Continuous Integration).
- Аналитика в Jira:
    - Отчет о количестве сделанной и оставшейся работы (burndown chart).
    - Отчет об изменении оценок времени выполнения задач (story points).
- Метрики сервиса Waydev [6]:
    - Графики по количеству событий в команде и разным типам (изменения в коде, запросы на слияние, изменения в задачах).
- Метрики от команды сервиса GitLean [7]:

[1] Gitlab – веб-сервис для управления репозиториями и задачами, а также организации автоматической сборки проекта (https://gitlab.com).

[2] Jira – система для управления задачами от компании Atlassian (https://www.atlassian.com/software/jira).

- Churn кода – напрасно написанный код (количество строчек кода, удаленных до написанных во время выполнения задачи и удаленных в итоговой реализации задачи).
- Количество багов.

Несмотря на широкое разнообразие и применение инструментов проектной аналитики в коммерческих проектах, в образовательных проектах такие инструменты применяются редко. Однако такие инструменты могли бы помочь быстрее выявлять проблемы процесса разработки, а также производить более качественную оценку работы студентов в команде, так как вручную подобная аналитика не может быть произведена на достаточном уровне качества [1, 5]. В рамках работы было принято решение разработать инструмент для анализа проектной работы с учетом специфики работы студенческих команд.

ОРГАНИЗАЦИЯ РАБОТЫ СТУДЕНЧЕСКИХ КОМАНД

Разрабатываемый инструмент анализа проектной работы планируется к применению в учебно-практической лаборатории, в которой студенты объединяются в команды до 5 человек для обучения разработке ИТ-проектов.

У каждой команды есть куратор, являющийся сотрудником института либо студентом старшего курса, основными задачами которого являются консультирование по организации процесса проектной работы, помощь в принятии организационных и технологических решений, а также в организации встреч с заказчиками и заинтересованными лицами. У студентов есть возможность получить консультацию по используемым технологиям у преподавателя.

Работа команд разделяется на спринты – недельные промежутки времени, в рамках которых нужно выполнить запланированный объем задач. В начале спринта происходит выбор задач для выполнения за спринт, к концу спринта все запланированные работы должны быть закончены.

Задачи по разработке проекта описываются в Trello[3], статус которых изменяется по следующему принципу:

[3] Сервис управления задачами. Сайт: https://trello.com.

- Backlog – этот статус получают задачи, планирующиеся к выполнению в будущем;
- ToDo – задачи этого статуса будут выполнены в текущем спринте;
- Doing – задачи, выполняющиеся в данный момент;
- Test – задачи, готовые для проверки;
- Done – выполненные задачи.

Код проектов хранится в Gitlab. Каждая задача разрабатывается в отдельной ветке. Происходит автоматическое развертывание тестовой версии проекта на сервер, чтобы участники команды и заинтересованные лица могли увидеть текущее состояние работы проекта.

Обсуждения производятся в Telegram, Microsoft Teams и в очном формате.

Информация из перечисленных сервисов будет анализироваться в инструменте анализа проектной работы.

СБОР ИНФОРМАЦИИ О РАБОТЕ КОМАНД

Инструмент анализа проектной работы собирает данные по проекту, корректирует и сопоставляет данные из различных сервисов, позволяет контролировать общие задачи по всем проектам. Участники процесса вносят следующую информацию (см. рис. 1). Ниже эти процессы описаны подробнее.

Рис. 1. Диаграмма взаимодействия с инструментом анализа проектной работы

Преподаватель вносит следующую информацию о проектах:

- название и описание проекта, ссылки на документы;
- состав команды проекта с учетом мнения студентов;
- список заинтересованных лиц;
- устанавливает связи с чатом проекта в Telegram, репозиториями в Gitlab, доской с задачами в Trello;
- устанавливает соответствие между именами пользователей в сервисах и карточкой студента, чтобы данные о работе студента в различных сервисах отображались в объединенном формате.

Информация, собираемая с сервисов проектной работы, разделяется на события. Некоторые из них представлены ниже.

- Trello: создание и закрытие задачи, изменение статуса, добавление комментария, изменения чеклиста;
- Gitlab: отправка кода, развертывание проекта, комментарии в code review[4];
- Telegram: сообщение в чате команды.

Процесс сбора информации запускается каждые 5 минут, после этого информация появляется в отчетах.

ПОСТРОЕНИЕ ОТЧЕТОВ

Для отображения сводной информации по всем проектам был построен отчет о датах последнего обновления информации о проекте и задачах, а также дате последних событий в сервисах (см. рис. 2). По клику на ячейку отчета можно перейти к соответствующему сервису. Эмпирическим путем были выбраны граничные значения для отображения статусов, представленные в таблице 1.

Таблица 1. Статусы обновления данных

Статус данные	**Дата последнего обновления**	**Цвет**
актуальные	до 7 дней назад	зеленый
устаревающие	от 7 до 14 дней назад	желтый
неактуальные	более 14 дней назад	красный

[4] Процесс проверки исходного кода программы.

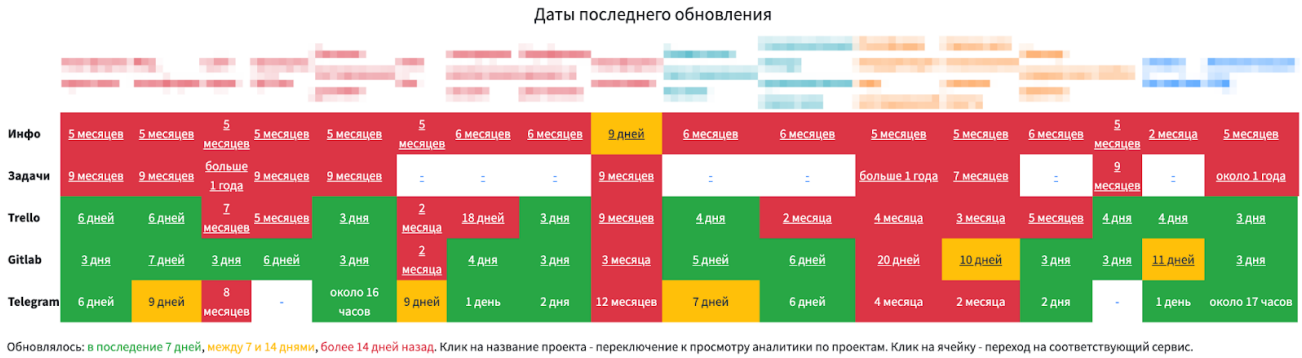

Рис. 2. Отчет о дате последнего обновления

Преподаватель оценивает проекты на еженедельной основе. Чтобы упростить эту задачу, в инструменте были построены графики о количестве событий, происходящих каждую неделю. Есть возможность посмотреть график по проектам (рис. 3), членам команды (рис. 4) и сервисам (рис. 5). Дополнительно доступны группировка по дням и фильтрация событий, сделанных определенным студентом.

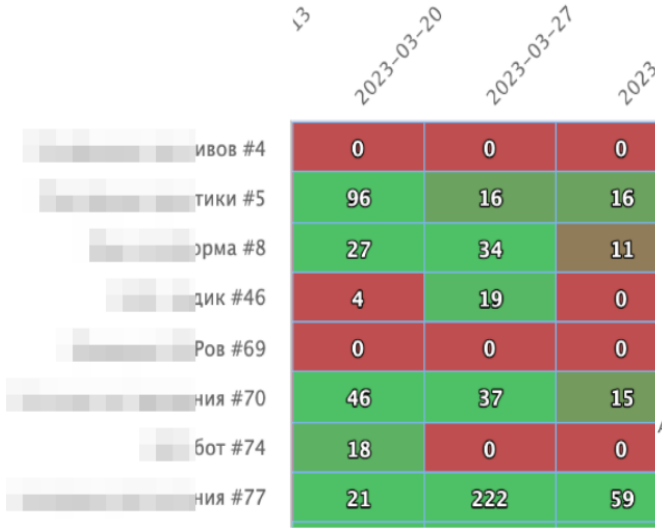

Рис. 3. График событий в проектах по неделям

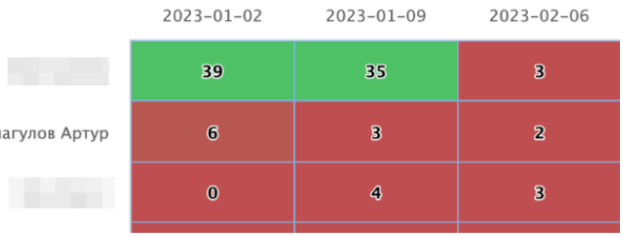

Рис. 4. График событий по членам команды проекта

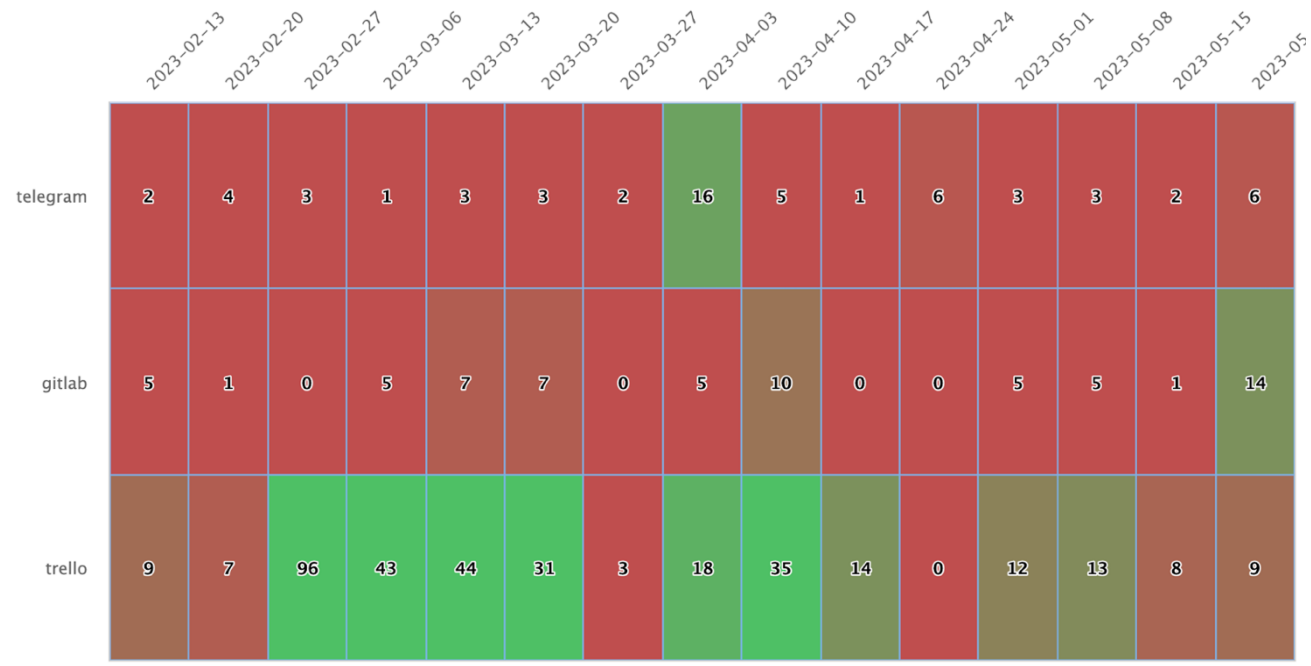

Рис. 5. График событий проекта по сервисам

Инструмент анализа проектной работы не выставляет итоговых оценок, но помогает преподавателю принять решение об оценке на основе данных проектной работы.

ТЕХНИЧЕСКАЯ РЕАЛИЗАЦИЯ

Техническая реализация инструмента анализа проектной работы выглядит следующим образом:

- данные о студентах, проектах и общих задачах проекта находятся в PostgreSQL, в Redis размещены очереди сообщений, в Clickhouse – информация, полученная из сервисов проектной работы;
- визуальная часть реализована с помощью Vue.js, для визуализации данных использована библиотека Highcharts;
- серверная часть разработана с помощью веб-фреймворка Django, Celery использован для выполнения задач очереди сообщений;
- развертывание проведено с помощью средств Docker и Ansible.

ЗАКЛЮЧЕНИЕ

Разработаны подходы и инструмент анализа проектной работы, позволяющие преподавателю выставлять оценки на основе собранных данных о работе студентов. За 2 года использования инструмента был обработан следующий объем данных:

- 46 проектов, из них 18 находятся в активной разработке, разработка 13 проектов завершена;
- 144 студента участвовали в разработке проекта;
- было обработано более 52000 событий, из них 16 000 – из Telegram, 27 000 – из Trello, 9218 – из Gitlab;
- данные собирались из 53 репозиториев в Gitlab, 43 чатов в Telegram и 31 доски в Trello.

Опыт использования информации из инструмента анализа проектной работы показал, что данные об активности студентов за неделю напрямую коррелируют с результатом работы команды за спринт. Замечания о том, что в определенную неделю было проведено недостаточно работы, мотивирует студентов в дальнейшем ответственно относиться к выполнению задания.

СПИСОК ЛИТЕРАТУРЫ

1. *Гергерт Д.В., Артемьев Д.И.* Практика внедрения проектно-ориентированного обучения в вузе // Университетское управление: практика и анализ. 2019. №4. URL: https://cyberleninka.ru/article/n/praktika-vnedreniya-proektno-orientirovannogo-obucheniya-v-vuze (дата обращения: 16.06.2023).

2. *Назаренко Н.В.* Проектная деятельность как средство формирования профессиональных компетенций ИТ-специалиста //Молодежь XXI века: Шаг в будущее. 2019. С. 237–238.

3. *Евстратова Л.А., Исаева Н.В., Лешуков О.В.* Проектное обучение: практики внедрения в университетах. Москва, 2018. 153 с.

4. *Мишин И.Н.* Реализация проектной деятельности в системе студентоориентированного обучения // Высшее образование в России. 2022. №3. URL: https://cyberleninka.ru/article/n/realizatsiya-proektnoy-deyatelnosti-v-sisteme-studentotsentrirovannogo-obucheniya (дата обращения: 16.06.2023).

5. What is a code review? [Электронный ресурс] // about.gitlab.com.

URL: https://about.gitlab.com/topics/version-control/what-is-code-review/ (дата обращения: 16.06.2023).

6. Как измерить и оценить производительность разработчиков [Электронный ресурс] // habr.com.
URL: https://habr.com/ru/companies/otus/articles/500282/ (дата обращения: 16.06.2023).

7. Оцениваем разработчика на основе объективных данных [Электронный ресурс] // habr.com.
URL: https://habr.com/ru/companies/oleg-bunin/articles/417411/ (дата обращения: 16.06.2023).

8. Культура разработки: как оценивают производительность и эффективность [Электронный ресурс] // habr.com.
URL: https://habr.com/ru/companies/vk/articles/481930/ (дата обращения: 16.06.2023).

STUDENT TEAMS PROJECT WORK ANALYSIS TOOL DEVELOPMENT

A. A. Atnagulov[1] [0000-0001-9766-4804], M. M. Abramskiy[2] [0000-0003-3063-8948]

[1, 2] *Institute of Information Technology and Intelligent Systems of Kazan Federal University*

[1]i@atnartur.dev, [2]mabramsk@kpfu.ru

Abstract

The project approach is widely used in organizing the training of IT specialists in universities. Despite the fact that the organized development processes are extremely close to the processes used in commercial IT companies, the analysis of the development process of student projects is not carried out. This article proposes an approach for performing analytics of the development process in a student project, and also discusses options for using the results of analytics in assessing student work.

Keywords*: project work, IT specialists training, project development, result assessment*

REFERENCES

1. *Gergert D.V., Artemyev D.I.* The practice of introducing project-oriented education at a university // University management: practice and analysis. 2019. No. 4. URL: https://cyberleninka.ru/article/n/praktika-vnedreniya-proektno-orientirovannogo-obucheniya-v-vuze (date of access: 16.06.2023).

2. *Nazarenko N.V.* Project activity as a means of developing professional competencies of an IT specialist // Youth of the 21st century: Step into the future. 2019. P. 237–238.

3. *Evstratova L.A., Isaeva N.V., Leshukov O.V.* Project-based learning: implementation practices in universities. Moscow, 2018. 153 p.

4. *Mishin I.N.* Implementation of project activities in the system of student-oriented education // Higher education in Russia. 2022. No. 3.
URL: https://cyberleninka.ru/article/n/realizatsiya-proektnoy-deyatelnosti-v-sisteme-studentotsentrirovannogo-obucheniya (date of access: 16.06.2023).

5. What is a code review? [Electronic resource] // about.gitlab.com.
URL: https://about.gitlab.com/topics/version-control/what-is-code-review/ (date of access: 16.06.2023).

6. How to measure and evaluate developer productivity [Electronic resource] // habr.com. URL: https://habr.com/ru/companies/otus/articles/500282/ (date of access: 16.06.2023).

7. Evaluating the developer based on objective data [Electronic resource] // habr.com. URL: https://habr.com/ru/companies/oleg-bunin/articles/417411/ (date of access: 16.06.2023).

8. Development culture: how productivity and efficiency are measured [Electronic resource] // habr.com.
URL: https://habr.com/ru/companies/vk/articles/481930/ (date of access: 16.06.2023).

СВЕДЕНИЯ ОБ АВТОРАХ

АТНАГУЛОВ Артур Александрович – студент магистратуры Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Artur Aleksandrovich ATNATULOV – Master's student at the Institute of Information Technology and Intelligent Systems, Kazan Federal University.

email: i@atnartur.dev

ORCID: 0000-0001-9766-4804

АБРАМСКИЙ Михаил Михайлович – директор Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, кандидат технических наук.

Mikhail Mikhailovich ABRAMSKIY – director of the Institute of Information Technology and Intelligent Systems, Kazan Federal University, PhD (Cand Sci. – Tech.)

email: mabramsk@kpfu.ru

ORCID: 0000-0003-3063-8948

Материал поступил в редакцию 3 августа 2023 года

УДК 004.550

РАЗРАБОТКА СИСТЕМЫ УПРАВЛЕНИЯ САЙТАМИ ДЛЯ СОЗДАНИЯ НАУЧНО-ПОПУЛЯРНЫХ ПОРТАЛОВ ПО ГЕОЛОГИИ (НА ПРИМЕРЕ ПОРТАЛА «ИСТОРИЯ ЗЕМЛИ: ГЕОЛОГИЧЕСКИЙ РАКУРС»)

А. С. Еременко[1, 2, 3] [0000-0003-1923-8417], **М. В. Гринёв**[3] [0009-0007-9611-6946],

Е. А. Одновил[4] [0009-0007-2368-4379]

[1] Государственный геологический музей им. В. И. Вернадского РАН, г. Москва

[2] Институт автоматики и процессов управления ДВО РАН, г. Владивосток

[3] Владивостокский государственный университет, г. Владивосток

[4] Дальневосточный федеральный университет, г. Владивосток

[1]academy21@gmail.com, [2]maxim-grinev.it@yandex.ru, [3]mr.odnovil@mail.ru

Аннотация

Работа посвящена разработке и реализации системы управления сайтами (CMS) для создания научно-популярных порталов по геологии на примере адаптивной версии научно-популярного портала «История Земли: геологический ракурс». В качестве базовых сущностей разрабатываемого движка выбраны и реализованы следующие: «главная страница», «статья», «галерея», «видео», «3D-Земля», «временная линия», «компонент времени» и «3D Экспонат». В результате создан научно-популярный портал, позволяющий изучать нашу планету в разрезе временных интервалов, событий и артефактов.

Ключевые слова*: история Земли, геология Земли, научно-популярный портал, популяризация науки, научно-образовательный ресурс*

АКТУАЛЬНОСТЬ

Цифровая трансформация 21-го века привела к повсеместной автоматизации процессов в различных сферах человеческой деятельности. С одной стороны, это позволяет упростить решение различных задач, но в то же время приводит к незаинтересованности людей в понимании устройства, процессов и знании исто-

рии нашего мира. Одним из важных аспектов для понимания являются наша планета и её строение. Этими вопросами занимается такая наука, как геология. Она играет важную роль в понимании истории и свойств нашей планеты и процессов, происходящих в ней.

О геологии знают многие несмотря на то, что она является, пожалуй, единственной естественнонаучной дисциплиной, не изучаемой в школьных курсах. Развитие геологических знаний сопутствовало развитию человечества на всех этапах его истории. Достаточно вспомнить, что общая периодизация истории основана на характере орудий труда и материалов, используемых для производства: каменный, бронзовый и железный века [1, 2].

Геология помогает нам разгадать историю нашей планеты. Изучая горные породы, окаменелости и геологические процессы, геологи могут реконструировать события прошлого и понять, как развивалась Земля на протяжении миллиардов лет. Эти знания дают представление о формировании континентов, развитии жизни, изменении климата и крупных геологических событиях, таких как землетрясения, извержения вулканов и падение метеоритов.

Для решения вопроса популяризации научного знания по геологии Земли в Государственном геологическом музее им. В. И. Вернадского РАН были разработаны и реализованы две различных версии интерактивного научно-популярного портала «Живая Земля: геологический ракурс» [3 – 5]. В основу этих разработок легли такие интерактивные инструменты, как «линия времени», «3D-Земля», «информационные слои» и «назад в прошлое». С их помощью возможно выбирать геологический промежуток времени для изучения и получать имеющуюся в нём информацию в интерактивной форме. При этом для работы с этим порталом необходимо использование устройств с размерами экрана 10 и более дюймов (планшеты, ноутбуки и настольные компьютеры). Также разработанное решение не позволяет изменять структуру и наполнение портала в интерактивном режиме.

В связи с вышесказанным возникла идея разработки движка для научно-популярных порталов, позволяющего оперировать различными видами данных, которые используются в геологии: подготовленные структурированные тексты, 3D-Земля, 3D-экспонаты, видеоматериалы по геологии и т. д. При этом такой движок должен позволять устанавливать связи между размещённой информацией, а

также обладать адаптивностью интерфейса для различных типов устройств, включая мобильные. В основе навигации для разрабатываемого движка лежит принцип динамического времени, позволяющий перемещаться во времени и изучать информацию, относящуюся к выбранному периоду.

ОСНОВНЫЕ ЦЕЛИ

Целями проведенного исследования были разработка и реализация движка CMS по созданию научно-популярных порталов по геологии. Основным результатом стало создание полнофункционального веб-портала по геологии «История Земли: геологический ракурс», способного в удобной форме отображать научно-популярную информацию.

Для достижения поставленных целей были реализованы следующие этапы:

1. Анализ и проектирование: разработка дизайна и общего макета приложения; проектирование навигации в веб-приложении; составление структуры управления данными в нем; построение архитектуры API для взаимодействия клиентской части приложения с серверной [6]; реализация макета, футера и хедера веб-приложения; реализация навигации и маршрутизации в портале; реализация API.
2. Реализация: формирование макета, футера, хедера приложения; организация навигации и маршрутизации в портале; реализация API.

ИНСТРУМЕНТАРИЙ ДВИЖКА CMS

Для достижения поставленной целей по разработке основы для движка по созданию научно-популярных порталов необходимо было спроектировать архитектуру разрабатываемого веб-приложения, принимая во внимание его потенциальную эффективность и лаконичность. Благодаря продуманной архитектуре веб-приложение будет легче масштабировать, изменять, тестировать и отлаживать. При проектировании основы движка обращалось внимание на несколько ключевых критериев:

- Эффективность;
- Гибкость;
- Расширяемость;
- Масштабируемость процесса разработки;

- Возможность повторного использования;
- Хорошо структурированный и читаемый код.

В современном мире существуют различные типы архитектур веб-приложений в зависимости от того, как логика приложения распределяется между клиентской и серверной частями. К одной из наиболее распространенных архитектур веб-приложений относятся Одностраничные приложения (SPA). SPA – это веб-приложения, которые загружают одну HTML-страницу и динамически обновляют ее содержимое без перезагрузки всей страницы. SPA в значительной степени полагаются на JavaScript-фреймворки, такие как React, Angular или Vue.js, для обработки рендеринга на стороне клиента и управления состоянием приложения. Они обеспечивают бесшовный и отзывчивый пользовательский опыт, асинхронно получая данные с сервера и обновляя содержимое страницы в режиме реального времени. SPA хорошо подходят для интерактивных приложений и приложений с большим объемом данных. Поэтому для реализации целей работы и была выбрана архитектура SPA.

Для серверной части, аналогично клиентской, эффективным решением стало использование backend-фреймворков. Backend-фреймворки — это программные инструменты, которые облегчают разработку серверной части веб-приложений. Они помогают разработчикам создавать и управлять серверными приложениями, обрабатывать запросы клиентов, взаимодействовать с базами данных и реализовывать бизнес-логику.

В результате анализа существующих backend-фреймворков был выбран фреймворк Django [7]. Этот фреймворк основан на языке программирования Python и предоставляет все необходимое для разработки веб-приложений, включая модели данных, ORM, систему маршрутизации URL, автоматический административный интерфейс и многое другое. Данный фреймворк также известен своей простотой использования и мощными инструментами для разработки. Также названный фреймворк предоставляет административный интерфейс и множество функций без необходимости дополнительного конфигурирования. Ещё одной особенностью Django является наличие такого инструмента, как Migrations. Migrations — это функция, которая позволяет управлять базой данных (БД) и при-

менять изменения схемы БД с течением времени структурированным и автоматизированным образом. Миграции позволяют изменять схему базы данных по мере изменения моделей Django без ручного написания SQL-кода или изменения схемы базы данных напрямую. Миграции хранятся в виде файлов Python в директории "migrations/" каждого приложения Django. Каждый файл миграции содержит серию операций, которые определяют, как схема базы данных должна быть изменена или обновлена. Операции могут включать создание таблиц, добавление или изменение столбцов, создание индексов и многое другое.

Django Migrations значительно упрощает процесс управления и изменения схемы базы данных по мере развития приложения. Они помогают отслеживать изменения, автоматизировать процесс применения и отмены миграций, а также гарантировать, что схема БД остается синхронизированной с используемыми моделями Django.

Проектирование серверной части

Модели данных

В Django модели — это классы Python, которые определяют структуру и поведение таблиц базы данных. Модель представляет собой одну таблицу в БД и инкапсулирует поля, отношения и методы, связанные с этой таблицей. Она также обеспечивает уровень абстракции, который позволяет работать с базами данных без необходимости записывать SQL-код напрямую. Названная модель следует парадигме объектно-реляционного отображения (ORM), где таблицы БД представлены как классы Python, а записи базы данных – как экземпляры этих классов.

Таким образом, созданная БД имеет 6 моделей (сущностей), описание которых приведено ниже.

Статья (Article)

title	название статьи
time_ago	количество лет, прошедшее с события или времени, о котором рассказывает статья
image	изображение
text	текст статьи
src_article	ссылка на статью

src_magazine	журнал, из которого была взята статья
Time	период, к которому принадлежит статья (выбирается из всех доступных периодов времени)

Экспонат (Exhibit)

title	название экспоната
time_ago	количество лет, прошедшее с события или времени, к которому относится экспонат
image	изображение
text	текст экспоната
src_article	ссылка на информацию об экспонате
time	период, к которому принадлежит экспонат (выбирается из всех доступных периодов времени)

Земля (Earth)

title	название
time_ago	количество лет, прошедших с данного периода земли
text	информация о земле данного периода
time	период, к которому принадлежит данная версия земли (выбирается из периодов Земли)
text_more	текст для страницы «Узнать больше»
image_more	изображение для страницы «Узнать больше»
src_article	ссылка на информацию
src_magazine	журнал, из которого взята информация
baseMap	текстура, представляющая собой основную текстуру (base texture) объекта; она определяет основной цвет или рисунок, который будет виден на поверхности объекта
ambientMap	текстура, представляющая собой карту окружающей освещенности (ambient lighting); она определяет, каким образом объект будет отражать окружающее освещение

Heightmap	текстура, представляющая собой карту высот (height map) объекта; она используется для создания рельефного эффекта или имитации трехмерности на плоской поверхности
metallicMap	текстура, определяющая степень металличности (metallic) объекта, показывая, где объект является металлическим, а где нет, влияя на его отражательные свойства
normalMap	текстура, представляющая собой карту нормалей (normal map) объекта; она используется для создания рельефных деталей и имитации поверхностных неровностей на объекте
roughnessMap	данная текстура определяет степень шероховатости (roughness) объекта, влияя на его отражательные свойства
cloudMap	текстура, представляющая собой карту распределения облачности (cloud map) на поверхности планеты или сферы

Видео (Video)

time_ago	количество лет, прошедшее с события, показанного в видео
Video	само видео в формате .mp4
time	период, к которому относится видео (выбирается из всех доступных периодов времени)

Локация (Location)

Title	название локации
Image	изображение
Time	период, к которому принадлежит данная локация (выбирается из всех доступных периодов времени)

Реконструкция (Reconstruction)

Title	название реконструкции
time_ago	количество лет, прошедшее с данной реконструкции
Position	местоположение
Coordinates	координаты

Image	изображение
Text	текст
Location	локация (выбирается из доступных объектов модели «Локация»)

Проектирование интерфейса

Реализация внешнего вида портала основана на проекте интерфейса, разработанного в рамках предыдущей работы [5]. Концепция интерфейса основана на проведённом исследовании различных фокус-групп по удобству подачи материала научно-популярной тематики. Далее приведена структура всех блоков, разработанных на портале, с их реализацией на основе имеющегося дизайна.

Дизайн-макет Главной страницы выглядит следующим образом (рис. 1).

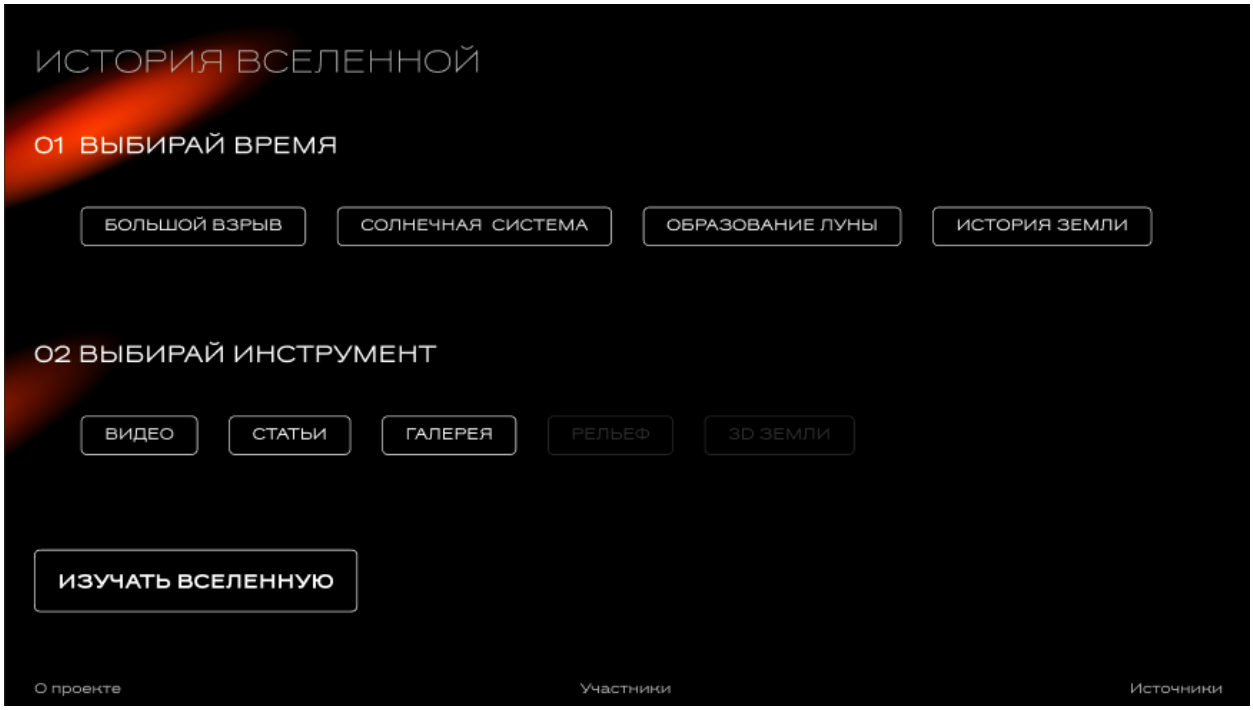

Рис. 1. Дизайн-макет Главной страницы

Требования к функционалу главной страницы – возможности:

- выбора времени;
- выбора инструмента;
- перехода в разделы «О проекте», «Участники», «Источники»;
- перехода на определенную страницу, в зависимости от выбранных времени и инструмента.

Для реализации предложенного макета Главной страницы был разработан шаблон html-страницы (рис. 2), состоящий из следующих элементов:

- <h1></h1> – заголовок стартового экрана – «История вселенной»;
- <h2></h2> – заголовок с призывом для выбора времени;
- <button/> – кнопки выбора временных периодов;
- <EarthTypeMenu/> – выпадающий список для выбора периода жизни Земли, например, «Черная земля», «Белая земля», «Голубая земля» и т.д.
- <h2></h2> – заголовок с призывом для выбора инструмента;
- <button/> – кнопки выбора инструмента;
- <button/> – кнопка перехода на выбранный период времени и соответствующий инструмент – «Изучать Вселенную»;
- <Footer/> – блок, содержащий ссылки на разделы «О проекте», «Участники» и «Источники».

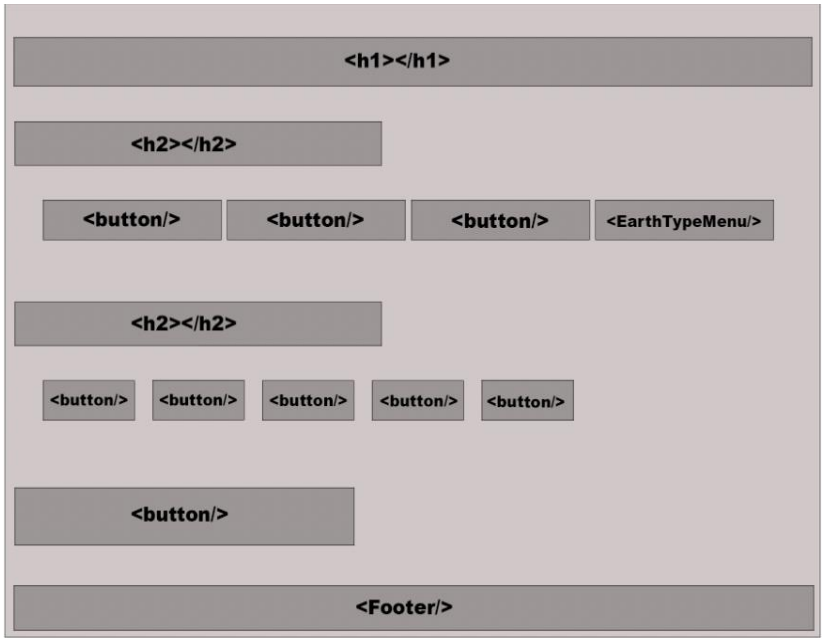

Рис. 2. Визуализация компонента «Главная страница»

Раздел «Статьи»

Этот раздел состоит из двух страниц: страница со списком всех статей (рис. 4) и страница для отображения статьи, выбранной для чтения (рис. 6). На

рис. 3 представлен шаблон html-страницы, описывающий раздел «Список статей». В процессе разработки данного шаблона были учтены следующие функциональные требования – возможности:

- выбора статьи для перехода;
- предпросмотра статьи (название и изображение);
- смены времени;
- смены инструмента;
- перехода на главную страницу;
- перехода на статьи, привязанные к следующему/предыдущему времени.

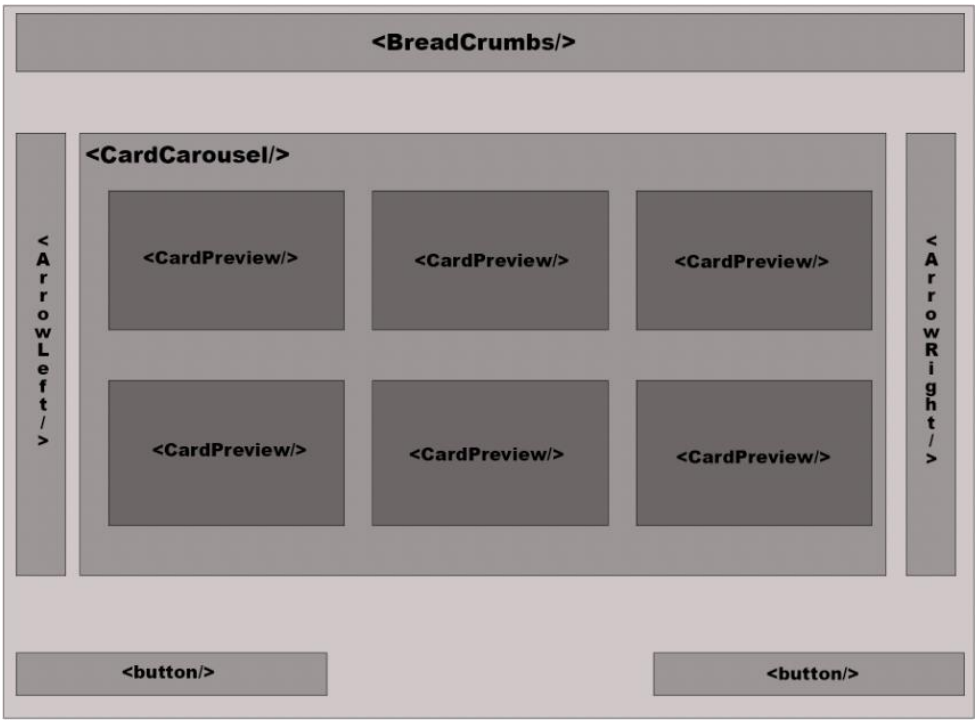

Рис. 3. Визуализация шаблона «Список статей»

С целью реализации обозначенных функциональных требований для раздела «Список статей» были разработаны следующие компоненты:

- <BreadCrumbs/> – компонент «хлебные крошки», позволяющий видеть, где вы находитесь на сайте, а также реализующий переход на главную страницу, смену времени и смену инструмента;
- <ArrowLeft/> – компонент, являющийся кнопкой для перелистывания между группами в компоненте <CardCarousel/>;

- <CardCarousel/> – компонент «карусель»; принимает на вход массив карточек и отображает элементы массива по одному, позволяя перемещаться между ними с помощью кнопок; такой компонент упрощает навигацию между статьями;
- <ArrowRight/> – аналогичен <ArrowLeft/>, но позволяет перелистывать вправо;
- <CardPreview/> – компонент, представляющий собой предварительный просмотр статьи; показывает изображение и название статьи;
- <button/> – кнопки, позволяющие переходить на следующее/предыдущее время, чтобы в компоненте «карусель» были показаны статьи, принадлежащие другому времени.

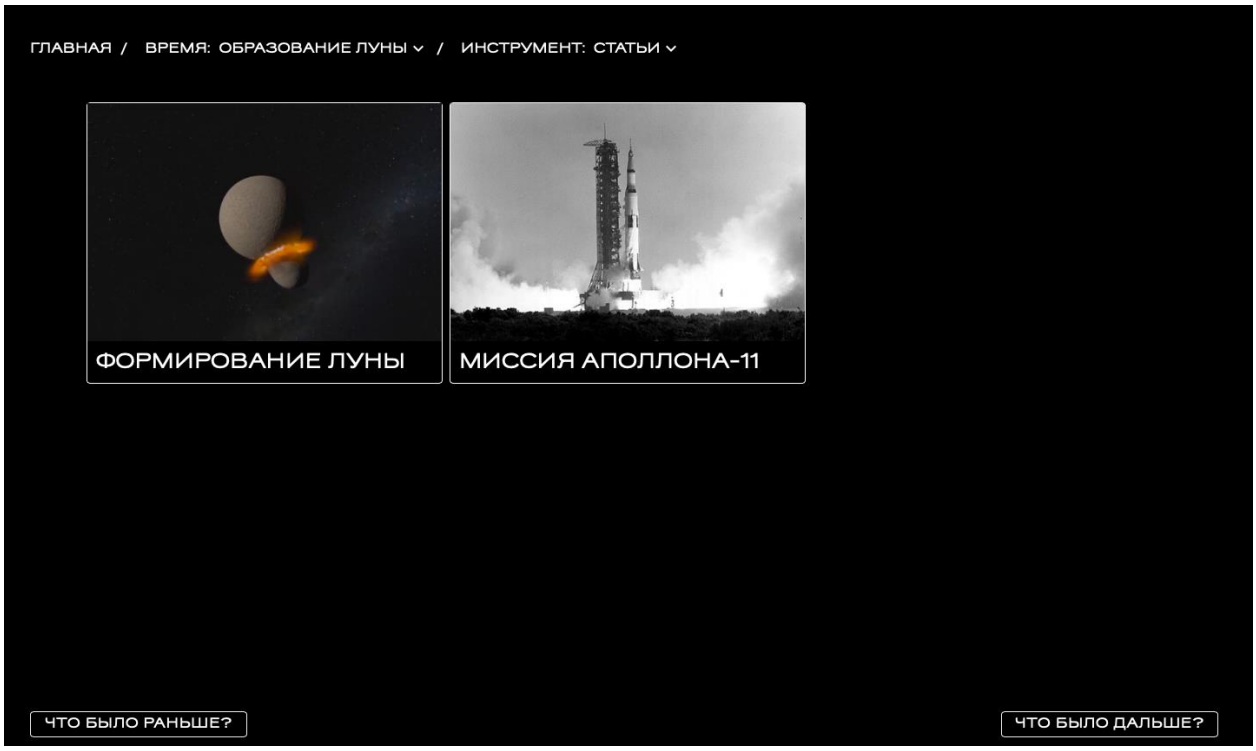

Рис. 4. Реализованная страница «Список статей»

Страница отдельной статьи

Для раздела «Страница статьи» был разработан шаблон html-страницы (рис. 5), реализующий следующие функциональные требования – возможности:

- просмотра названия;
- просмотра текста;

- просмотра источников (ссылка, журнал);
- перехода на главную страницу;
- смены времени;
- смены инструмента;
- перехода на другие статьи, связанные с этим временем;
- перехода обратно на страницу списка статей.

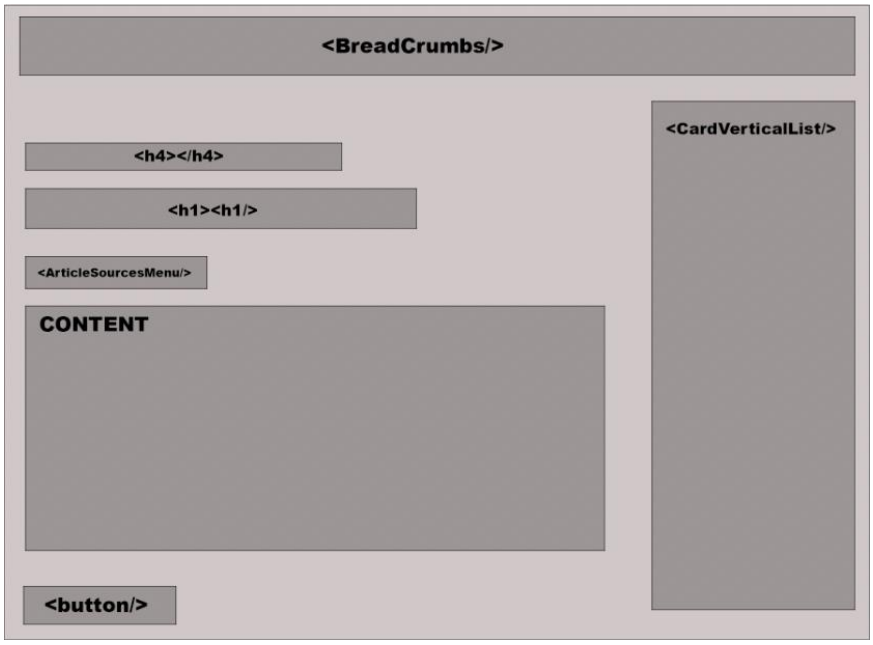

Рис. 5. Визуализация шаблона «Страница статьи»

С целью реализации названных функциональных требований для раздела «Страница статьи» (рис. 6) были разработаны следующие компоненты:

- <h4></h4> – подзаголовок, отвечающий за количество лет, прошедших с события, рассказанного в статье;
- <h1></h1> – название статьи;
- <ArticleSourcesMenu/> – выпадающее меню, отвечающее за показ ссылки на статьи и журнала, в котором она была опубликована;
- CONTENT – текст статьи, представленный абзацами, завернутыми в тег <p>;
- <button/> – кнопка «Назад», отвечающая за переход на страницу списка статей;

- <CardVerticalList/> – виртуализированный вертикальный список, содержащий карточки (предпросмотр) других статей, связанных с данным временем.

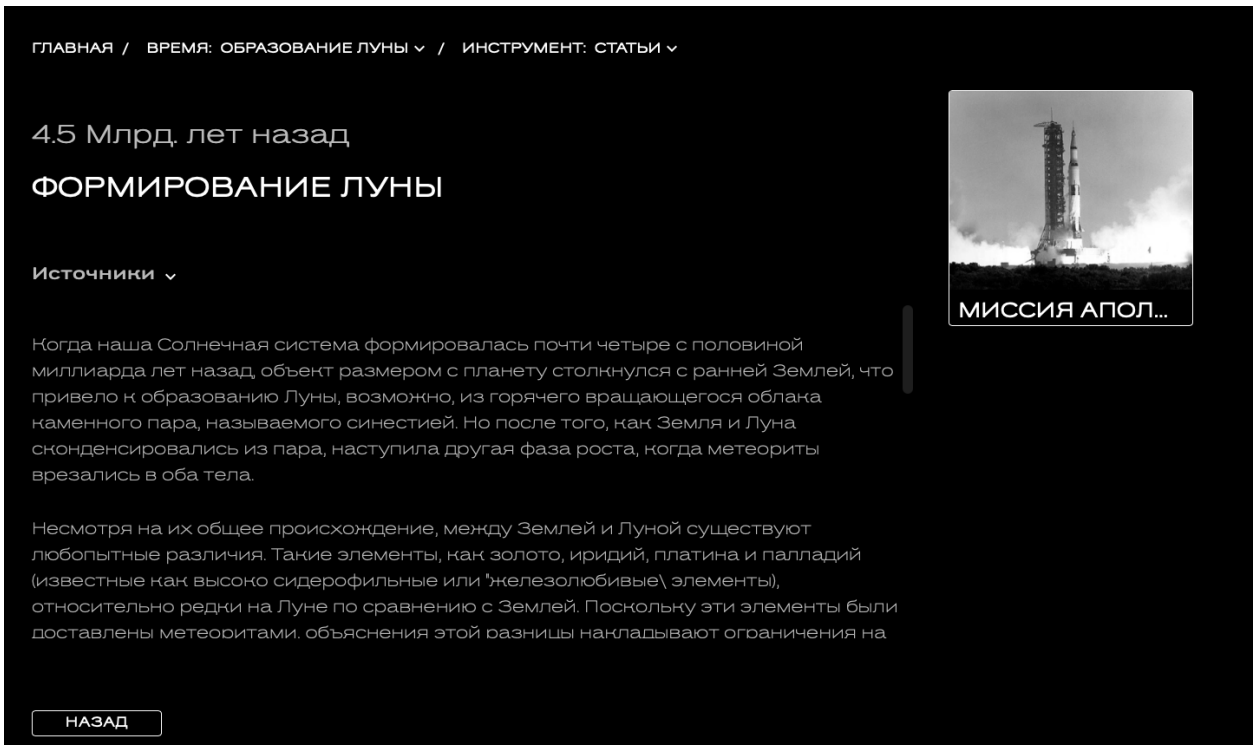

Рис. 6. Реализованная раздел «Страница статьи»

Страница со списком экспонатов

Эта страница аналогична странице «Список статей» и имеет такую же структуру (рис. 7). Названные разделы на портале имеют различия в источниках отображаемых данных, получаемых компонентом «карусель».

Рис. 7. Реализованная страница «Список экспонатов»

Страница экспоната

Для страницы «Экспонат» были разработаны html-шаблон (рис. 8) и его реализация (рис. 9) со следующим списком функциональных требований – возможностей:

- просмотра изображения экспоната;
- просмотра источника;
- просмотра названия;
- просмотра текста;
- перехода на другие экспонаты;
- перехода обратно на страницу «Список экспонатов»;
- смены времени;
- смены инструмента;
- перехода на главную страницу.

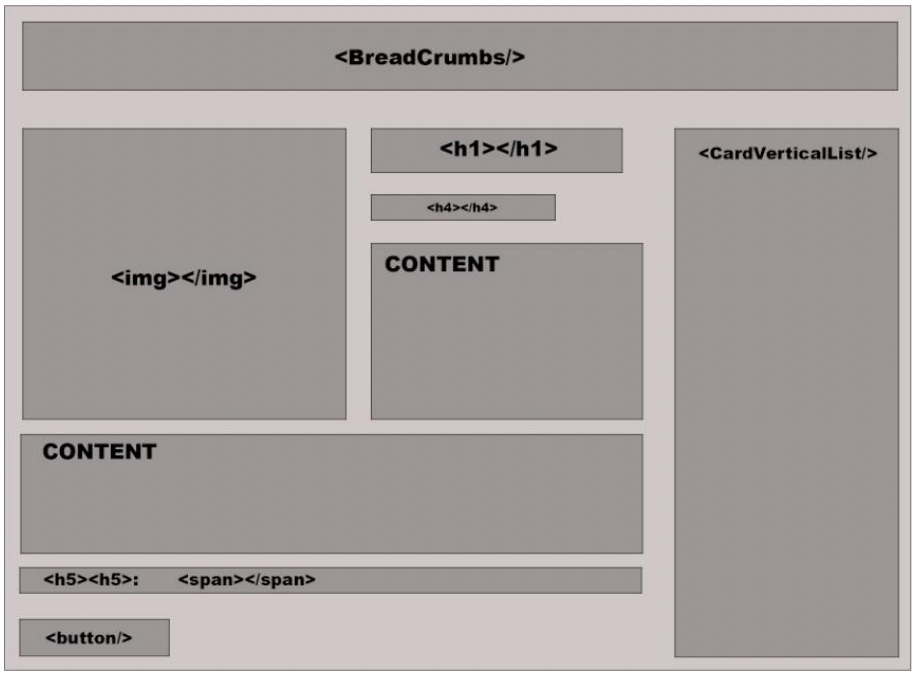

Рис. 8. Визуализация шаблона «Экспонат»

Html-шаблон, разработанный для страницы «Экспонат», состоит из следующих элементов:

- <h1></h1> – название экспоната;
- <h4></h4> – количество лет;
- <img></img> – изображение экспоната;
- CONTENT – текст экспоната, являющийся абзацами, завёрнутыми в тег <p>, также текст обёрнут снизу и справой стороны изображения;
- <h5></h5> – надпись «Источник»;
- – тег, в котором находится название источника или ссылка;
- <button/> – кнопка «Назад»;
- <CardVerticalList/> – вертикальный виртуализированный список других экспонатов, принадлежащих данному времени.

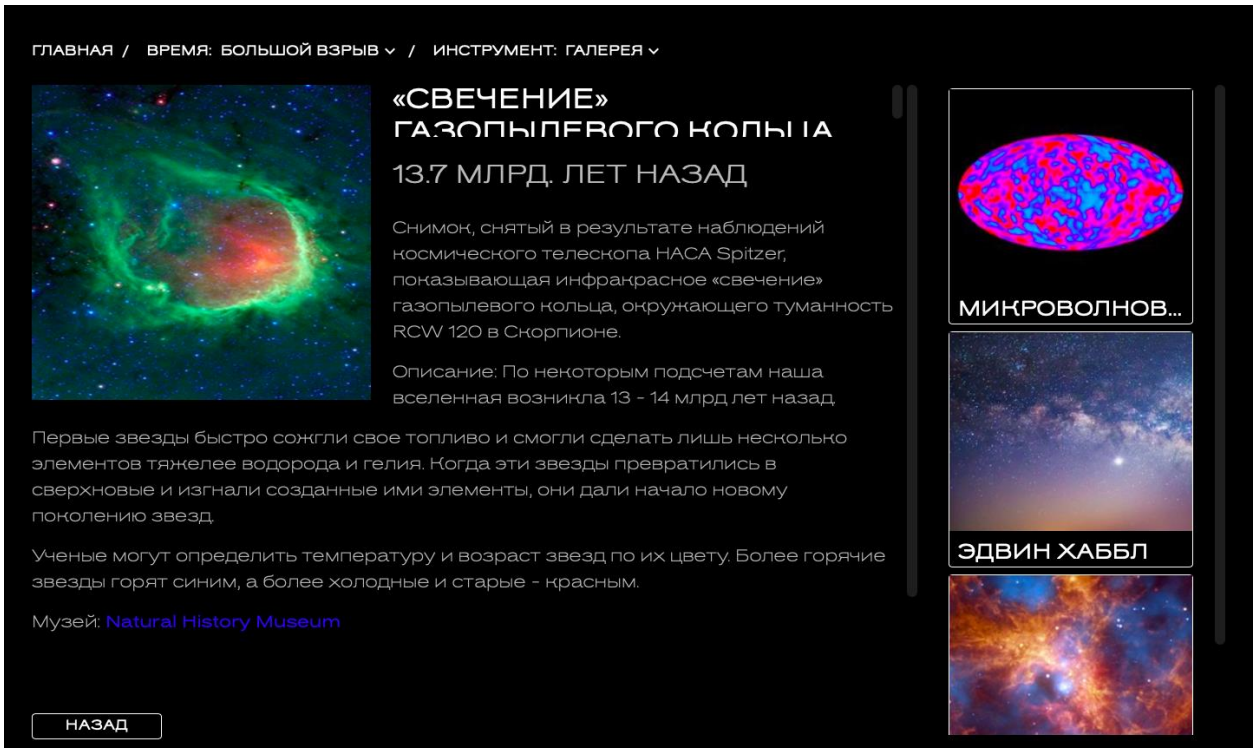

Рис. 9. Реализованная страница «Страница экспоната»

Раздел Видео

Для этого раздела разработаны html-шаблон (рис. 10) и его реализация (рис. 11) со следующим списком функциональных требований – возможностей:

- просмотра видео;
- перематывания видео на 5 секунд вперёд/назад;
- использования полноэкранного режима;
- перехода на главную страницу;
- смены времени;
- смены инструмента;
- перехода на следующее/предыдущее время

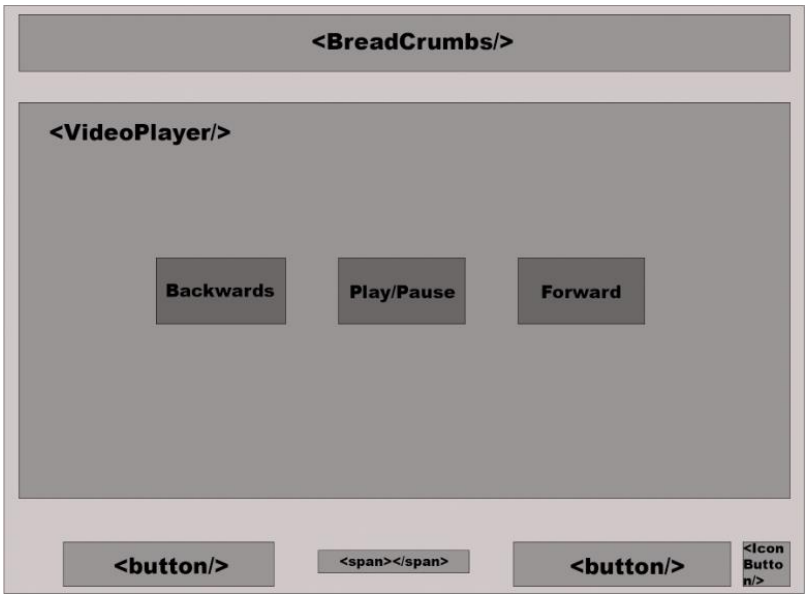

Рис. 10. Визуализация шаблона «Видео»

Разработанный html-шаблон состоит из следующих компонентов:

- <VideoPlayer/> – компонент, показывающий видео;
- Backwards – кнопка, перематывающая видео на 5 секунд назад, работает на кнопке «стрелка влево» на клавиатуре;
- Play/Pause – кнопка, проигрывающая/приостанавливающая видео, работает на кнопке «пробел» на клавиатуре;
- Forward – кнопка, перематывающая видео на 5 секунд вперёд, работает на кнопке «стрелка вправо» на клавиатуре;
- <button/> – кнопки, отвечающие за переход на следующее/предыдущее время;
- – тег, содержащий количество лет, прошедшее с события, показанного на видео;
- <IconButton/> – кнопка, выглядящая как иконка полноэкранного режима, выполняющая функцию включения полноэкранного режима.

Рис. 11. Реализованная страница раздела «Видео»

Раздел «3D-модель Земли»

Этот раздел состоит из двух страниц – страница, отображающая 3D-модель Земли (рис. 13) с краткой справочной информацией, и страница «Узнать больше» (рис. 15), где представлена более развёрнутая информация о выбранном периоде из жизни Земли.

3D-модель Земли

Для реализации раздела «3D-модель Земли» разработаны html-шаблон (рис. 12) и его реализация (рис. 13) со следующим списком функциональных требований – возможностей:

- просмотра названия периода данной модели Земли;
- просмотра текста, описывающего данную модель Земли;
- перехода на страницу «Узнать больше»;
- взаимодействия с 3D-моделью Земли;
- перехода на следующую/предыдущую 3D-модель Земли;
- перехода на главную страницу;
- смены времени;
- смены инструмента.

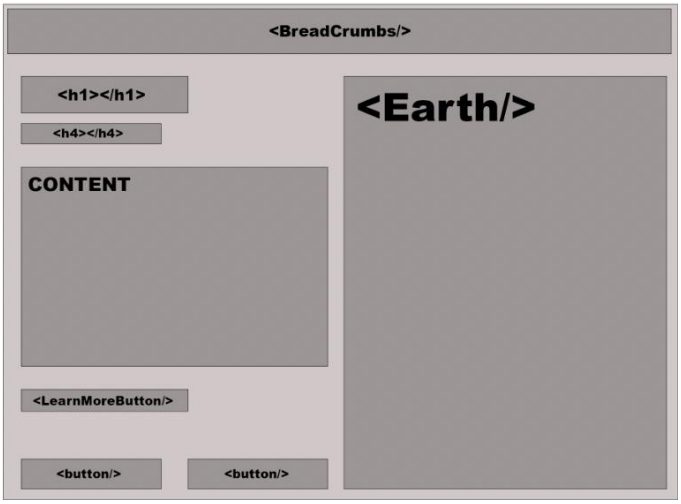

Рис. 12. Визуализация шаблона «3D модель Земли»

Разработанный html-шаблон состоит из следующих элементов:

- <h1></h1> – название периода земли, модель которой представлена на странице;
- <h2></h2> – количество лет, прошедших с данного периода Земли;
- CONTENT – текст, описывающий данную землю;
- <LearnMoreButton/> – кнопка, по нажатию которой осуществляется переход на страницу «Узнать больше»;
- <button/> – кнопки, осуществляющие переход на следующий/предыдущий период изучения Земли;
- <Earth/> – 3D-модель Земли; модель является интерактивной с возможностью её осмотра путём её вращения; также при отсутствии действий со стороны пользователя модель Земли начинает вращаться сама.

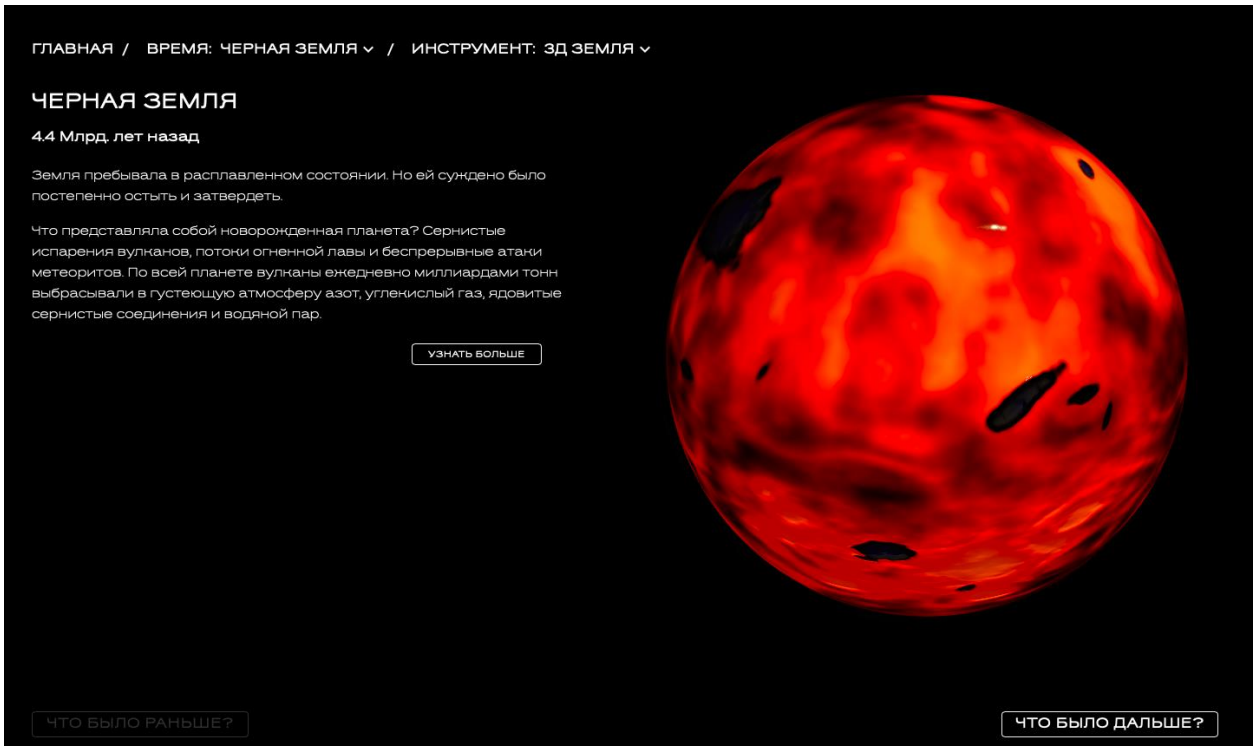

Рис. 13. Реализованная страница «3D-модель Земли»

Раздел «Узнать больше»

Для реализации раздела «Узнать больше» разработаны html-шаблон (рис. 14) и его реализация (рис. 15) со следующим списком функциональных требований – возможностей:

- просмотра названия статьи, связанной с этим периодом Земли;
- просмотра текста статьи;
- просмотра изображения события или объекта, описываемого в статье;
- просмотра источников (ссылка, журнал);
- перехода на главную страницу;
- перехода обратно на страницу 3D-модели Земли;
- смены времени;
- смены инструмента.

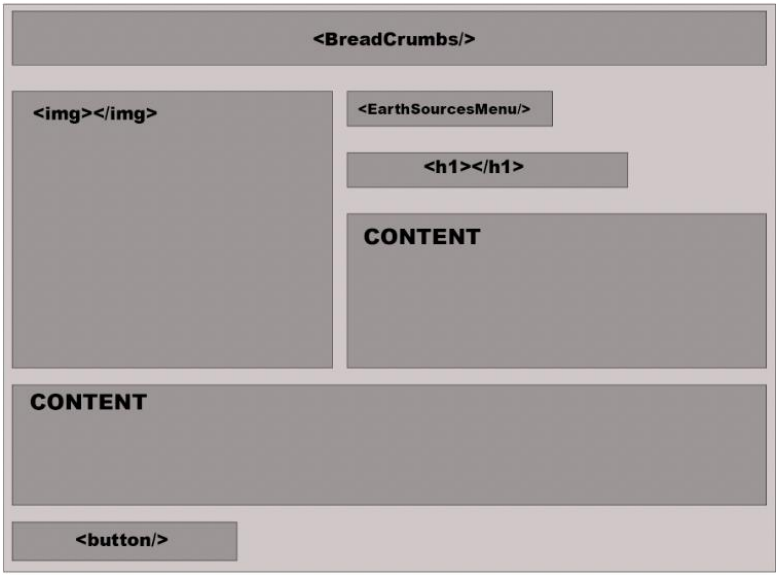

Рис. 14. Визуализация шаблона «Узнать больше»

Разработанный html-шаблон для страницы «Узнать больше» состоит из следующих элементов:

- <img></img> – изображение;
- <EarthSourcesMenu/> – выпадающий список, содержащий источники в виде ссылки на статью и журналы;
- <h1></h1> – название изучаемого периода Земли;
- CONTENT – текст статьи;
- <button/> – кнопка «Назад», осуществляющая переход на страницу 3D-модель Земли.

Рис. 15. Дизайн макет страницы «Узнать больше».

Страница «Реконструкции»

Для раздела «Реконструкции» разработан следующий дизайн-макет (рис. 16).

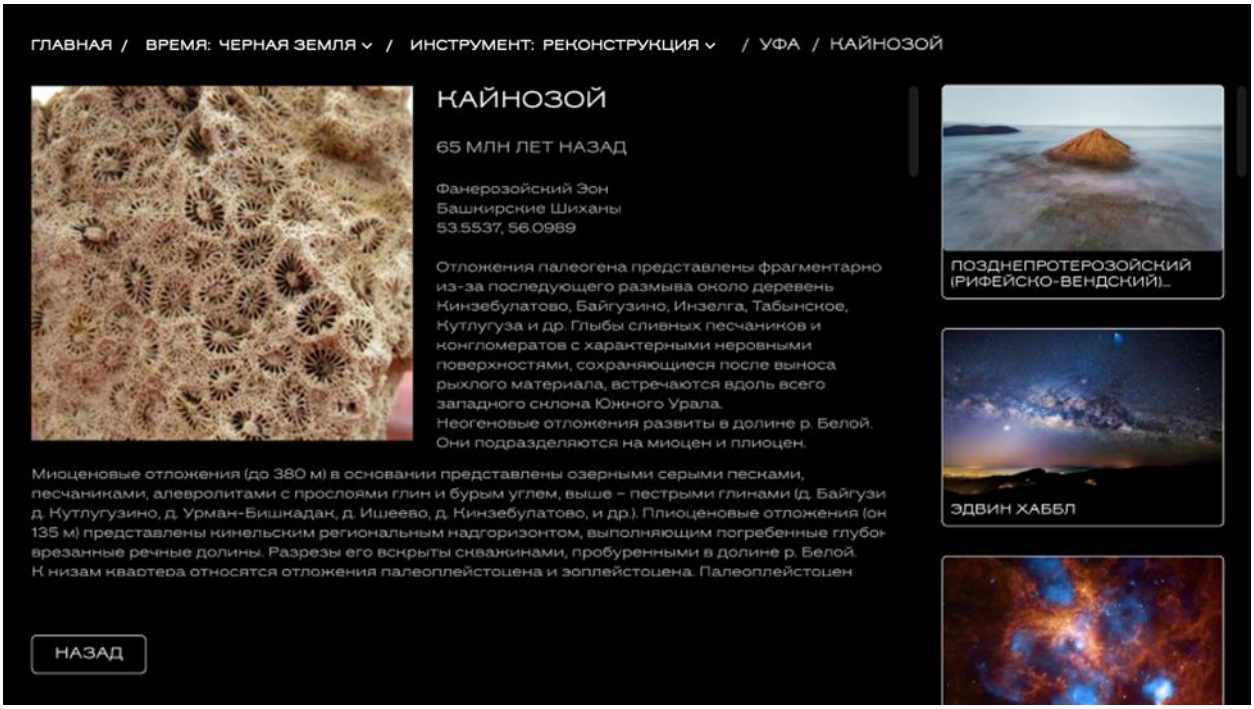

Рис. 16. Дизайн макет страницы «Реконструкция»

К данной странице предъявляются следующие функциональные требования – возможности:

- просмотра изображения реконструкции;
- просмотра источника;
- просмотра названия реконструкции;
- просмотра текста;
- перехода на другие реконструкции данной локации;
- перехода обратно на страницу «Список локаций»;
- смены времени;
- смены инструмента;
- перехода на главную страницу.

Для реализации предложенного макета страницы «Реконструкция» разработан шаблон html-страницы (рис. 17), состоящий из следующих элементов:

- <img></img> – изображение реконструкции;
- <h1></h1> – название реконструкции;
- <h4></h4> – количество лет, прошедших с описываемого периода;
- – местоположение и координаты, расположенные друг под другом;
- CONTENT – текст, описывающий реконструкцию;
- <h5></h5> – надпись «Источник»;
- – источник;
- <button/> – кнопка «Назад», возвращающая на страницу «Список локаций».

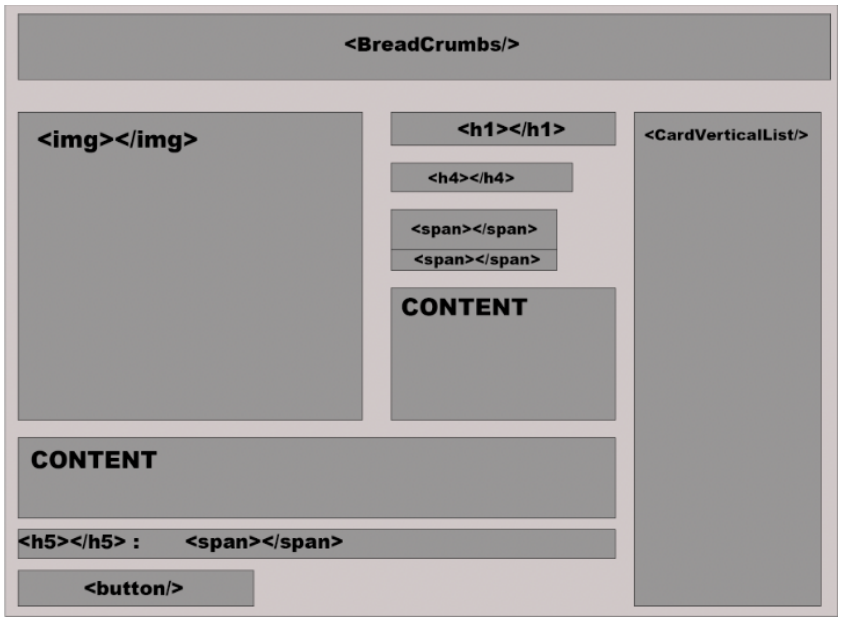

Рис. 17. Визуализация компонента страница «Реконструкция»

Страница «3D экспонат»

Для обеспечения возможности дополнительного виртуального погружения в изучение геологии Земли был спроектирован раздел портала, позволяющий пользователю изучать 3D-сканы различных геологических экспонатов. Прототип интерфейса данного раздела представлен на рис. 18. В данный момент раздел находится на этапе реализации и в скором времени будет доступен на основном сайте портала.

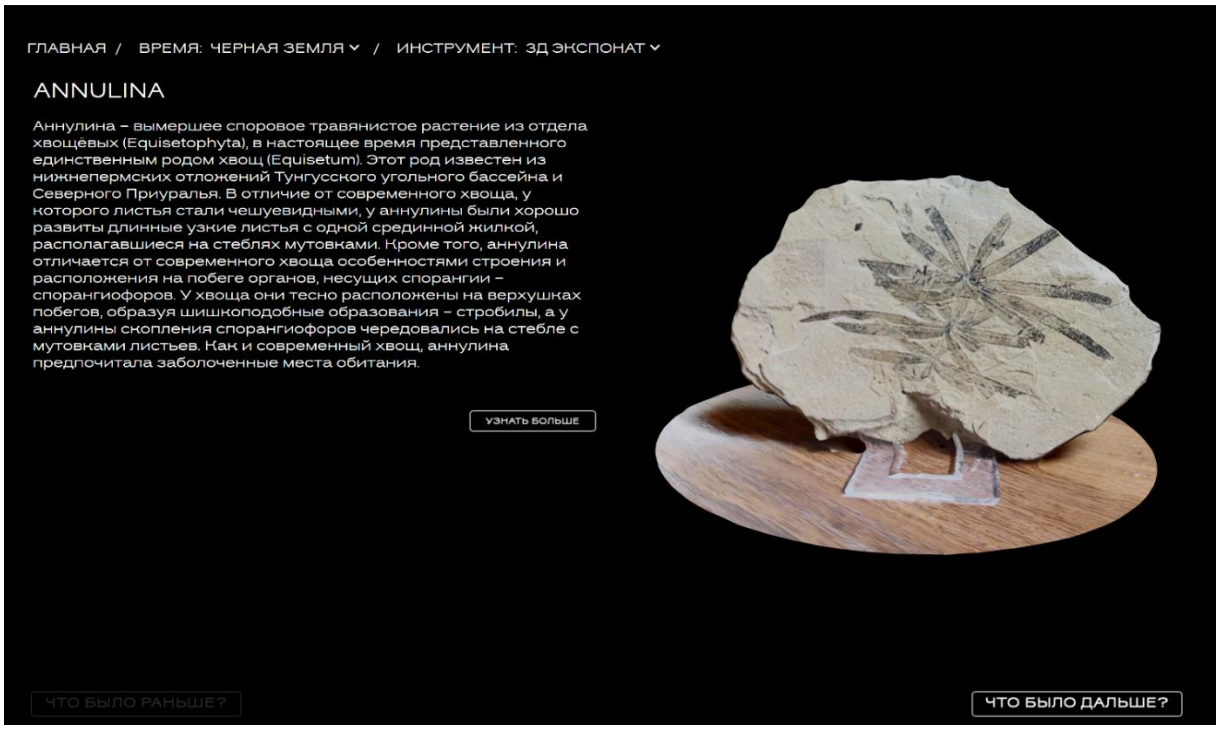

Рис. 18. Макет раздела «3D-экспонат»

ЗАКЛЮЧЕНИЕ

В результате выполнения проекта успешно разработан и внедрен веб-портал (https://new.populargeology.ru), который обеспечивает эффективное управление и представление геологической информации в онлайн-среде. Созданный портал по геологии полностью функционален и способен эффективно отображать и предоставлять доступ к данным в различных форматах.

Работы выполняются в рамках Государственного задания ГГМ РАН по Теме № 0140-2019-0005 «Разработка информационной среды интеграции данных естественнонаучных музеев и сервисов их обработки для наук о Земле», а также Государственной темы № 1021061009468-8-1.5.1 «Цифровая платформа интеграции и анализа геологических и музейных данных».

Авторы работы выражают благодарность за постановку задачи Вере Викторовне Наумовой, главному научному сотруднику, заведующей Научным отделом Государственного геологического музея им. В.И. Вернадского РАН.

За разработку концепции дизайна внешнего вида портала авторы выражают благодарность выпускницам Института математики и компьютерных наук

Дальневосточного федерального университета Д.Е. Лещиковой и Л.С. Романенковой.

За сканирование и обработку 3D-моделей музейных экспонатов выражаем благодарность магистру МГРИ Александру Безкоровайному.

СПИСОК ЛИТЕРАТУРЫ

1. Образовательный геологический сайт Юрия Попова [Электронный ресурс] // URL: https://popovgeo.sfedu.ru (дата обращения: 01.10.2023)

2. Все о геологии [Электронный ресурс] // URL: https://geo.web.ru (дата обращения: 01.10.2023)

3. *Eremenko A.S., Naumova V.V.* The development of popular-science portal "LIVING EARTH: GEOLOGICAL PERSPECTIVE" // Proceedings of the V International Conference “Information Technologies in Earth Sciences and Applications for Geology, Mining and Economy” (ITES&MP-2019), Moscow, Russia, October 14–18, 2019, CEUR-WS.org/Vol-2527/short2.pdf

4. *Еременко А.С., Наумова В.В., Загумённов А.А., Ерёменко В.С., Злобина А.Н.* Интернет-портал "История Земли: геологический ракурс". Высокотехнологичная популяризация научных геологических знаний // Электронные библиотеки. 2021. Т. 24. № 4. С. 604–621.

5. *Ерёменко А.С., Лещикова Д.Е., Романенкова Л.С.* Проектирование дальнейшего развития научно-популярного портала «История Земли: геологический ракурс» // Электронные библиотеки. 2022. Т. 25. № 4. С. 317–335.

6. Django REST Framework [Электронный ресурс] // URL: https://www.django-rest-framework.org (дата обращения 01.10.2023)

7. Django documentation [Электронный ресурс] // URL: https://docs.djangoproject.com/en/4.2 (дата обращения 01.10.2023)

IMPLEMENTATION OF AN ENGINE FOR CREATING POPULAR SCIENCE PORTALS ON GEOLOGY (USING THE EXAMPLE OF THE PORTAL “HISTORY OF THE EARTH: GEOLOGICAL PERSPECTIVE”)

Aleksandr Eremenko[1, 2, 3] [0000-0003-1923-8417], **Maksim Grinev**[3] [0009-0007-9611-6946], **Evgeniy Odnovil**[4] [0009-0007-9611-6946]

[1]*Vernadsky State Geological Museum of the Russian Academy of Sciences, Moscow*

[2]*Institute of Automation and Control Processes FEB RAS, Vladivostok*

[3]*Vladivostok State University, Vladivostok*

[4]*Far Eastern Federal University, Vladivostok*

[1]academy21@gmail.com, [2]maxim-grinev.it@yandex.ru, [3]mr.odnovil@mail.ru

Abstract

The work is devoted to the development and implementation of a CMS engine for creating popular science portals on geology with the subsequent implementation of an adaptive version of the popular science portal “History of the Earth: a geological perspective”. The following were selected and implemented as the basic entities of the engine being developed: “main page”, “article”, “gallery”, “video”, “3D-Earth”, “time-line”, “time component” and “3D Exhibit”. As a result of the work done, a popular science portal was created that allows us to study our planet in the context of time intervals, events and artifacts.

Keywords*: history of the Earth, geology of the Earth, popular science portal, popularization of science, scientific and educational resource*

REFERENCES

1. Educational geological site of Yuri Popov [Electronic resource] // URL: https://popovgeo.sfedu.ru (access date: 10/01/2023)

2. All about geology [Electronic resource] // URL: https://geo.web.ru (access date: 10/01/2023)

3. *Eremenko A.S., Naumova V.V.* THE DEVELOPMENT OF POPULAR-SCIENCE PORTAL "LIVING EARTH: GEOLOGICAL PERSPECTIVE" // Proceedings of the V Interna-

tional Conference “Information Technologies in Earth Sciences and Applications for Geology, Mining and Economy” (ITES&MP- 2019), Moscow, Russia, October 14–18, 2019, CEUR-WS.org/Vol-2527/short2.pdf

4. *Eremenko A.S., Naumova V.V., Zagumennov A.A., Eremenko V.S., Zlobina A.N.* Internet portal "History of the Earth: geological perspective." High-tech popularization of scientific geological knowledge // Electronic libraries. 2021. T. 24. No. 4. P. 604–621.

5. *Eremenko A.S., Leshchikova D.E., Romanenkova L.S.* Designing the further development of the popular science portal “History of the Earth: geological perspective” // Electronic libraries. 2022. T. 25. No. 4, P. 317–335.

6. Django REST Framework [Electronic resource] //
URL: https://www.django-rest-framework.org (last access 01.10.2023)

7. Django documentation [Electronic resource] //
URL: https://docs.djangoproject.com/en/4.2 (last access 01.10.2023)

СВЕДЕНИЯ ОБ АВТОРАХ

ЕРЕМЕНКО Александр Сергеевич – кандидат технических наук, старший научный сотрудник лаборатории спутникового мониторинга Института автоматики и процессов управления ДВО РАН, внештатный сотрудник Государственного геологического музея им. В.И. Вернадского РАН

Aleksandr EREMENKO – Senior Researcher, Candidate of Technical Sciences, Satellite Monitor-ing Laboratory, Institute of Automation and Control Processes, Far Eastern Branch of the Rus-sian Academy of Sciences, Freelance employee of the V.I. Vernadsky State Geological Museum of the Russian Academy of Sciences.

email: academy21@gmail.com

ORCID: 0000-0003-1923-8417

ГРИНЁВ Максим Владимирович – студент магистерского направления «Бизнес информатика» Владивостокского государственного университета.

Maksim GRINEV – Student of the Vladivostok State University Master's program «Business Informatics», Vladivostok State University.

email: maxim-grinev.it@yandex.ru

ORCID: 0009-0007-9611-6946

ОДНОВИЛ Евгений Александрович – студент Магистерского направления ДВФУ «Программирование игр, цифровых развлечений, виртуальной и дополненной реальности (совместно с ЦК НТИ по нейротехнологиям), технологиям виртуальной и дополненной реальности».

Evgeniy ODNOVIL – Student of the Far Eastern Federal University Master's program «Programming of games, digital entertainment, virtual and augmented reality (together with the NTI Central Committee on neurotechnologies), virtual and augmented reality technologies».

email: mr.odnovil@mail.ru

ORCID: 0009-0007-2368-4379

Материал поступил в редакцию 2 октября 2023 года

УДК 004.738.52

ИНДЕКСЫ ЦИТИРОВАНИЯ И ОЦЕНКА ПУБЛИКАЦИОННОЙ АКТИВНОСТИ АВТОРОВ

А. С. Козицын[1] [0000-0002-8065-9061], С. А. Афонин[2] [0000-0003-3058-9269], Д. А. Шачнев[3] [0000-0002-5940-9180]

[1-3]*НИИ механики МГУ им. М.В. Ломоносова*

[1]alexanderkz@mail.ru, [2]serg@msu.ru, [3]mitya57@gmail.com

Аннотация

В современном научном мире одним из способов оценки успешности научной деятельности ученого является вычисление различных показателей, основанных на количестве его публикаций и их цитируемости. При этом каждый соавтор публикации получает за нее одинаковое количество баллов. Подобный способ оценки приводит к искусственному увеличению количества соавторов, что, в свою очередь, влечет за собой искажение рейтинговых оценок научной деятельности в организации, а также значительно снижает качество тематического поиска по библиографическим данным экспертов, конференций и журналов. Представленный в работе метод позволяет оценить степень влияния указанного фактора на показатели, основанные на учете количества и цитируемости научных публикаций. Апробация метода проводилась на данных наукометрической системы ИАС «ИСТИНА».

***Ключевые слова**: ранжирование, наукометрия, наукометрические системы, соавторство, системы цитирования, научный рейтинг*

ВВЕДЕНИЕ

В современном научном мире одним из способов оценки успешности научной деятельности ученого является подсчет различных показателей, основанных на количестве его публикаций и их цитируемости. Наиболее распространенным показателем оценки деятельности ученого является h-индекс (индекс Хирша). Несмотря на популярность этого показателя, необходимо учитывать имеющиеся у него недостатки. Во-первых, индекс Хирша существенно зависит от области охвата публикаций – в зависимости от выбранной системы цитирования (WoS, Scopus,

Google Scholar, РИНЦ) результаты оценки для любого ученого будут разными. Во-вторых, этот индекс не учитывает тематическую направленность работ. Вместе с тем, средний индекс цитирования в области физики и медицины в несколько раз превышает средний индекс цитирования в области математики и информационных технологий [1]. В-третьих, индекс Хирша не учитывает самоцитирования и цитирования соавторов. В-четвертых, не учитывается авторитетность цитирующей статьи или издания, что особенно важно для систем с очень большим охватом разнородных источников, например, Goggle Scholar. Поэтому на основе индекса Хирша было построено множество модификаций [2], создатели которых пытались в той или иной мере устранить недостатки, перечисленные выше.

Ниже приведены наиболее известные персональные индексы цитирования [3], использующие данные о цитировании каждой из N статей автора C_i, отсортированных в порядке убывания, количестве соавторов статьи A_i, а также о количестве лет Y_i, прошедших с публикации статьи.

h-индекс. Классический индекс Хирша [4]: $h = \max(x : C_x \geq x)$.

h_α-индекс. Небольшая модификация индекса Хирша, предложенная в 2008 году [5]: $h_\alpha = \max(x : C_x \geq \alpha * x)$.

h2-индекс. Предложен [6] для компенсации самоцитирований статей и рассчитывается по формуле $h2 = \max(x : C_x \geq x^2)$.

g-индекс. Учитывает количество ссылок на наиболее цитируемые работы [7]: $g = \max(x : \sum_{i=1}^{x} C_i \geq x^2)$

g1-индекс. Расширяет диапазон возможных значений g-индекса и использует действительные числа [8]:

$$g1 = g + \frac{(g+1)^2 - \sum_{i=1}^{g+1} C_i}{(g+1)^2 - g^2}.$$

hg—индекс: рассчитывается как среднее геометрическое между h- и g-индексом [9]: $hg = \sqrt{h \cdot g}$.

ghp-индекс: рассчитывается по всем цитированиям [10]

$$ghp = \sqrt{h^2 + \sum_{i=1}^{h}(C_i - h) + \sum_{i=h+1}^{N} C_i} = \sqrt{\sum_{i=1}^{N} C_i}\ .$$

a-индекс: это среднее количество цитирований

$$a = \frac{\sum_{i=1}^{N} C_i}{N}.$$

ISI-индекс. Учитывает самоцитирования и количество соавторов в статье без учета порядка [11]. Вычисляется по формуле

$$isi = \sum_{i=1}^{N} \frac{C_i - 0.75 \cdot SC_i}{A_i}.$$

hi-индекс. Учитывает количество соавторов в статьях без учета порядка. Предлагались два варианта:

$hi = \max(x : \frac{C_x}{A_x} \geq x)$ [12, 13] и $hi = \frac{h^2}{\sum_{i=1}^{h} A_i}$ [14].

sN*-индекс предложен О.В. Михайловым [15] для возможности учета позиции автора k_i в списке соавторов каждой статьи:

$$sN^* = \max\left(x : C_x \frac{\sqrt{A_i - k_i + 1}}{\sum_{i=1}^{k} \sqrt{i}} \geq x \right).$$

hc-индекс. Учитывает изменение активности ученого, присваивая старым статьям меньший вес [16]:

$$hc = \max(x : (C_x \cdot \frac{4}{Y_x + 1}) \geq x\ .$$

m-индекс. Определяет среднюю успешность автора [17]

$$m = \frac{h}{\max(Y_i + 1)}.$$

q2-индекс определяется как среднее геометрическое между h- и m-индексом [18]:

$$q2 = \sqrt{h \cdot m}\ .$$

i10-индекс: это количество статей, которые имеют не менее 10 цитирований.

Тренд h-индекса определяет тенденцию публикации статей автором [16]:

$$Tr = 4 * \sum_{i:C_i \geq h} (Y_i + 1)\ .$$

r-индекс и ar-индекс предложены в 2007 году для учета количества цитирований статей, входящих в h-индекс автора с учетом и без учета количества соавторов [19]:

$$r = \sqrt{\sum_{i=1}^{h} C_i}\ ,\ ar = \sqrt{\sum_{i=1}^{h} \frac{C_i}{A_i}}\ .$$

ra-индекс дополняет r-индекс учетом давности публикаций:

$$ra = \sqrt{\sum_{i=1}^{h} \frac{C_i}{Y_i + 1}}\ .$$

Индексы для организаций

i-индекс. Это аналог персонального индекса Хирша для ученых. Научная организация имеет индекс i, если не менее i учёных из этой организации имеют h-индекс не менее i.

Комплексный балл публикационной результативности (КБПР) введен для оценки деятельности организаций:

$$КБПР = \sum_{i=1}^{N} K_i \frac{1}{A_i} \sum_{j=1}^{M_i} \frac{1}{w_{ij}} s_{ij}\ ,$$

где N – количество статей в организации, M_i – количество авторов в статье, w_{ij} – количество указанных автором аффилиаций, s_{ij} равно 1, если одна из аффилиаций совпадает с организацией, для которой проводится расчет, иначе равно 0. Коэффициент качества статьи K_i задается на основе квантиля журнала в WoS, наличия в Scopus, РИНЦ и списке ВАК.

Следует отметить один существенный недостаток большинства индексов, представленных выше: для всех соавторов статья учитывается одинаково. Это со-

здает существенное искажение данных для коллабораций, когда в список соавторов включаются сотни и, даже, тысячи ученых (Atlas collaboration, Cms collaboration, Alice collaboration, Ecoteam4 и другие). Кроме того, при публикации статьи в список соавторов могут по разным причинам включаться авторы, которые в реальности не принимали существенного участия в написании статьи и получении описанных в ней научных результатов. Все эти факторы приводят к искусственному увеличению количества соавторов. По данным наукометрической системы ИАС «ИСТИНА» [20], среднее количество соавторов в статьях без коллабораций в период 2012–2022 года увеличилось на 35% (Табл. 1). Также после внедрения системы автоматического подсчета рейтингов в три раза выросло количество соавторов, которые представляют на одной конференции более трех докладов.

Таблица 1. Распределение среднего числа соавторов статей за период 2010–2022 гг.

Год	Среднее число соавторов
2010	2.74
2011	2.88
2012	3.02
2013	3.05
2014	3.03
2015	3.13
2016	3.21
2017	3.18
2018	3.24
2019	3.34
2020	3.53
2021	3.65
2022	3.7

Поскольку каждый соавтор публикации получает за статью или тезисы докладов одинаковое количество баллов, представленные выше способы учета цитируемости автора приводят к искажению рейтинговых оценок. Кроме того, искусственное добавление соавторов в публикации существенно ухудшает качество работы алгоритмов тематического поиска авторов.

КРИТЕРИЙ ОЦЕНКИ ВКЛАДА АВТОРОВ

Одним из методов оценки вклада авторов в опубликованные им работы может являться оценка статистического распределения его позиции в списке соавторов статей. В работе [21] проводилась оценка достоверности гипотез о распределении авторов в библиографическом описании. Рассматривались два варианта: «Первый соавтор располагается по алфавиту» и «Первый соавтор располагается по вкладу в работу». Для оценки использовались данные наукометрической системы ИАС «ИСТИНА», к которой на настоящий момент подключено более 30 организаций, в том числе МГУ им. М.В. Ломоносова.

В работе [21] показано, что для статей, имеющих четырех и более соавторов, процент библиографических описаний с указанием первого автора на основе алфавита составляет менее 9%. Оценка производилась в предположении, что в библиографическом описании публикации при распределении соавторов по алфавиту первый соавтор всегда имеет наименьший лексический порядок, а при распределении по смыслу вероятность этого события обратно пропорциональна количеству соавторов в статье. Таким образом, долю статей, в которых позиция первого соавтора определяется лексикографическим порядком, можно подсчитать по следующей формуле

$$L_k = \frac{r_k - \frac{a_k - r_k}{k-1}}{a_k},$$

где

r_k – количество статей с правильным лексикографическим порядком для первого соавтора, имеющих k соавторов;

a_k – общее количество статей, имеющих k соавторов.

В таблице 2 приведены значения расчета по публикациям сотрудников МГУ им. М.В. Ломоносова, зарегистрированных в наукометрической системе ИАС «ИСТИНА».

Таблица 2. Доля статей с лексикографическим определением первого соавтора.

Количество авторов K	L_k
2	24%
3	16%
4	9%
5	6%
6	6%
7	3%

Этот факт необходимо учитывать как при разработке алгоритмов тематического анализа наукометрических данных [22], в том числе, алгоритмов поиска экспертов по заданным предметным областям [23, 24] и тематического поиска с использованием графов соавторства [25], так и при оценке научной деятельности ученого.

Одним из наиболее простых критериев такой оценки являются количество и процент статей, в которых автор является первым соавтором. Однако подобные критерии не чувствительны к количеству соавторов статьи. Для построения нового индекса оценки предъявлялись следующие требования:

для авторов, которые все статьи опубликовали без соавторов, индекс должен быть равен 1;

для авторов, которые никогда не являлись первыми соавторами в библиографическом списке, индекс должен быть равен 0;

для «первых» соавторов статей индекс должен увеличиваться с увеличением количества соавторов.

Рассматривались два варианта индексов

$$P(a)=\frac{\sum\limits_{d\in D_a}\left(A(d)\right)\cdot\left|\left\{d\in D_a \mid ord\left(a,d\right)=1\right\}\right|}{\left|D_a\right|^2} \text{ и } K(a)=\frac{\sum A(d)}{\left|D_a\right|},$$

где $A(d)$ – количество соавторов в статье d, D_a – множество статей автора a, $ord(a,d)$ – порядковый номер автора a в статье d.

Апробация предложенных индексов производилась на данных наукометрической системы ИАС «ИСТИНА» [20]. Для проверки зависимости индексов от первой буквы фамилии автора были рассчитаны индексы за 5 лет для всех сотрудников организаций, зарегистрированных в системе, вся группа сотрудников была разбита на подгруппы по первой букве фамилии и был подсчитан средний индекс в каждой подгруппе (табл. 3)

Таблица 3. Распределение индексов P и K по первой букве фамилии.

	P	K		P	K
А	1.54	1.35	П	1.2	0.99
Б	1.44	1.2	Р	1.21	1.02
В	1.3	1.12	С	1.16	1
Г	1.45	1.17	Т	1.15	0.98
Д	1.3	1.1	У	1.24	1.01
Е	1.28	1.12	Ф	1.22	1.05
Ж	1.22	0.99	Х	1.21	1.02
З	1.31	1.08	Ц	1.02	0.86
И	1.21	1.02	Ч	1.22	1.01
К	1.22	1.03	Ш	1.14	0.93
Л	1.16	1	Щ	1.08	0.93
М	1.22	1.02	Э	1.2	1
Н	1.23	1.04	Ю	1.09	0.95
О	1.26	1.09	Я	1.17	1.03

Из приведенной таблицы можно видеть, что зависимость от первой буквы фамилии существует, но разница между сумами первых трех букв алфавита и последних трех букв алфавита составляет менее 30%. Можно сделать вывод, что индивидуальная оценка по такому индексу не является корректной, однако ее можно использовать для агрегированных оценок по научным коллективам в целом.

Принцип такой оценки аналогичен методу проведения тайного голосования с открытыми ответами, при котором каждый участник кидает кубик, и, если выпадает нечетное число, отвечает «Нет», если выпадает 6, отвечает правду, в остальных случаях отвечает «Да». По полученным ответам невозможно оценить

каждого человека по отдельности, но анализ статистики ответов позволяет достаточно точно оценить общую ситуацию.

ПРОГРАММНАЯ РЕАЛИЗАЦИЯ

Программная реализация аналитических отчетов для получения агрегированных оценок различных срезов выполнена с использованием модуля SQLReport [26]. Срезы могут выполняться по подразделениям, должностям, научным коллективам и другим достаточно ретроспективным признакам.

В табл. 4 приведено распределение в процентах количества сотрудников по должностям в зависимости от значения индекса K. В столбце «<0.25» указана доля занимающих эту должность сотрудников, для которых значения индекса K не превышает 0.25, в столбце «0.25–0.5» указана доля сотрудников, для которых значение индекса K находится в диапазоне от 0.25 до 0.5, и так далее.

Таблица 4. Гистограмма распределения для должностей.

Должность	**<0.25**	**0.25–0.5**	**0.5–0.75**	**0.75–1**	**1–1.5**	**1.5–2**	**2–3**	**>3**
Аспирант	20	4	8	7	35	8	11	8
Ассистент	17	7	12	13	33	6	7	5
Ведущий научный	22	11	11	12	23	10	7	4
Ведущий специалист	41	5	6	7	15	8	9	8
Доцент	14	5	9	14	42	8	5	2
Заведующий кафедрой	14	6	6	12	39	14	5	2
Заведующий лабораторией	24	11	11	17	18	7	6	6
Инженер 1-ой категории	30	6	6	10	17	11	11	9
Лаборант	28	4	4	6	25	6	17	10
Магистрант	17	2	5	8	46	6	7	9
Младший научный	19	6	8	9	22	11	14	11
Научный сотрудник	23	9	11	10	20	9	12	6
Преподаватель	7	2	9	15	58	4	4	1
Профессор	17	7	8	11	39	10	6	2
Специалист	30	7	11	6	16	6	11	12
Старший научный	22	10	13	10	22	11	8	4
Старший преподаватель	16	5	10	13	48	5	3	1

На основе анализа приведенных данных важно отметить, что для всех должностей выделяются два значимых столбца. Столбец $K \in [1,1.5)$ – сотрудники, фамилии которых оказываются в библиографическом списке на первой позиции немного чаще, чем при случайном распределении. $K < 0.25$ – сотрудники, которые

имеют статьи, но почти никогда или почти никогда не оказываются на первом месте в библиографическом списке статей. Для должностей из категории научно-вспомогательного персонала («Ведущий специалист», «Инженер 1-ой категории», «Лаборант», «Специалист») высокое значение показателя в этом столбце является вполне закономерным, поскольку эти сотрудники, как правило, обеспечивают техническое сопровождение научных исследований, но не проводят их самостоятельно. Однако для остальных должностей наличие больших значений в столбце $K < 0.25$ может означать наличие определенных проблем с учетом публикаций и, как следствие, с распределением финансирования в соответствии с публикационной активностью сотрудников.

Следует отметить, что 1013 сотрудников научного и профессорского состава из общего количества 9 тысяч имеют более 2 статей (в среднем по 18 статей) за последние 5 лет, но при этом они ни разу не являлись первым соавтором или единственным автором какой-либо публикации. Эти данные совместно с данными таблицы 1 о росте соавторов необходимо учитывать при построении наукометрических оценок для анализа результатов деятельности научной организации.

ЗАКЛЮЧЕНИЕ

Индексы, предложенные в настоящей работе, позволяют оценить степень участия авторов в написании статей. В силу наличия определенного влияния фамилии автора на значение индекса не рекомендуется применять подобные оценки для построения персональных рейтингов и проведения конкурсов. Однако эти индексы могут применяться для построения аналитических срезов с агрегированием данных по различным условиям, а также для уточнения результатов тематического поиска, в том числе, с использованием графов соавторства.

СПИСОК ЛИТЕРАТУРЫ

1. *Harzing A.-W.* The Publish or Perish Book // URL: https://harzing.com/popbook/ch16_3_1.htm (дата обращения: 11.09.2023)

2. Наукометрические показатели для авторов и организаций // URL: https://science.bsu.by/index.php/info/indexes/h-index (дата обращения: 10.04.2023)

3. h-index and Variants // URL: https://sci2s.ugr.es/hindex (дата обращения: 11.09.2023)

4. *Hirsch J.E.* An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences of the USA. V. 102, No. 46. P. 16569–16572.

5. *Eck N.J., Waltman L.* Generalizing the H- and G-Indices // ERIM Report Series Reference No. ERS-2008-049-LIS. URL: https://ssrn.com/abstract=1331777

6. *Kosmulski M. A* new Hirsch-type index saves time and works equally well as the original h-index // ISSI Newsletter. 2006. V. 2, No. 3. P. 4–6.

7. *Jin B.H., Liang L.M., Rousseau R., Egghe L.* The R- and AR-indices: Complementing the h-index // Chinese Science Bulletin. 2007. V. 52. No. 6. P. 855–863.

8. *Tol Richard*. A Rational, Successive g-Index Applied to Economics Departments in Ireland // Journal of Informetrics. 2008. No. 2. P. 149–155.

9. *Alonso S, Cabrerizo F.J, Herrera-Viedma E, Herrera F.* hg-index: A new index to characterize the scientific output of researchers based on the h- and g- indices // Scientometrics. 2010. V. 82(2). P. 391–400.

10. *Герасименко П.В.* Моделирование показателей результатов творческой деятельности ученого по его публикациям и их цитированиям // Автоматика на транспорте. 2019. №4. С. 493–504
URL: https://cyberleninka.ru/article/n/modelirovanie-pokazateley-rezultatov-tvorcheskoy-deyatelnosti-uchenogo-po-ego-publikatsiyam-i-ih-tsitirovaniyam (дата обращения: 11.09.2023).

11. *Назаров А.Д., Благинин В.А., Куликова Е.С.* Разработка модели интегрального наукометрического показателя публикационной активности ученых российских вузов // Московский экономический журнал. 2017. № 3.

URL: https://qje.su/otraslevaya-i-regionalnaya-ekonomika/moskovskij-ekonomicheskij-zhurnal-3-2017-6/

12. *Batista P.D., Campiteli M.G., Kinouchi O.* Is it possible to compare researchers with different scientific interests? // Scientometrics. 2006. V. 68(1). P. 179–189.

13. *Bi H.H.* Four problems of the h-index for assessing the research productivity and impact of individual authors // Scientometrics. 2023. V. 128. P. 2677–2691.

14. *Schreiber M.* A modification of the h-index: The hm-index accounts for multi-authored manuscripts // Journal of Informetrics. 2008. V. 2(3). P. C. 211–216.

15. *Михайлов О.В.*, Новая версия h-индекса с учетом числа соавторов и порядка их перечисления в научной публикации // Социология науки и технологий. 2015. №2. С. 24–32.
URL: https://cyberleninka.ru/article/n/novaya-versiya-h-indeksa-s-uchetom-chisla-soavtorov-i-poryadka-ih-perechisleniya-v-nauchnoy-publikatsii (дата обращения: 11.09.2023).

16. *Sidiropoulos A., Katsaros D., Manolopoulos Y.* Generalized Hirsch h-index for disclosing latent facts in citation networks // Scientometrics. 2008. V. 72(2). P. 253–280.

17. *Hirsch J.E.* An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences. 2005. V. 102. P. 16569–16572.

18. *Cabrerizo F.J., Alonso S., Herrera-Viedma E., Herrera F.* q2-Index: Quantitative and Qualitative Evaluation Based on the Number and Impact of Papers in the Hirsch Core // Journal of Informetrics. 2009. V. 4(1). P. 23–28.

19. *Jin B.H., Liang L.M., Rousseau R., Egghe L*. The R- and AR-indices: Complementing the h-index // Chinese Science Bulletin. 2007. V. 52(6). P. 855–863.

20. *Васенин В.А., Занчурин М.А., Козицын А.С., Кривчиков М.А., Шачнев Д.А.* Архитектурно-технологические аспекты разработки и сопровождения больших информационно-аналитических систем в сфере науки и образования // Программная инженерия. 2017. Т. 8, № 10. С. 448–455.

21. *Козицын А.С., Афонин С.А., Шачнев Д.А.* Метод оценки тематической близости научных журналов // Программная инженерия. 2020. № 6. С. 335–341.

22. *Козицын А.С.* Алгоритмы тематического поиска данных в наукометрических системах // Программная инженерия. 2022. Т. 13, № 6. С. 291–300.

23. *Shachnev D.A.* Searching for activity results and experts in a given subject area, taking results significance into account // Programmnaia inzheneriia. 2021. T. 12, № 5. S. 260–266.

24. *Козицын А.С., Афонин С.А., Шачнев Д.А.,* Алгоритм поиска по ключевым словам специалистов в заданной предметной области // Современные информационные технологии и ИТ-образование. 2021. Т. 17, № 1. С. 124–133.

25. *Козицын А.С., Афонин С.А., Шачнев Д.А.* Метод оценки тематической близости научных журналов // Программная инженерия. 2020. № 6. С. 335–341.

26. *Afonin S., Kozitsyn A., Astapov I.* Sqlreports: Yet another relational database reporting system // Proceedings of the 9th International Conference on Software Engineering and Applications. 2014. P. 529–534.

CITATION INDIXES AND ASSESSMENT OF AUTHORS PUBLICATION ACTIVITY

A. S. Kozitsyn[1] [0000-0002-8065-9061], S. A. Afonin[2] [0000-0003-3058-9269],
D. A. Shachnev[3] [0000-0002-5940-9180]

[1-3]*Institute of Mechcanics Lomonosov Moscow State University*

[1]*alexanderkz@mail.ru*, [2]serg@msu.ru, [3]mitya57@gmail.com

Abstract

One way to evaluate the scientific activity of a scientist is to calculate various indicators based on the number of his publications and their citation. In this case, each co-author of the publication receives the same number of points for it. A similar ranking method leads to an artificial increase in the number of co-authors in our publications. This leads to a distortion of ratings, and to a significant decrease in the quality of thematic search. The method presented in the paper allows us to evaluate the contribution of the author to his published works. Approbation of the method was done on the data of the scientometric system IAS ISTINA.

Keywords*: ranking, scientometrics, scientometric systems, co-authorship, citation systems, scientific rating*

REFERENCES

1. *Harzing A.-W.* The Publish or Perish Book // URL: https://harzing.com/popbook/ch16_3_1.htm

2. Naukometricheskie pokazateli dlya avtorov i organizacij // URL: https://science.bsu.by/index.php/info/indexes/h-index

3. h-index and Variants //URL: https://sci2s.ugr.es/hindex

4. *Hirsch J.E.* An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences of the USA. V. 102, No. 46. P. 16569–16572.

5. *Eck N.J., Waltman L.* Generalizing the H- and G-Indices // ERIM Report Series Reference No. ERS-2008-049-LIS. URL: https://ssrn.com/abstract=1331777

6. *Kosmulski M. A* new Hirsch-type index saves time and works equally well as the original h-index // ISSI Newsletter. 2006. V. 2, No. 3. P. 4–6.

7. *Jin B.H., Liang L.M., Rousseau R., Egghe L.* The R- and AR-indices: Complementing the h-index // Chinese Science Bulletin. 2007. V. 52. No. 6. P. 855–863.

8. *Tol Richard*. A Rational, Successive g-Index Applied to Economics Departments in Ireland // Journal of Informetrics. 2008. No. 2. P. 149–155.

9. *Alonso S, Cabrerizo F.J, Herrera-Viedma E, Herrera F.* hg-index: A new index to characterize the scientific output of researchers based on the h- and g- indices // Scientometrics. 2010. V. 82(2). P. 391–400.

10. *Gerasimenko P.V.* Modelirovanie pokazatelei rezultatov tvorcheskoi deiatelnosti uchenogo po ego publikatsiiam i ikh tsitirovaniiam // Avtomatika na transporte. 2019. №4. P. 493–504.
URL: https://cyberleninka.ru/article/n/modelirovanie-pokazateley-rezultatov-tvorcheskoy-deyatelnosti-uchenogo-po-ego-publikatsiyam-i-ih-tsitirovaniyam

11. *Nazarov A.D., Blaginin V.A., Kulikova E.S.* Razrabotka modeli integratsionnogo naukometricheskogo pokazatelia publikatsionnoi aktivnosti uchenykh rossiiskikh vuzov // Moskovskii ekonomicheskii zhurnal. 2017. No. 3.
URL: https://qje.su/otraslevaya-i-regionalnaya-ekonomika/moskovskij-ekonomicheskij-zhurnal-3-2017-6/

12. *Batista P.D., Campiteli M.G., Kinouchi O.* Is it possible to compare researchers with different scientific interests? // Scientometrics. 2006. V. 68(1). P. 179–189.

13. *Bi H.H.* Four problems of the h-index for assessing the research productivity and impact of individual authors // Scientometrics. 2023. V. 128. P. 2677–2691.

14. *Schreiber M.* A modification of the h-index: The hm-index accounts for multi-authored manuscripts // Journal of Informetrics. 2008. V. 2(3). P. C. 211–216.

15. *Mikhailov O.V*. Novaia versiia h-indeksa s uchetom chisla soavtorov i poriadka ikh perechisleniia v nauchnoi publikatsii // Sotsiologiia nauki i tekhnologii. 2015. №2. P. 24–32.
URL: https://cyberleninka.ru/article/n/novaya-versiya-h-indeksa-s-uchetom-chisla-soavtorov-i-poryadka-ih-perechisleniya-v-nauchnoy-publikatsii

16. *Sidiropoulos A., Katsaros D., Manolopoulos Y.* Generalized Hirsch h-index for disclosing latent facts in citation networks // Scientometrics. 2008. V. 72(2). P. 253–280.

17. *Hirsch J.E.* An index to quantify an individual's scientific research output // Proceedings of the National Academy of Sciences. 2005. V. 102. P. 16569–16572.

18. *Cabrerizo F.J., Alonso S., Herrera-Viedma E., Herrera F.* q2-Index: Quantitative and Qualitative Evaluation Based on the Number and Impact of Papers in the Hirsch Core // Journal of Informetrics. 2009. V. 4(1). P. 23–28.

19. *Jin B.H., Liang L.M., Rousseau R., Egghe L*. The R- and AR-indices: Complementing the h-index // Chinese Science Bulletin. 2007. V. 52(6). P. 855–863.

20. *Vasenin V.A., Zanchurin M.A., Kozitsyn A.S., Krivchikov M.A., Shachnev D.A.* Arkhitekturno-tekhnologicheskie aspekty razrabotki i soprovozhdeniia bolshikh informatsionno-analiticheskikh sistem v sfere nauki i obrazovaniia // Programmnaia inzheneriia. 2017. T. 8, № 10. S. 448–455.

21. *Kozitsyn A.S., Afonin S.A., Shachnev D.A.* Metod otsenki tematicheskoi blizosti nauchnykh zhurnalov // Programmnaia inzheneriia. 2020. № 6. S. 335–341.

22. *Kozitsyn A.S.* Algoritmy tematicheskogo poiska dannykh v naukometricheskikh sistemakh // Programmnaia inzheneriia. 2022. T. 13, № 6. S. 291–300.

23. *Shachnev D.A.* Searching for activity results and experts in a given subject area, taking results significance into account // Programmnaia inzheneriia. 2021. T. 12, № 5. S. 260–266.

24. *Kozitsyn A.S., Afonin S.A., Shachnev D.A*. Algoritm poiska po kliuchevym slovam spetsialistov v zadannoi predmetnoi oblasti // Sovremennye informatsionnye tekhnologii i IT-obrazovanie. 2021. T. 17, № 1. S. 124–133.

25. *Kozitsyn A.S., Afonin S.A., Shachnev D.A.* Metod otsenki tematicheskoi blizosti nauchnykh zhurnalov // Programmnaia inzheneriia. 2020. № 6. S. 335–341.

26. *Afonin S., Kozitsyn A., Astapov I.* Sqlreports: Yet another relational database reporting system // Proceedings of the 9th International Conference on Software Engineering and Applications. 2014. P. 529–534.

СВЕДЕНИЯ ОБ АВТОРАХ

КОЗИЦЫН Александр Сергеевич – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационного поиска и баз данных.

Alexander Sergeevich KOZITSYN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of information retrieval and database.

email: alexanderkz@mail.ru

ORCID: 0000-0002-8065-9061

АФОНИН Сергей Александрович – ведущий научный сотрудник, к. ф.-м. н., окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области регулярных языков и информационных систем.

Sergey Alexandrovich AFONIN – Leading Researcher, Ph.D., graduated from M.V. Lomonosov Moscow State University. Specialist in the field of regular languages and information systems.

email: serg@msu.ru

ORCID:0000-0003-3058-9269

Шачнев Дмитрий Алексеевич – программист, окончил мехмат МГУ им. М.В. Ломоносова. Специалист в области информационных систем.

Dmitry Alekseevich SHACHNEV – programmer, graduated from M.V. Lomonosov Moscow State University. Specialist in information systems.

email: mitya57@gmail.com

ORCID: 0000-0002-5940-9180

Материал поступил в редакцию 17 октября 2023 года

УДК 004.942

МЕТОДИКА СЕТЕВОГО АНАЛИЗА НАУЧНЫХ ПУБЛИКАЦИЙ

И. Г. Ольгина[0000-0002-9932-4552]

Омский государственный технический университет

inna_olgina@mail.ru

Аннотация

Актуальность вопросов анализа значимости научных публикаций обусловлена тем, что с появлением интернет-технологий стал возможен сбор данных о сети цитирования публикаций. Между тем, существующий сегодня подход к анализу значимости научных публикаций базируется на библиометрических показателях, учитывающих только количество цитирований. Однако все более широкое применение начинает получать сетевой анализ, применяемый преимущественно в исследованиях социальных сетей. Автором разработана методика, позволяющая осуществить эффективный анализ значимости научных публикаций, которая основана на методах сетевого анализа, альтернативных библиометрическим методам. В качестве критериев оценки значимости научных публикаций, основанных на сетевом анализе, установлены релевантные меры центральности узлов сети цитирования: центральность по степени связности; близости к другим узлам; посредничеству; авторитетности; концентрации. Приведен результат эксперимента, позволивший продемонстрировать адекватность разработанной методики анализа научных публикаций на основе сетевых метрик. В качестве первичных источников данных о публикациях использованы наукометрические базы данных, позволяющие отслеживать цитируемость публикаций и выявлять соответствующие сети цитирования. Применение предложенной методики способствует выявлению важных публикаций в развитии соответствующих научных направлений.

Ключевые слова*: сеть цитирования, публикации, наукометрия, библиометрический анализ, сетевой анализ, граф.*

ВВЕДЕНИЕ

С увеличением количества информации, представляемой в виде научных публикаций, появляется необходимость анализа разнообразных показателей уровня значимости этих источников, включая их информативность, авторитетность в научной среде, количества ссылок на них и т. д. На сегодняшний день существует очень большое количество международных систем цитирования (библиографических баз): РИНЦ, Web of Science, Scopus, Web of Knowledge, Chemical Abstracts, PubMed, Springer, Agris, GeoRef и др., которые используются для оценки уровня научных публикаций. При этом применяются признанные в мире библиометрические показатели, которые характеризуют авторов или журналы: импакт-фактор, индекс Хирша, индекс цитирования [1, 2, 3]. Но в расчете таких показателей и индексов учитывается только количество цитирований – для качественной оценки публикаций этого недостаточно.

Появление науки о сетях – Network Science позволяет исследовать сложные сетевые системы (в том числе социальные и информационные сети) посредством представления их в виде графовых моделей. Широкое применение приобрел метод сетевого анализа [4]. За последние десятилетия возрос интерес к науке о сетях, что повлекло за собой закономерное развитие всевозможных инструментов для исследований в данной области.

Рост в геометрической прогрессии общего объема публикаций обуславливает актуальность задач анализа взаимосвязей научных публикаций. В науке о сетях для решения этих задач разрабатываются модели и методы, относящиеся к сфере так называемых сетей цитирования. Однако оценить важность научных публикаций с учетом многоаспектного сетевого анализа не представляется возможным в силу отсутствия соответствующих инструментов. Исходя из этого, ставится задача комбинирования сетевых мер для выявления наиболее важных публикаций.

Существует различное программное обеспечение для визуализации и исследования сетей, например, VOSviewer (https://www.vosviewer.com/), Gephi (https://gephi.org/), Tom Sawyer Perspectives (http://www.tomsawyer.com/), Sentinel Visualizer (http://www.fmsasg.com/Products/SentinelVisualizer/). Эти инструменты широко применяются для исследования социальных сетей, в частно-

сти, сетей цитирования. Возможно построение сетей цитирования патентов [4], ключевых слов [5], научных публикаций [6], сетей авторского цитирования [7], сетей по совместному цитированию и соавторству [8, 9, 10]. Для анализа таких сетей разработано большое количество показателей (мер центральности), характеризующих значимость каждого узла в разных аспектах [11]. Исследование мер центральности представляет действительно большой интерес в исследованиях социальных сетей и сетей цитирования. По этой теме имеется достаточно много публикаций. В статье [12] дан обзор исследовательских работ по показателям центральности в социальных сетях.

На практике основной слабой стороной программных продуктов, названных выше, является отсутствие возможности определения суммарной оценки с учетом нескольких показателей. Это существенно ограничивает исследования сетей и визуализацию результатов с учетом многоаспектности. Таким образом, возникает потребность в решении задачи вывода и визуализации результатов многокритериальной оценки узлов сети.

Пример комбинирования методов для достижения лучшей производительности и получения более точных показателей по сравнению с традиционными методами продемонстрирован в работе [4]. Ее авторы объединили классический анализ основного пути (main path analysis) с алгоритмом PageRank и протестировали этот новый комбинированный метод на доступных данных о патентах. Стоит отметить, что анализ сети цитирования патентов является одним из наиболее важных методов измерения значимости интеллектуального анализа и идентификации содержания патентов.

ОПРЕДЕЛЕНИЕ ЗНАЧИМОСТИ ПУБЛИКАЦИИ НА ОСНОВЕ СЕТЕВЫХ МЕТРИК

Предметом настоящего исследования являются сети цитирования, которые представляют собой один из видов социальных сетей. Математическими моделями сетей цитирования являются ориентированные графы. Узлами сети являются научные публикации, а связями – коммуникации между ними, реализуемые путем цитирования [6]. Сеть цитирования научных публикаций представляется в виде ориентированного графа $G = (V, E)$, где V – множество вершин графа, а E – множество его дуг.

При вычислении значений параметров узлов сети, которыми являются меры центральности, удобно использовать ранжирование узлов сети по каждому из параметров в отдельности. Например, в статье [13] выполнено ранжирование коллекции периодических изданий на основе центральности по посредничеству и предложен алгоритм вычисления этой меры для взвешенных графов. В статье [14] подтверждена адекватность графа цитирования журналов цифровой библиотеки Math-Net.Ru как модели научных коммуникаций сравнением ранжирования журналов в графе цитирования с их рейтингом SCIENCE INDEX в eLIBRARY.RU. Ранжирование научных журналов осуществлялось по значению меры Page Rank.

По значениям мер центральности упорядочиваются по важности узлы сети при следующих ограничениях. Первое заключается в том, что центральность, которая оптимальна для одного приложения, часто не оптимальна для другого. Следовательно, не нужно использовать столько различных центральностей. Второе ограничение состоит в том, что центральность вершины отражает относительную важность вершин в графе. Меры центральности для измерения узлов в общем виде не разрабатывались [15].

При исследовании значимости публикации в соответствующей научной области с помощью сетевого подхода необходимо применить многокритериальный анализ. Следовательно, недостаточно оценить публикацию с помощью единственного показателя – одной из перечисленных мер центральности, а необходимо учесть важность соответствующего узла сети с помощью релевантных мер центральности *C*, исходя из целей исследования. Поэтому необходимо вычислить «обобщенный показатель важности», который представляет собой взвешенную сумму частных показателей C_i, в которую каждый из них входит с определенным весом k_i, отражающим его значимость:

$$C = k_1 C_1 + k_2 C_2 + \cdots + k_n C_n \,, \qquad (1)$$

где $0 \le k_i \le 1, i = \overline{1,n}\,; n \in \mathbf{N}$. Весовые коэффициенты, с которыми входят в расчет разные показатели, не постоянны, а изменяются в зависимости от ситуации.

МЕТОД ОПРЕДЕЛЕНИЯ ВАЖНОСТИ УЗЛА СЕТИ ЦИТИРОВАНИЯ

Данный метод можно также назвать методом комплексной оценки центральности узлов сети. В соответствии с представленным выше способом (формула (1)) строится математическая модель, которая позволяет определить комплексную оценку центральности узлов сети цитирования как сумму рангов узлов сети по каждой мере центральности с учетом их релевантности, по следующей формуле:

$$C_{sum}(h) = \sum_{i=1}^{n} k_i(h) R_i, \qquad (2)$$

где R_i – ранг *i*-го показателя; *n* – число показателей; k_i – весовой коэффициент – индекс релевантности *i*-го показателя, $0 \leq k_i \leq 1$; *h* – профиль исследования.

Для удобства проведения экспериментов в настоящей работе индекс релевантности *k*, соответствующий весу показателя, принимается равным 1, если данный показатель соответствует профилю исследования, и 0, если он не принимается во внимание, т. е. $k_i \in \{0, 1\}$.

ОТБОР ПОКАЗАТЕЛЕЙ ДЛЯ АНАЛИЗА СЕТЕЙ ЦИТИРОВАНИЯ НАУЧНЫХ ПУБЛИКАЦИЙ

Нами проведен анализ существующих мер центральностей узлов для социальных сетей. На этой основе выбраны пять показателей для анализа сетей цитирования: центральность по степени связности (degree centrality) [16]; центральность по близости к другим узлам (closeness centrality) [17]; центральность по посредничеству (betweenness centrality) [18]; центральность по авторитетности (authority centrality) и центральность по информативности (hub centrality) [19, 20]. В работе [21] приведено подробное описание этих мер центральности узлов сети. В [15] дана интерпретация значений каждой меры центральности относительно того, как она может повлиять на показатель важности публикации. Для проведения исследований по обнаружению зарождения инноваций в определенной научной области с помощью сетевого анализа применяют центральность по близости к другим узлам и центральность по посредничеству. Примером могут служить исследования [22]. Для исследования социальных сетей в целом применяют центральность по посредничеству [23]. Центральность по авто-

ритетности в социальных сетях применяют для рекомендаций пользователю, на кого стоит ориентироваться, как продемонстрировано в [24]. Если использовать этот показатель в контексте сетей цитирования, то можно интерпретировать его так: данная статья располагает важной информацией по определенной теме. Центральности по авторитетности и информативности фокусируются на структуре сети и определяют ее важность в соответствии с их позициями на графе [25] и зависят от общего количества отношений, которые имеются за пределами узла [24]. Узел является центром, когда он имеет ребра с авторитетными узлами.

Является важным, что в связи с тем, что граф сети цитирования является ориентированным, входящие и исходящие связи можно анализировать раздельно. При определении центральности по близости к другим узлам для орграфов будем рассматривать как дистанции от определяемой вершины до всех остальных (исходящие связи – режим *out*), так и дистанции от всех вершин до определяемой (входящие связи – режим *in*) [15]. В случае, когда цитируемым объектам придается большая значимость, имеет смысл пользоваться вторым определением. При рассмотрении центральности по степени связности аналогично рассматриваются раздельно входящие связи – (полустепень захода), которые соответствуют количеству цитирований рассматриваемой публикации, и исходящие связи (полустепень исхода), отражающие количество ссылок публикации на другие.

Стоит отметить, что при анализе социальных сетей часто применяют показатель PageRank. Эта центральность изначально применялась поисковой системой Google для ранжирования веб-страниц и рассчитывается по формуле

$$PR(A) = (1 - d) + d \sum_{i-1}^{n} \frac{PR(T_i)}{C(T_i)} \ , \qquad (3)$$

где *A* – оцениваемая веб-страница, T_i – веб-страницы, ссылающиеся на страницу *A*, *d* – коэффициент демпфирования (вероятность перехода по ссылке, имеющейся на странице *A*), *C(T)* – число ссылок на веб-странице *T*.

В случае анализа научных публикаций эту метрику можно не принимать во внимание в силу особенности данной меры и целей, для которых она создавалась. В формуле (3) расчета значений названной меры применяется коэффициент демпфирования *d*, который в сети цитирования может быть проинтерпретирован как

вероятность того, что читатель просмотрит источник, ссылка на который есть в публикации.

МЕТОДИКА СЕТЕВОГО АНАЛИЗА НАУЧНЫХ ПУБЛИКАЦИЙ

Предложим новую методику ранжирования публикаций на основе сетевых метрик, комбинация которых задает профиль ранжирования. Этот профиль определяется целью исследования.

1. Сбор данных о цитировании

Для проведения анализа научных публикаций необходимы данные о цитировании, которые можно получить следующими способами:

– сбор данных с использованием покластерного обхода всей сети цитирования (с помощью написанной программы);

– сбор подсети по заданной тематике;

– использование данных о сети цитирования публикаций определенной научной организации; например, в Омском государственном техническом университете создан с сервис science.omgtu.ru, содержащий необходимые данные для сетевого анализа публикаций всех авторов университета.

2. Выбор и формирование профиля исследования

Нами рассмотрены меры центральности узлов сети, которые следует использовать в качестве характеристик узлов сети цитирования научных публикаций.

Для математического описания задачи введем следующие обозначения:

$K = \{k_1, k_2, \ldots, k_n\}(n = 9)$ – кортеж коэффициентов, задающих степень влияния соответствующей центральности на профиль исследования (индекс релевантности *i*-го показателя);

$H = \{h_1, h_2, \ldots, h_m\}$ – множество профилей исследования, которые формируются, исходя из поставленных целей. К примеру, если важность публикации характеризуется ее высокой информативностью, то такие публикации будут содержать большое количество ссылок или иметь ссылки в основном на авторитетные публикации.

Для характеристики узлов сетей цитирования используем следующие сетевые метрики:

– k_1 – центральность по степени связности (degree centrality);

– k_2 – полустепень исхода (out-degree centrality);

– k_3 – полустепень захода (in-degree centrality);

– k_4 – центральность по близости к другим узлам (closeness centrality);

– k_5 – центральность по близости к другим узлам (closeness centrality в режиме *out*);

– k_6 – центральность по близости к другим узлам (closeness centrality в режиме *in*);

– k_7 – центральность по посредничеству (betweenness centrality);

– k_8 – центральность по авторитетности (authority centrality);

– k_9 – центральность по информативности (hub centrality).

Таким образом, k_i является весовым коэффициентом и принадлежит множеству рациональных чисел, $0 \leq k_i \leq 1, i = \overline{1,n}$.

Выбор значения k_i зависит от предпочтений лица, принимающего решение (ЛПР). Для простоты проведения экспериментальных исследований рекомендуется использовать k_i = 1, если показатель принимается во внимание, и k_i = 0 в противоположном случае.

Профиль исследования h_j формируется с учетом множества значений весовых коэффициентов K_j в виде кортежа (k_1, k_2, …,k_i), i = 9, где j = 9^2 при $k_i \in$ {0, 1}. Соответственно количество профилей исследования будет составлять 81 вариант.

ЛПР необходимо определить сочетание показателей, которые будут составлять профиль исследования, а также определить весовые коэффициенты данных показателей.

Одним из подходов к решению таких задач является принятие решения на основе методов ранговой корреляции Спирмена и Стьюдента [26, 27]. Методы ранговой корреляции могут быть использованы для определения тесноты связи не только между количественными, но и качественными переменными при условии, если их значения можно ранжировать и упорядочить. Согласно этим методам значения комплексной оценки узлов сети, полученные по формуле (2), сравниваются с известными альтернативными показателями, применяемыми в базах цитирования. Затем требуется выбрать профиль исследования, при кото-

ром значения коэффициента корреляции наибольшие, при условии отклонения гипотезы H_0 об отсутствии зависимости.

3. Алгоритм ранжирования научных публикаций по важности

Алгоритм, реализующий метод ранжирования публикаций по важности согласно профилю исследования на основе сетевого анализа, представлен на рисунке 1.

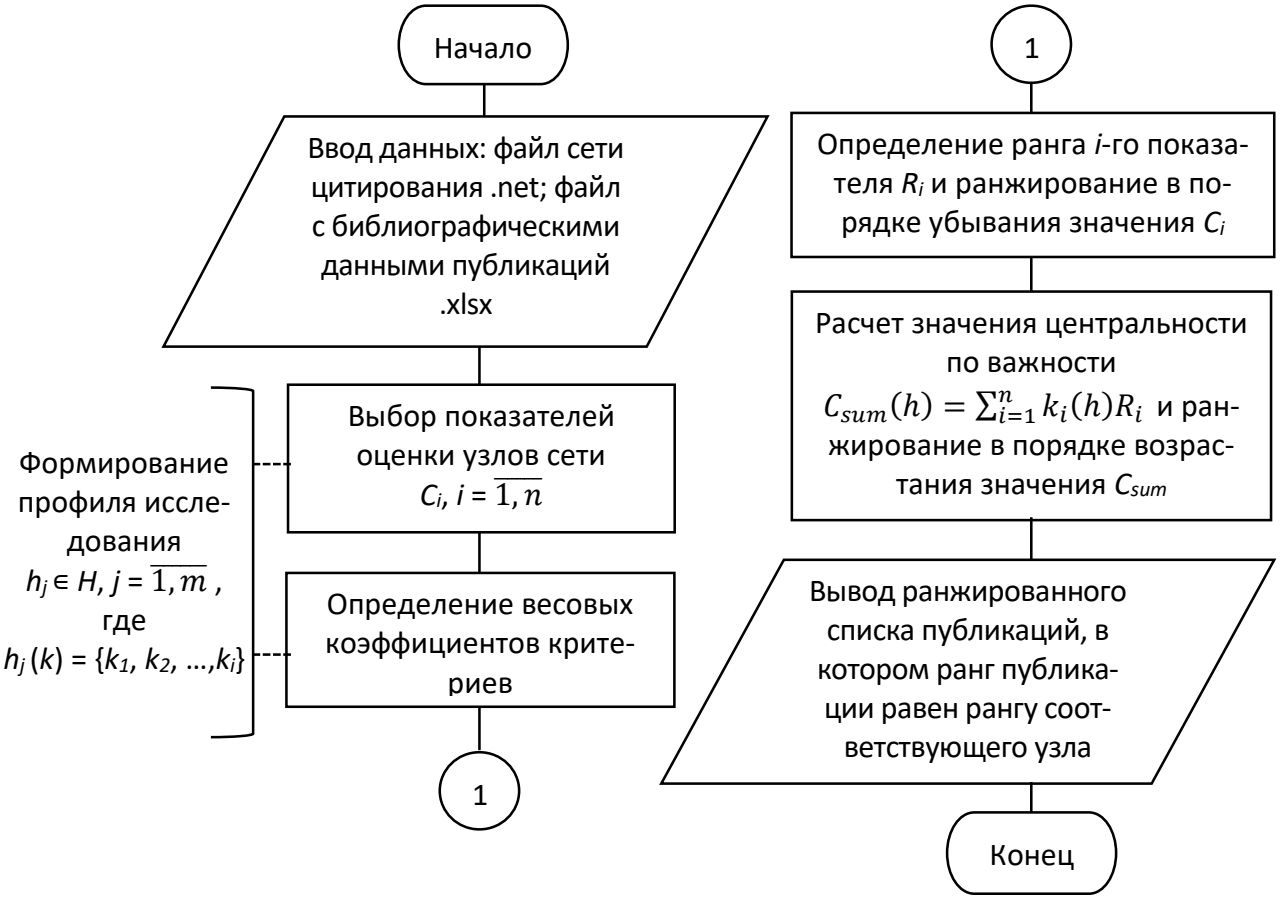

Рис. 1. Алгоритм ранжирования научных публикаций по важности

С целью реализации предложенной методики анализа сетей цитирования разработаны «Программный комплекс LinkAnalyzer 1.0 для сбора и анализа информации о цитировании научных публикаций» № 2020615709 от 29.05.2020 г.; «Генератор списка источников информации в сетях цитирования» № 2021661693 от 14.07.2021 г.; «Визуализатор сетей цитирования» № 2023666387 от 31.07.2023 г. Интерфейс программы для анализа научных публикаций представлен на рисунке 2.

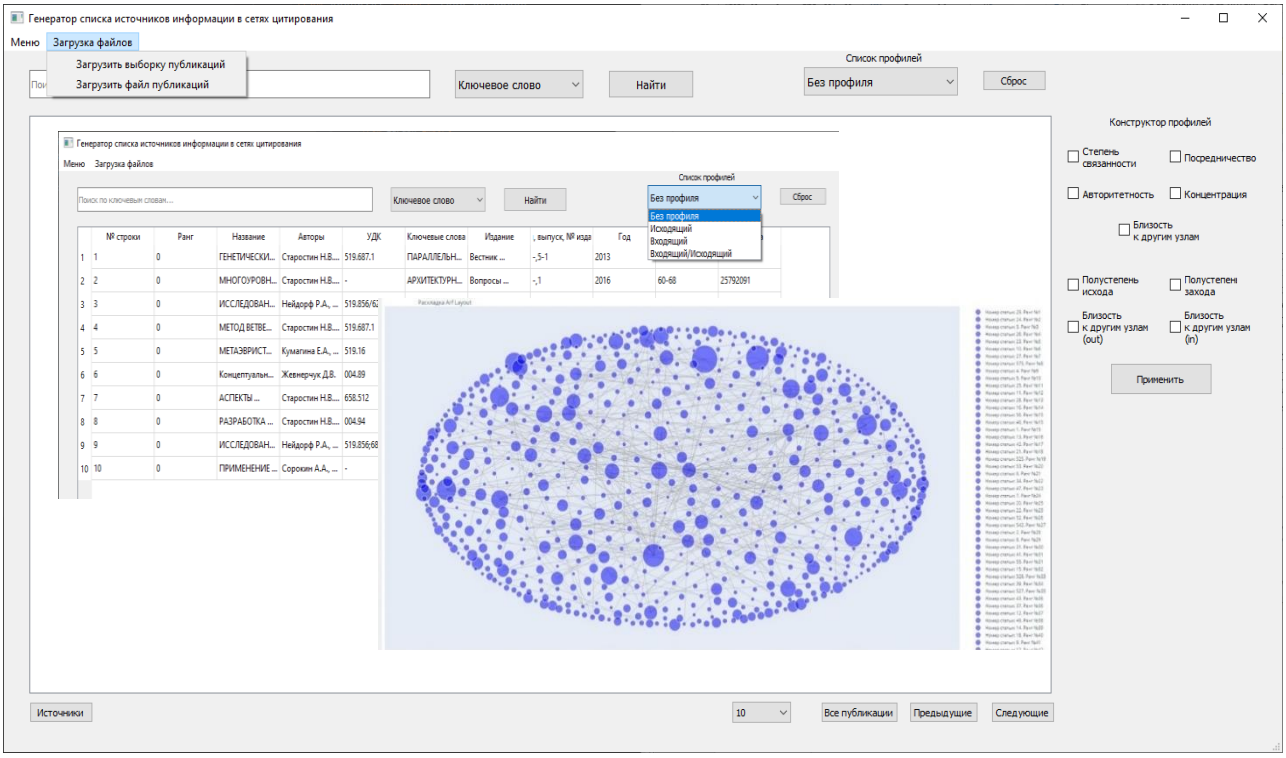

Рис. 2. Интерфейс программы «Генератор списка источников информации в сетях цитирования»

ВАРИАНТЫ ФОРМИРОВАНИЯ ПРОФИЛЯ ИССЛЕДОВАНИЯ

В зависимости от целей подбора и анализа публикаций в сети, используя различные комбинации параметров (мер центральностей), можно получить различные профили исследования *h* (формула (2)). При ранжировании публикаций по важности необходимо осуществить формирование профиля исследования, соответственно следует определить релевантные меры центральностей. Приведем примеры профилей исследования.

Профиль «Входящий/Исходящий». Включает в расчет показателя важности (C_{sum}) все меры центральностей, перечисленные в разделе «Отбор показателей для анализа сетей цитирования научных публикаций», без учета направленности графа. Выбор этого профиля позволяет отобрать публикации, которые могут быть как первоисточниками, так и обзорными, имеющими множество ссылок на другие публикации. Характеризует в целом уровень публикации, согласно предложенным критериям оценки с применением сетевого подхода.

Профиль «Исходящий». Включает в расчет показателя важности (C_{sum}) меры центральности, рассчитанные с учетом только исходящих связей, таких как полустепень исхода, центральность по близости к другим узлам в режиме *out*, центральность по информативности. Данный профиль может использоваться для отбора реферативных или обзорных публикаций.

Профиль «Входящий». Включает в расчет показателя важности (C_{sum}) меры центральности, рассчитанные с учетом только входящих связей, таких как полустепень захода, центральность по близости к другим узлам в режиме *in*, центральность по авторитетности. При выборе данного профиля представляется возможным отбор эмпирических публикаций, содержащих результаты оригинальных исследований.

Можно получить и другие профили, например, обратить в максимум только один показатель, а другие свести к минимуму, что характерно для любой сложной задачи исследования операций. В результате получим профиль «Авторитетность», ориентируясь только на меру центральности по авторитетности или профиль «Посредничество», учитывая меру центральности по посредничеству, которая управляет информацией среди других вершин графа сети через соединительный путь. При комбинировании двух показателей (меры центральности по близости к другим узлам и меры центральности по посредничеству) получим оценку важности узла относительно его положения в сети. Эти две меры центральности относят к геометрическим. Центральность по близости к другим узлам основана на кратчайших путях в графе и имеет простой физический смысл: чем меньше расстояния от исследуемой вершины до остальных вершин графа, тем больше будет значение самой центральности. Центральность по посредничеству позволяет достаточно хорошо определять «узкие места» в графе – вершины, входящие в состав ребра или набора ребер, соединяющих два ярко выраженных кластера.

РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

В настоящем разделе представлены полученные результаты анализа публикаций на основе данных международной базы цитирования Scopus. Сеть цитирования содержит статьи, изданные по 2023 год. Из данных о цитировании публикаций базы данных Scopus осуществлена выборка по ключевому слову

«Network Science». На основе полученных данных построен ориентированный граф, который содержит 850 вершин и 2169 дуг. На рисунке 3 представлен этот граф, для его визуализации использована программа VOSviewer.

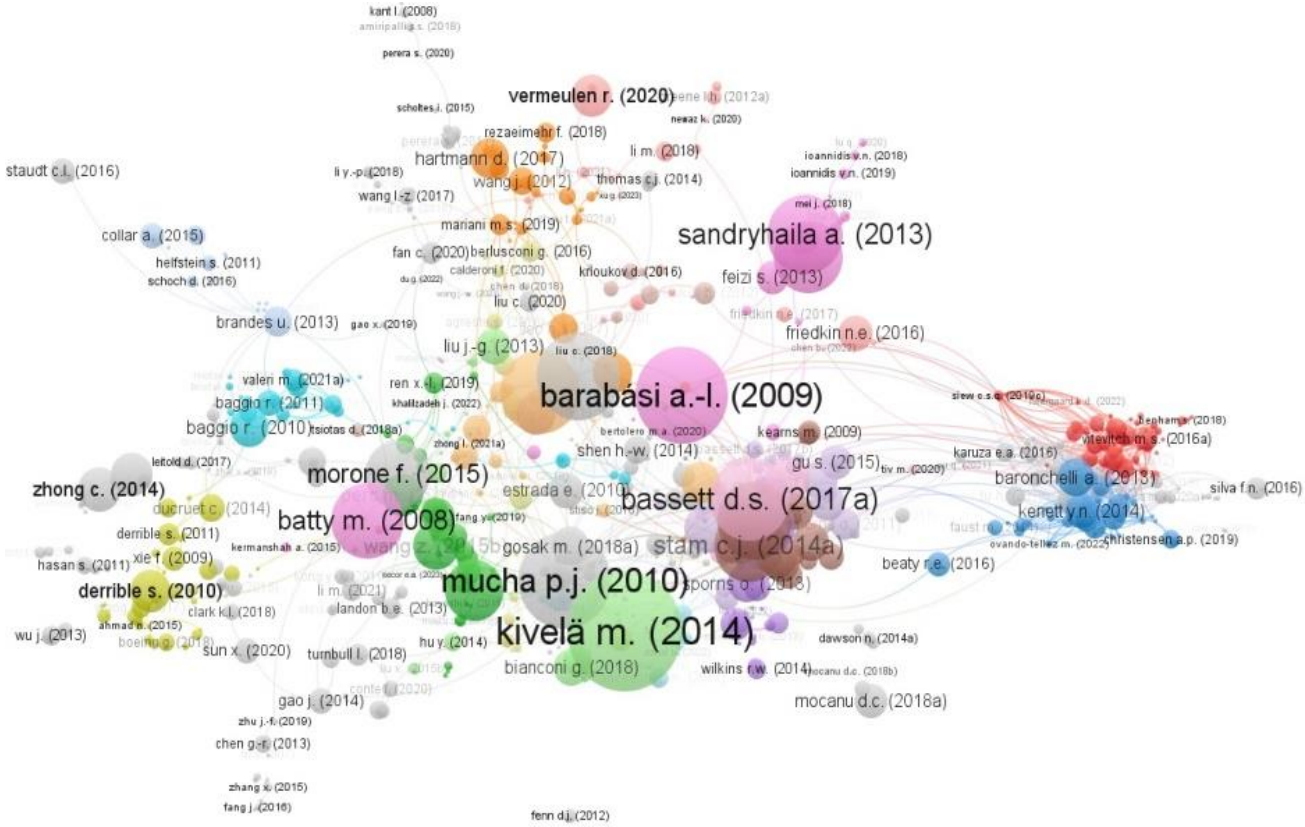

Рис. 3. Сеть цитирования по ключевому слову «Network Science»

На рисунке 4 представлен укрупненный фрагмент кластера данной сети цитирования, на который стоит обратить внимание. Из предыдущего изображения видно, что этот фрагмент даже визуально можно выделить в отдельный кластер. Диаметр узла сети пропорционален показателю количества цитирований публикации на рисунке 4 (входящие связи). На данном кластере произведем подробный анализ публикаций с помощь комбинирования показателей центральности узлов сети.

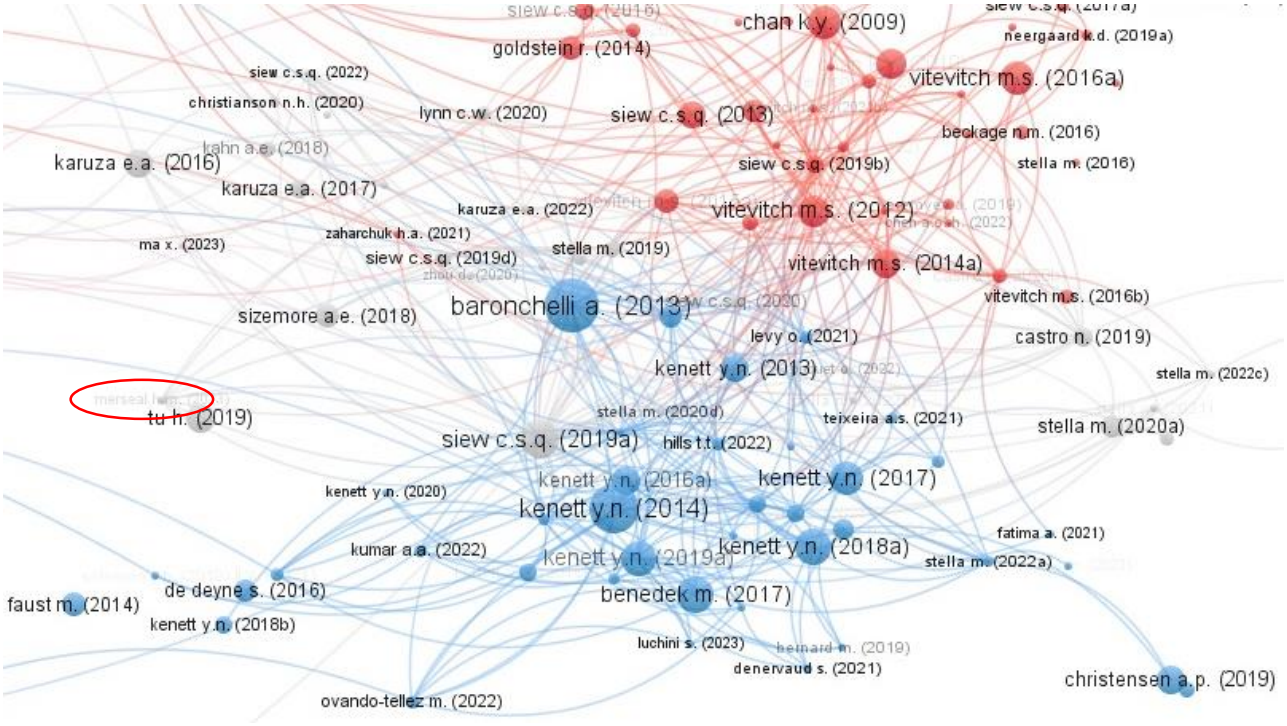

Рис. 4. Фрагмент сети цитирования по ключевому слову «Network Science»

Стоит обратить внимание на статью 2023 года автора Merseal H.M. (на рисунке 3 данный узел обведен красным маркером). Если использовать только возможности программы VOSviewer для исследования и визуализации сетей цитирования, то эту публикацию можно упустить из виду.

Выделенная публикация еще не имеет большого количества цитирований в связи с тем, что она только что опубликована. Однако она имеет важное значение в изучении и развитии науки о сетях. Тема данной статьи: «Представление мелодических взаимосвязей с использованием сетевой науки». Исходя из содержания публикаций, на которые она ссылается, можно судить о широком применении сетевого анализа и развитии науки о сетях (рисунок 5). Она цитирует авторитетные статьи, получившие большое количество цитирований, по следующим направлениям исследования:

- как сетевая наука может пролить свет на наше понимание познания;
- карты, транспортные средства и скайхуки в когнитивной сетевой науке;
- что сетевая наука может рассказать нам о фонологии и обработке языка;
- вклад современной сетевой науки в когнитивные науки;

– использование сетевой науки для понимания лексикона старения: связывание индивидуального опыта, семантических сетей и когнитивных способностей;

– значение статистического обучения для когнитивной сетевой науки.

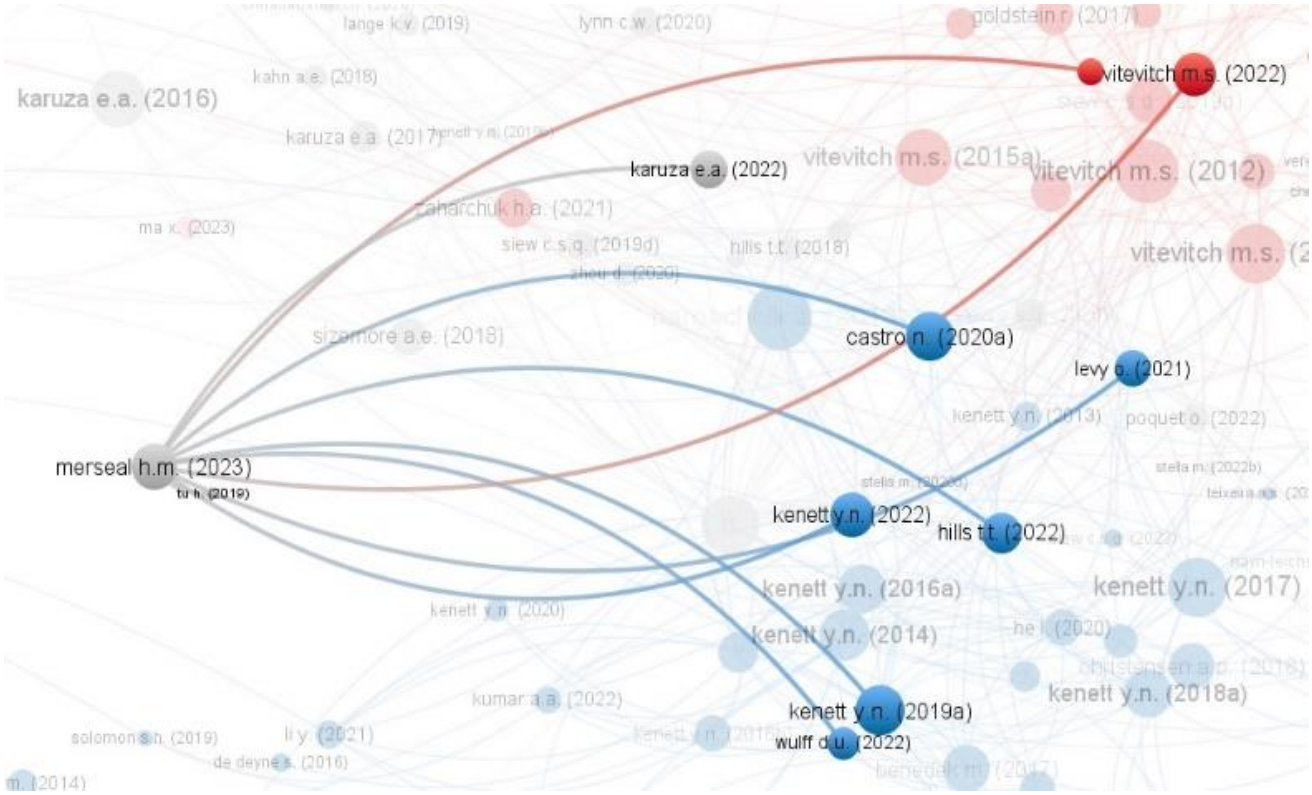

Рис. 5. Ссылки автора H.M. Merseal на публикации

При изучении данной сети цитирования с помощью разработанной нами технологии, лежащей в основе программного комплекса анализа сетей цитирования, подобные статьи легко обнаружить. При использовании профиля исследования «Входящий/Исходящий» данная публикация получила 3 ранг из 390 (таблица 1).

В таблице 1 публикации проранжированы по убыванию их важности. Результат визуализирован с помощью разработанной программы визуализации графа цитирования и представлен на рисунке 6. Рассматриваемая публикация выделена красным маркером. Программа позволяет масштабировать размер вершины с учетом ранга публикации, полученного в результате комплексной оценки, с использованием математической модели (формула (2)). Радиус вер-

шины вычисляется через функцию натурального логарифма, что позволяет делать самые значимые вершины визуально различимыми. На данном рисунке показаны только те узлы, которые имеют ранг ≥ 50, остальные вершины скрыты. Для демонстрации отображены подписи узлов с 1 по 3 ранг.

Таблица 1. Ранжированный список публикаций согласно профилю исследования «Входящий/Исходящий»

№ строки	Ранг	Название
468	1	Baronchelli A. (2013)
563	2	Karuza E.A. (2016)
369	**3**	**Merseal H.M. (2023)**
239	4	Kenett Y.N. (2022)
590	5	Vitevitch M.S. (2015)
466	6	Medaglia J.D. (2015)
440	7	Bassett D.S. (2017)
446	8	Stam C.J. (2014)
438	9	Mucha P.J. (2010)
535	10	Vitevitch M.S. (2012)
…	…	…
702	390	Leitold D. (2018)

С помощью данной программы визуализировать граф и осуществить масштабирование вершины графа согласно занимаемому публикацией рангу возможно на основе любого выбранного профиля исследования.

В таблице 2 представлены результаты ранжирования публикаций согласно нескольким профилям исследования. Для удобства описания название вершины заменено обозначением – S_i. С целью сравнения и анализа взяты первые три публикации из таблицы 1, занимающие 1, 2 и 3 ранги соответственно по профилю «Входящий/Исходящий» [28–30], и публикация [31] известного ученого по сетевой науке – Network Science, автора книги «Наука о сетях» (Cambridge, 2016).

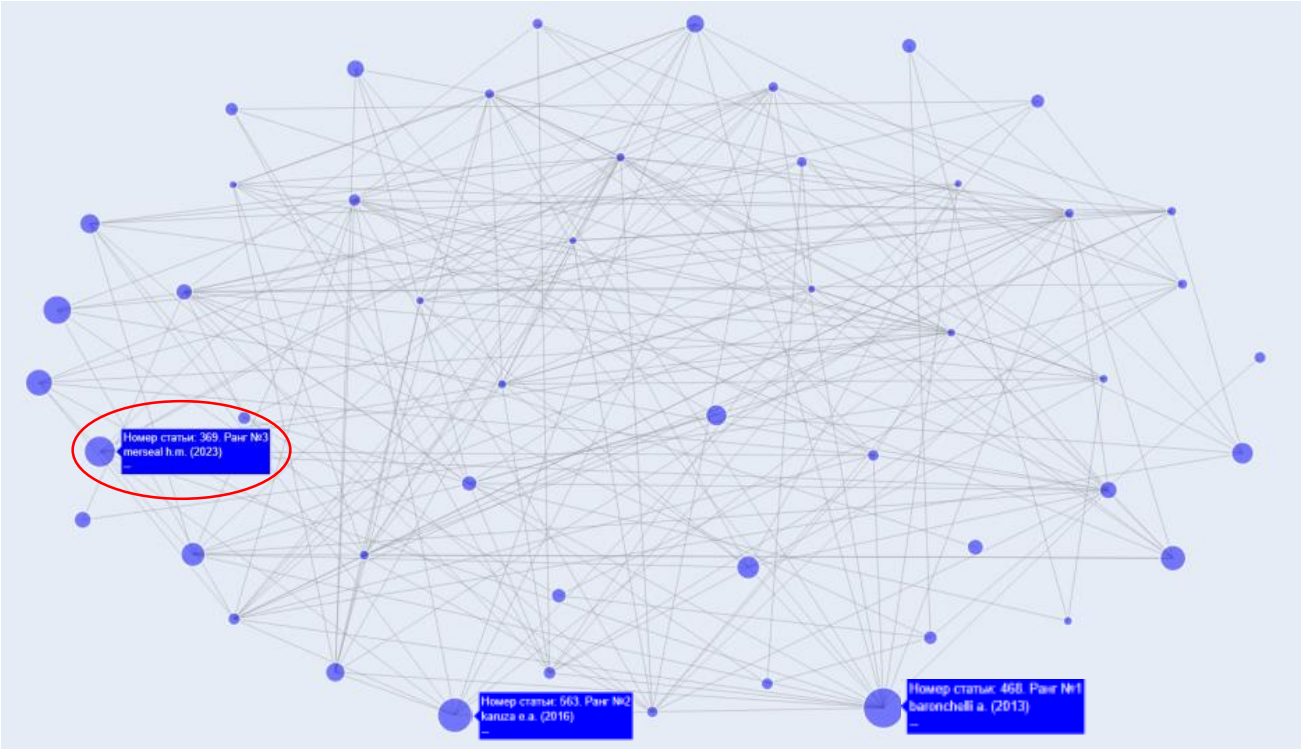

Рис. 6. Масштабирование вершины графа с учетом комплексной оценки (C_{sum}) центральности узлов сети согласно профилю «Входящий/Исходящий»

Таблица 2. Ранги публикаций по профилям исследования

Профиль исследования	Ранг вершин графа			
	Baronchelli A. (2013), S_1	Karuza E.A. (2016), S_2	Merseal H.M. (2023), S_3	Barabasi A.-L. (2009), S_4
Входящий/ Исходящий	1	2	3	120
Входящий	46	91	81	7
Исходящий	98	163	82	38
Авторитетность	4	7	37	76
Информативность	6	35	16	93
Степень связности	4	9	21	35
Полустепень захода	3	6	19	25
Полустепень исхода	7	14	15	23
Посредничество	17	3	12	137
Близости к другим узлам	85	78	56	393

Публикация S_4 занимает высокий ранг по профилю «Входящий», что говорит о важности данной работы как первоисточника информации о принятии теории сетей в качестве общей парадигмы, развитии науки о сетях как новой области исследований с особым набором задач и достижений. Ссылки на Albert-Laszlo Barabasi как авторитетного ученого имеются во всех трех публикациях, представленных в таблице. По сравнению с рассматриваемыми публикациями S_3 имеет наивысший ранг по профилю «Близость к другим узлам», что может характеризовать важность статьи относительно ее положения в исследуемом фрагменте графа. Количество ссылок в публикации S_1 составляет 174 источника, S_2 – 108 источников, S_3 – 101 источник, S_4 – всего 11. Публикация S_3 ссылается на S_1, S_2 и S_3 и занимает достаточно высокий ранг по профилю «Информативность». Однако важность публикации по информативности у публикации S_3 значительно выше, чем у S_2, это говорит о том, что S_3 ссылается на большее количество авторитетных статей. Значение меры центральности по информативности зависит от количества ребер, исходящих из исследуемой вершины, соединяющих вершины с высоким значением меры по авторитетности, и учитывает глубину связей, т. е. насколько эти вершины являются информативными. Публикация S_1 имеет высокий ранг по профилю «Авторитетность», что свидетельствует о том, что данная работа цитируется авторитетными авторами. Стоит отметить, что показатели рассчитаны, исходя из данных собранного фрагмента сети цитирования. В данном случае ограничивается данными о цитировании, полученными из базы данных Scopus.

Следовательно, можно сделать следующий вывод: исследуя топологию сети и применяя комбинированный метод анализа сетей цитирования, предложенный в данной работе, можно определить уровень научных публикаций в разных аспектах анализа, что дает представление о вкладе публикаций в исследуемой области науки.

ЗАКЛЮЧЕНИЕ

Выполнено систематическое исследование используемых в современной практике способов анализа важности научных публикаций. Отметим, что применение инструментов анализа научных публикаций на основе топологических свойств сети дает возможность осуществить их эффективный анализ. Для проведения экспериментальных исследований построена модель реальной сети цитирования публикаций. В связи с тем, что для любой прикладной задачи нахождения оптимального решения определяется свой набор критериев, разработана методика сетевого анализа научных публикаций на основе комплексной оценки узлов сети с учетом релевантных мер центральности, исходя из целей исследования. Внедрение подобных инструментов в наукометрические базы данных даст возможность осуществлять оценку публикаций, определяя их важность, не только на основе библиометрических показателей, в определении которых учитывается лишь количество цитирований. Ведь только библиометрические данные не могут служить критерием эффективности исследований или ценности публикаций.

Использование предложенных методов и алгоритмов анализа значимости научных публикаций позволяет повысить эффективность решения множества прикладных задач. Появилась возможность очень быстро проводить библиографические исследования для составления списка наиболее значимых литературных источников по какой-либо сфере научных исследований, определенной предметной области, по конкретной учебной дисциплине, составления литературного обзора и др. Использование предложенной методики применимо для: оптимизации комплектования литературой научных библиотек; подбора различных видов документов, в том числе патентов и патентных ландшафтов (объектов интеллектуальной собственности); поиска потенциальных соавторов или научных сообществ в исследуемой области знаний; наукометрии и инфометрики; управления научной и инновационной политикой.

Благодарность

Автор выражает благодарность Евгению Борисовичу Юдину, к. т. н., доценту кафедры «Математические методы и информационные технологии в экономике», начальнику Управления научной информацией ОмГТУ, за важнейшие советы при проведении исследования и оформления статьи.

СПИСОК ЛИТЕРАТУРЫ

1. *Гонашвили А.С.* Наукометрические базы данных и работа с ними : науч.-метод. пособие / ун-т при Межпарламент. ассамблее ЕврАзЭС. СПб.: Изд-во ун-та при МПА ЕврАзЭС, 2020. 57 с.

2. H-index. URL: https://en.wikipedia.org/wiki/H-index (дата обращения: 15.04.2023).

3. Impact factor. URL: https://ru.wikipedia.org/wiki/Impact factor (дата обращения: 15.04.2023).

4. *Lu Z., Ma Y., Song L*. Patent Citation Network Analysis Based on Improved Main Path Analysis: Mapping Key Technology Trajectory // Advances in Artificial Intelligence and Security (ICAIS 2021): Communications in Computer and Information Science. Springer: Cham, 2021. Vol. 1423. P. 158–171. https://doi.org/10.1007/978-3-030-78618-2_13.

5. *Wang J., Cheng Q., Lu W. et al.* A term function–aware keyword citation network method for science mapping analysis // Information Processing & Management. Vol. 60, no. 4. P. 103405. https://doi.org/10.1016/j.ipm.2023.103405.

6. *Ольгина И.Г., Пронин И.В., Абдрахманов А.Н.* Построение графовых моделей сети цитирования научных публикаций // Системы управления, информационные технологии и математическое моделирование: материалы II Всерос. науч.-практ. конф. с междунар. участием (Омск, 19–20 мая 2020 г.). Омск: ОмГТУ, 2020. Т. I. С. 118–125.

7. *Zhao F., Zhang Y., Lu J. et al.* Measuring academic influence using heterogeneous author-citation networks // Scientometrics. 2019. Vol. 118. P. 1119–1140. https://doi.org/10.1016/10.1007/s11192-019-03010-5.

8. *Ji P., Jin J., Ke Z. T., L. W.* Co-citation and Co-authorship Networks of Statisticians // Journal of Business & Economic Statistics. 2022. Vol. 40, no. 2. P. 469–485.

https://doi.org/10.1080/07350015.2021.1978469.

9. *Luc P.T., Lan P.X., Le A.N.H., Tran B.T.* A Co-Citation and Co-Word Analysis of Social Entrepreneurship Research // Journal of Social Entrepreneurship. 2022. Vol. 13, No. 3. P. 324–339. https://doi.org/10.1080/19420676.2020.1782971.

10. *Печников А.А., Чебуков Д.Е.* Анализ соавторства в математических журналах Math-Net.Ru // Научный сервис в сети Интернет: тр. XXIV Всерос. науч. конф. (19-22 сент. 2022 г.). М.: ИПМ им. М.В. Келдыша, 2022. С. 190-202. https://doi.org/10.20948/abrau-2022-5.

11. *Gómez S*. Centrality in Networks: Finding the Most Important Nodes // Business and Consumer Analytics: New Ideas / P. Moscato, N. Jane de Vries. Springer: Cham, 2019. P. 401–433. https://doi.org/10.1007/978-3-030-06222-4_8.

12. *Das K., Samanta S., Pa M.* Study on centrality measures in social networks: a survey // Social Network Analysis and Mining. 2018. Vol. 8. P. 13. https://doi.org/10.1007/s13278-018-0493-2.

13. *Бредихин С.В., Ляпунов В.М., Щербакова Н.Г.* Мера важности научной периодики – «центральность по посредничеству» // Проблемы информатики. 2014. №3. С. 53–64.

14. *Печников А.А., Чебуков Д.Е.* Структура графа цитирования журналов Math-Net.Ru // Научный сервис в сети Интернет: тр. XXIII Всерос. науч. конф. (20–23 сент. 2021 г.). М.: ИПМ им. М.В. Келдыша, 2021. С. 265–278. https://doi.org/10.20948/abrau-2021-2.

15. *Ольгина И.Г.* Метод определения важных узлов сети цитирования научных публикаций // Вестник компьютерных и информационных технологий. 2021. Т. 18, № 5 (203). С. 3–10. https://doi.org/10.14489/vkit.2021.05.pp.003-010.

16. *Freeman L.C.* Centrality in social networks conceptual clarification // Social Networks. 1978. No. 31. P. 215–239.

17. *Newman M.E.J.* Scientific collaboration networks. I. Network construction and fundamental results // Physical Review. 2001. Vol. 64, No. 1. P. 016131. https://doi.org/10.1103/PhysRevE.64.016131.

18. *Brandes U.* A faster algorithm for betweenness centrality // The Journal of Mathematical Sociology. 2001. Vol. 25, No. 2. P. 163–177.

19. *Kleinberg J.* Authoritative sources in a hyperlinked environment // Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA 98). 1998. P. 668–677.

20. *Leon C., Perez J.* Authority Centrality and Hub Centrality as metrics of systemic importance of financial market infrastructures // Borradores de Economia. 2013. Vol. 754. P. 1–25. https://doi.org/10.2139/ssrn.2290271.

21. *Бредихин С.В., Ляпунов В.М., Щербакова Н.Г., Юргенсон А.Н.* Параметры «центральности» узлов сети цитирования научных статей // Проблемы информатики. 2016. № 1. С. 39–57.

22. *Shibata N., Kajikawa Y., Takeda Y. et al.* Early detection of innovations from citation networks // International Conference on Industrial Engineering and Engineering Management (Hong Kong, 8–11 December 2009). IEEE, 2009. https://doi.org/10.1109/ieem.2009.5373444.

23. *Baglioni M., Geraci F., Pellegrini M., Lastres E.* Fast Exact Computation of betweenness Centrality in Social Networks // Advances in Social Networks Analysis and Mining: International Conference 2012 IEEE/ACM (Istanbul, 26–29 August 2012). P. 450–456. https://doi.org/10.1109/ASONAM.2012.79.

24. *Farhan M.T., Darwiyanto E., Asror I.* Analysis of Hubs and Authorities Centrality Using Probabilistic Affinity Index (PAI) on directed-weighted graph in Social Network Analysis // Journal of Physics: Conference Series. 2019. Vol. 1192. P. 012005. https://doi.org/10.1088/1742-6596/1192/1/012005.

25. *Marra A., Antonelli P., Dell'Anna L., Pozzi C.* A network analysis using metadata to investigate innovation in clean-tech – Implications for energy policy // Energy Policy. 2015. Vol. 86. P. 17–26.

26. *Baronchelli A., Ferrer-i-Cancho R, Pastor-Satorras R., Chater N., Christiansen Morten N.* Networks in cognitive science // Trends in cognitive sciences. 2013. Vol. 17. Iss. 7. P. 348–360. https://doi.org/10.1016/j.tics.2013.04.010.

27. *Karuza E.A., Thompson-Schill Sh.L., Bassett Danielle S.* Local patterns to global architectures: influences of network topology on human learning // Trends in cognitive sciences. 2016. Vol. 20. Iss. 8. P. 629–640. https://doi.org/10.1016/j.tics.2016.06.003.

28. *Merseal Hannah M., Beaty Roger E., Kenett Yoed N., Lloyd-Co J., Orjan de Manzano, Norgaard Martin*. Representing melodic relationships using network science // Cognition. 2023. Vol. 233. P. 105362.
https://doi.org/10.1016/j.cognition.2022.105362.

29. *Barabasi A.-L.* Scale-free networks: Aa decade and beyond // Science. 2009. Vol. 325. Iss. 5939. P. 412–413.

30. *Кошелева Н.Н.* Корреляционный анализ и его применение для подсчета ранговой корреляции Спирмена // Актуальные проблемы гуманитарных и естественных наук. 2012. № 5. С. 23–26.

31. *Ермолаев О.Ю.* Математическая статистика для психологов. М.: Московский психолого-социальный институт: Флинта, 2003. 366 с.

METHODOLOGY OF NETWORK ANALYSIS OF SCIENTIFIC PUBLICATIONS

I. G. Olgina[0000-0002-9932-4552]

Omsk State Technical University

inna_olgina@mail.ru

Abstract

The relevance of the issues of the analysis of scientific publications is due to the fact that with the advent of Internet technologies, it became possible to collect data on the publication citation network. Meanwhile, the current approach to the analysis of scientific publications is based on bibliometric indicators that take into account only the number of citations. However, network analysis, which is mainly used in the study of social networks, is becoming increasingly widely used. The author has developed a methodology that allows for an effective analysis of scientific publications based on network analysis methods alternative to bibliometric methods. As criteria for evaluating scientific publications based on network analysis, relevant measures of the centrality of the citation network nodes are established: centrality by degree of connectivity; centrality by proximity to other nodes; centrality by mediation; centrality by authority; centrality by concentration. The author presented the experiment re-

sult that allows validating the developed methodology of network analysis of the scientific publications significance. Scientometric databases were used as primary sources of data on publications, which make it possible to track the citation of publications and identify relevant citation networks. The application of the proposed network analysis methodology contributes to the identification of important publications in the development of the scientific direction.

Keywords*: citation network, publications, scientometrics, bibliometric analysis, network analysis, graph*

REFERENCES

1. Gonashvili A.S. Naukometricheskie bazy dannyh i rabota s nimi [Scientometric databases and working with them]. SPb.: MPA EvrAzJeS. 2020. 57 p. (In Russ.).

2. H-index. Available at: https://en.wikipedia.org/wiki/H-index (accessed 15.04.2023).

3. Impact factor. Available at: https://ru.wikipedia.org/wiki/ Impact factor (accessed 15.04.2023).

4. *Lu Z., Ma Y., Song L*. Patent Citation Network Analysis Based on Improved Main Path Analysis: Mapping Key Technology Trajectory // Advances in Artificial Intelligence and Security (ICAIS 2021): Communications in Computer and Information Science. Springer: Cham, 2021. Vol. 1423. P. 158–171. https://doi.org/10.1007/978-3-030-78618-2_13.

5. *Wang J., Cheng Q., Lu W. et al.* A term function–aware keyword citation network method for science mapping analysis // Information Processing & Management. Vol. 60, No. 4. P. 103405. https://doi.org/10.1016/j.ipm.2023.103405.

6. *Ol'gina I.G., Pronin I.V., Abdrahmanov A.N.* Postroenie grafovyh modelej seti citirovanija nauchnyh publikacij [Graph models construction of the citation network of scientific publications] // Sistemy upravlenija, informacionnye tehnologii i matematicheskoe modelirovanie: materialy Vtoroj Vseros. nauch.-prakt. konf. s mezhdunar. uchast [Control systems, information technologies and mathematical modeling: Collected papers]. Omsk: OmGTU, 2020. Vol. 1. P. 118–125 (In Russ.).

7. *Zhao F., Zhang Y., Lu J. et al.* Measuring academic influence using heterogeneous author-citation networks // Scientometrics. 2019. Vol. 118. P. 1119–1140. https://doi.org/10.1016/10.1007/s11192-019-03010-5.

8. *Ji P., Jin J., Ke Z. T., L. W.* Co-citation and Co-authorship Networks of Statisticians // Journal of Business & Economic Statistics. 2022. Vol. 40, No. 2. P. 469–485. https://doi.org/10.1080/07350015.2021.1978469.

9. *Luc P.T., Lan P.X., Le A.N.H., Tran B.T.* A Co-Citation and Co-Word Analysis of Social Entrepreneurship Research // Journal of Social Entrepreneurship. 2022. Vol. 13, No. 3. P. 324–339. https://doi.org/10.1080/19420676.2020.1782971.

10. *Pechnikov A.A., Chebukov D.E.* Analiz soavtorstva v matematicheskih zhurnalah Math-Net.Ru [Analysis of co-authorship in mathematical journals of Math-Net.Ru]. Nauchnyj servis v seti Internet: trudy dvadcat' tret'ej Vserossijskoj nauchnoj konferencii. [Scientific service on the Internet: Conference proceedings]. M., 2022. P. 190-202. https://doi.org/10.20948/abrau-2022-5.

11. *Gómez S*. Centrality in Networks: Finding the Most Important Nodes // Business and Consumer Analytics: New Ideas / P. Moscato, N. Jane de Vries. Springer: Cham, 2019. P. 401–433. https://doi.org/10.1007/978-3-030-06222-4_8.

12. *Das K., Samanta S., Pa M.* Study on centrality measures in social networks: a survey // Social Network Analysis and Mining. 2018. Vol. 8. P. 13. https://doi.org/10.1007/s13278-018-0493-2.

13. *Bredihin S.V., Ljapunov V.M., Shherbakova N.G.* Mera vazhnosti nauchnoj periodiki – central'nost' po posrednichestvu [Measurement of the scientific periodicals importance: cooperation centrality]. Problemy informatiki = Problems of Informatics. 2014. No. 3. P. 53–64.

14. *Pechnikov A.A., Chebukov D.E.* Struktura grafa citirovanija zhurnalov Math-Net.Ru [Structure of the citation graph of Math-Net.Ru journals]. Nauchnyj servis v seti Internet: trudy dvadcat' tret'ej Vserossijskoj nauchnoj konferencii. [Scientific service on the Internet: Conference proceedings]. M., 2021. P. 265–278. https://doi.org/10.20948/abrau-2021-2 (In Russ.).

15. *Ol'gina I.G.* Metod opredelenija vazhnyh uzlov seti citirovanija nauchnyh publikacij [Method for determining important nodes of the citation network of scientific publications]. Vestnik komp'juternyh i informacionnyh tehnologij = Herald of Computer and Information Technologies. 2021. Vol. 18, No. 5 (203). P. 3–10. https://doi.org/10.14489/vkit.2021.05.pp.003-010 (In Russ.).

16. *Freeman L.C.* Centrality in social networks conceptual clarification // Social Networks. 1978. No. 31. P. 215–239.

17. *Newman M.E.J.* Scientific collaboration networks. I. Network construction and fundamental results // Physical Review. 2001. Vol. 64, No. 1. P. 016131. https://doi.org/10.1103/PhysRevE.64.016131.

18. *Brandes U.* A faster algorithm for betweenness centrality // The Journal of Mathematical Sociology. 2001. Vol. 25, No. 2. P. 163–177.

19. *Kleinberg J.* Authoritative sources in a hyperlinked environment // Proceedings of the 9th ACM-SIAM Symposium on Discrete Algorithms (SODA 98). 1998. P. 668–677.

20. *Leon C., Perez J.* Authority Centrality and Hub Centrality as metrics of systemic importance of financial market infrastructures // Borradores de Economia. 2013. Vol. 754. P. 1–25. https://doi.org/10.2139/ssrn.2290271.

21. Бредихин С.В., Ляпунов В.М., Щербакова Н.Г., Юргенсон А.Н. Параметры «центральности» узлов сети цитирования научных статей // Проблемы информатики. 2016. № 1. С. 39–57.

22. *Shibata N., Kajikawa Y., Takeda Y. et al.* Early detection of innovations from citation networks // International Conference on Industrial Engineering and Engineering Management (Hong Kong, 8–11 December 2009). IEEE, 2009. https://doi.org/10.1109/ieem.2009.5373444.

23. *Baglioni M., Geraci F., Pellegrini M., Lastres E.* Fast Exact Computation of betweenness Centrality in Social Networks // Advances in Social Networks Analysis and Mining: International Conference 2012 IEEE/ACM (Istanbul, 26–29 August 2012). P. 450–456. https://doi.org/10.1109/ASONAM.2012.79.

24. *Farhan M.T., Darwiyanto E., Asror I.* Analysis of Hubs and Authorities Centrality Using Probabilistic Affinity Index (PAI) on directed-weighted graph in Social Network Analysis // Journal of Physics: Conference Series. 2019. Vol. 1192. P. 012005. https://doi.org/10.1088/1742-6596/1192/1/012005.

25. *Marra A., Antonelli P., Dell'Anna L., Pozzi C.* A network analysis using metadata to investigate innovation in clean-tech – Implications for energy policy // Energy Policy. 2015. Vol. 86. P. 17–26.

26. *Kosheleva N.N.* Korreljacionnyj analiz i ego primenenie dlja podscheta rangovoj korreljacii Spirmena [Correlation analysis and its application for calculate Spearman's rank correlation] // Aktual'nye problemy gumanitarnyh i estestvennyh nauk = Current problems in the humanities and natural sciences. 2012. No. 5. P. 23–26. (In Russ.).

27. *Ermolaev O.Ju.* Matematicheskaja statistika dlja psihologov [Mathematical statistics for psychologists]. M.: Moscow Psychological and Social Institute: Flinta, 2003. 366 p. (In Russ.).

28. *Baronchelli A., Ferrer-i-Cancho R, Pastor-Satorras R., Chater N., Christiansen Morten N.* Networks in cognitive science // Trends in cognitive sciences. 2013. Vol. 17. Iss. 7. P. 348–360. https://doi.org/10.1016/j.tics.2013.04.010.

29. *Karuza E.A., Thompson-Schill Sh.L., Bassett Danielle S.* Local patterns to global architectures: influences of network topology on human learning // Trends in cognitive sciences. 2016. Vol. 20. Iss. 8. P. 629–640. https://doi.org/10.1016/j.tics.2016.06.003.

30. *Merseal Hannah M., Beaty Roger E., Kenett Yoed N., Lloyd-Co J., Orjan de Manzano, Norgaard Martin*. Representing melodic relationships using network science // Cognition. 2023. Vol. 233. P. 105362.

https://doi.org/10.1016/j.cognition.2022.105362.

31. *Barabasi A.-L.* Scale-free networks: Aa decade and beyond // Science. 2009. Vol. 325. Iss. 5939. P. 412–413.

СВЕДЕНИЯ ОБ АВТОРЕ

ОЛЬГИНА Инна Геннадьевна – директор библиотеки ОмГТУ, старший преподаватель кафедры «Математические методы и информационные технологии в экономике», ОмГТУ. Область научных интересов: теория графов, применение методов системного анализа для исследования сетей цитирования.

Inna Gennadevna OLGINA – Director of the library in Omsk State Technical University, Senior Lecturer of the Mathematical Methods and Information Technologies in Economics Department, Omsk State Technical University. Research interests: graph theory, application of systems analysis methods for the citation networks research.

email: inna_olgina@mail.ru

ORCID: 0000-0002-9932-4552

Материал поступил в редакцию 20 августа 2023 года

УДК 004.85

МЕТОДЫ ИСКУССТВЕННОГО ИНТЕЛЛЕКТА ДЛЯ НАУЧНЫХ ИССЛЕДОВАНИЙ В ГЕОЛОГИИ

М. И. Патук[1] [0000-0003-3036-2275], В. В. Наумова[2] [0000-0002-3001-1638]

ФГБУН Государственный геологический музей им. В.И. Вернадского Российской академии наук, Москва, Россия

[1]patuk@mail.ru, [2]naumova_new@mail.ru

Аннотация

Приведен краткий обзор некоторых методов искусственного интеллекта в области наук о Земле. Отмечены перспективы применения указанных методов для получения новых знаний. Приведены результаты первых попыток авторов в применении методов обработки естественного языка для обработки научных статей по геологии. Обсуждены возможности развития работ в этом направлении.

***Ключевые слова**:* Искусственный интеллект, машинное обучение, обработка естественного языка, геология.

ВВЕДЕНИЕ

Известно много определений термина искусственный интеллект (ИИ). Например:

- ИИ — общий термин, описывающий системы, выполняющие когнитивные, познавательные функции, например, решение тематических проблем [1].
- ИИ – способность системы правильно интерпретировать внешние данные, извлекать уроки из таких данных и использовать полученные знания для достижения конкретных целей и задач посредством гибкой адаптации [2].

Уровень познания, необходимый для выполнения определенной задачи, определяется ее характером, поэтому рассматриваемый термин можно применять в отношении любого процесса поиска решения или интерпретации данных с использованием компьютера. Таким образом, понятие «искусственный интеллект» охватывает широкий спектр процессов, используется в контексте программного обеспечения и соответствующих услуг, в том числе связанных с машинным

обучением. В области ИИ существует много подходов и направлений. Мы остановимся на трех их них: экспертные системы, обработка изображений и обработка естественного языка (Рис. 1).

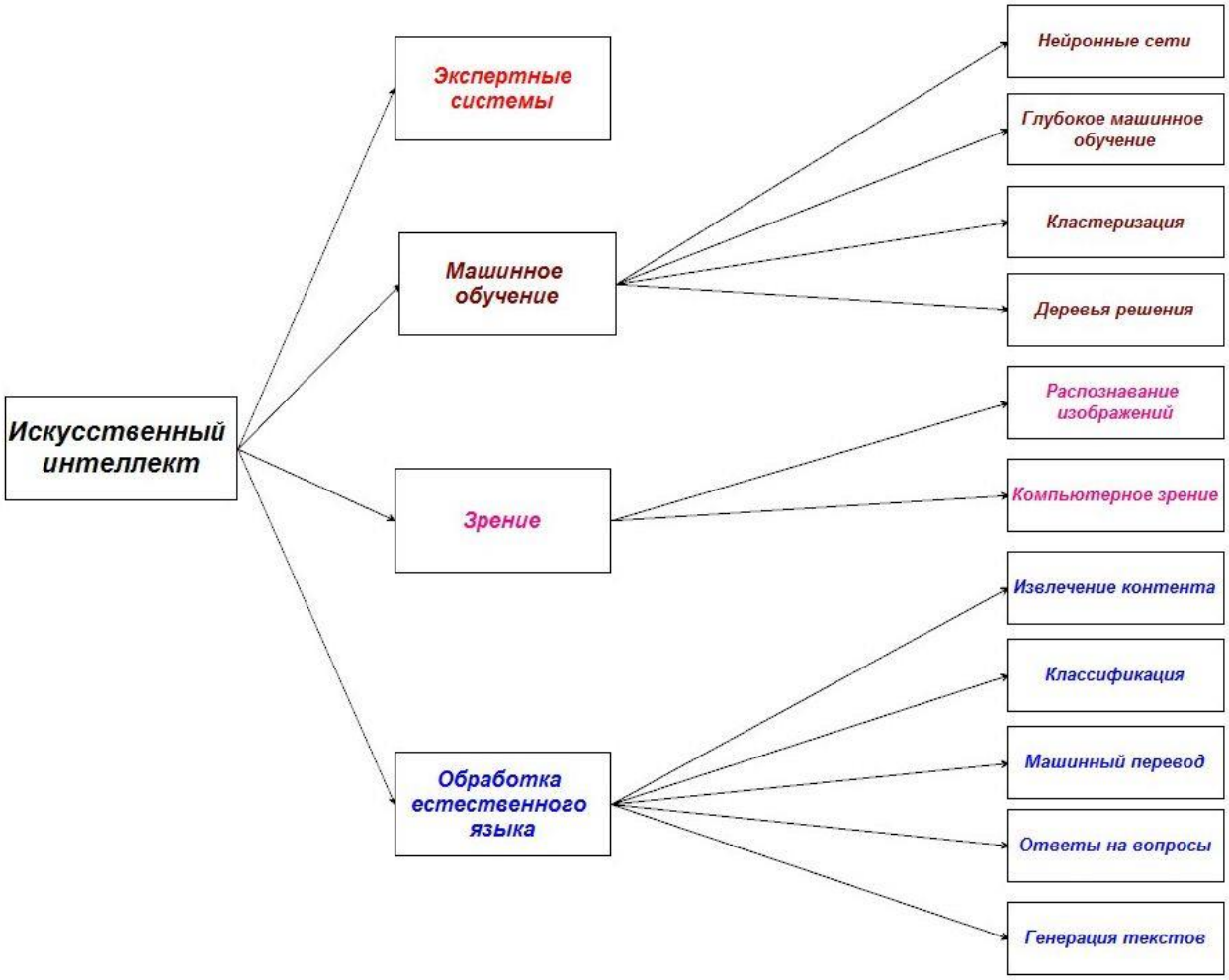

Рис. 1. Обозреваемые направления искусственного интеллекта

Экспертные системы в геологии

Экспертная система (ЭС) – это интеллектуальная программа, способная делать логические выводы на основании знаний в конкретной предметной области и обеспечивающая решение специфических задач. Для этого ее необходимо наделить функциями, позволяющими решать задачи, которые в отсутствие эксперта (специалиста в данной конкретной предметной области) невозможно решить правильно.

В области геологии одной из первых является экспертная система

PROSPECTOR [3], разработанная для оказания помощи геологам-поисковикам. Эта программа создана компаниями ESRI [4] (совместно с консультантами по геологии) и USGS [5]. Она способна давать три типа «советов»: оценку местности на предмет существования определенных залежей; оценку прогнозных ресурсов региона и выбор перспективных участков для бурения. Эксперт-геолог применяет очень ограниченные модели при выявлении территорий с возможными запасами, например, золота. В ЭС PROSPECTOR эти модели в закодированном виде находятся в памяти компьютера, они определенным образом интерпретируются при оценке того или иного участка. Одна из главных особенностей знаний экспертов-геологов заключается в том, что они неполны и неопределенны. В связи с этим используются специальные технические приемы, отличные от тех, что применяются в других (более определенных) экспертных системах.

Экспертная система «ОЛОВО» [6]

Для работы с ЭС «Олово» не требуется специальной подготовки картографического или иного материала. Единственное условие – квалифицированное владение пользователем совокупностью сведений, относящихся к комплексной многоуровневой характеристике объекта прогноза.

Экспертная система «Олово» воспринимает новые данные для решений одной или комплекса прогнозных задач по какому-либо конкретному объекту или участку территории в виде геологических знаний пользователя о данном объекте. Система непрерывно запрашивает пользователя, пока не будут заполнены все отсутствующие компоненты новой структуры (модели экзаменуемого объекта).

В процессе усвоения знаний об экзаменуемом объекте на каждом шаге развития диалога с пользователем происходят поэтапное формирование модели этого объекта, а также поэтапное последовательное сопоставление всех ее элементов с соответствующими элементами знаний, заложенных в базу знаний (БЗ) экспертных моделей.

Непосредственное решение прогнозных задач выполняется в системе с использованием нескольких, различных по своему смысловому содержанию методических приемов, каждый из которых обеспечен соответствующим математическим аппаратом. Один из таких приемов основан на методе аналогий. При его реализации за каждым шагом усвоения новых знаний об экзаменуемом объекте

следует определение вероятной схожести (в численном выражении) конструируемой модели объекта с той или иной экспертной моделью. В процессе диалога с пользователем по мере накопления новых знаний эти вероятности могут изменяться. Все текущие изменения учитываются при формировании прогнозного заключения, в котором указывается степень соответствия данного объекта определенной экспертной модели.

Другой методический прием содержит в своей основе метод распознавания образов. Эталонные образы в виде экспертных моделей заложены в БЗ. Поэтапный анализ положения конкретных признаков экзаменуемого объекта в многомерном пространстве экспертных признаков позволяет системе производить классификацию знаний, получаемых в процессе диалога с пользователем. Наличие в системе специфических решающих правил обеспечивает возможность использования в процессе формирования прогнозного заключения наиболее значимых признаков с оценкой степени их влияния на окончательные выводы и демонстрировать пользователю численное значение вероятностей распознавания образа. Таким путем достигается дифференциация окончательных результатов решения задачи по степени их надежности.

Таким образом, ЭС «Олово» позволяет решать задачи прогнозирования на стадии регионального изучения недр. Более локальные прогнозы, касающиеся количественной оценки ресурсов оловорудного узла, уровня эрозионного среза объекта и возможности обнаружения промышленных скоплений руд решается с меньшей долей вероятности. Это связано со спецификой созданной базы знаний, которая включает широкий спектр разнообразной геологической информации, но лишена знаний экономического характера. Тем не менее, по ряду косвенных характеристик и различных критериев (минералого-геохимических, геофизических, геолого-минералогических и др.) эта система дает количественные прогнозы и определяет вероятность их подтверждения.

Сейчас количество экспертных систем исчисляется тысячами и десятками тысяч. В развитых зарубежных странах сотни фирм занимаются их разработкой и внедрением.

SOLSA [7] – первая автоматизированная экспертная система для анализа

керна на месте. Благодаря доступу к данным в режиме онлайн ожидается значительная экономия на количестве буровых скважин, точности геомоделей и экономической оценке запасов руды. ЭС SOLSA идеально отвечает потребности в «Новых технологиях устойчивой разведки и геомоделей» SC5-11d-2015. Целью ее создания была «разработка новых или усовершенствованных высокоэффективных и рентабельных, устойчивых технологий разведки», включая:

- комплексное бурение, оптимизированное для работы в сложных латеритных условиях со сложной смесью твердых и мягких пород, распространяемое также на другие типы руд,
- полностью автоматизированный сканер и анализатор фазовой идентификации, а также программное обеспечение, которое можно использовать и в других отраслях.

ЭС SOLSA впервые объединила неразрушающие датчики рентгеновской флуоресценции, рентгеновской дифракции, колебательной спектроскопии, 3D- и гиперспектральной визуализации вдоль керна скважины. Для этой цели SOLSA разработала инновационное, удобное для пользователя и интеллектуальное программное обеспечение на уровнях TRL 4-6. Чтобы минимизировать риск и извлечь выгоду из новейших технологий, на рынке миниатюрных датчиков были выбраны подсистемы для аппаратного обеспечения.

ЭС SOLSA призвана произвести революцию в геологической разведке, сократив ее время на 50%, а время анализа – с 3–6 месяцев до реального времени и, таким образом, снизить воздействие на окружающую среду.

Машинное обучение в геологии

Машинное обучение — это класс количественных методов (под которыми зачастую понимают алгоритмы), предназначенных для ускорения процесса прогнозирования определенных показателей на основе некоторого прецедента [1]. В отличие от остальных направлений в ИИ, машинное обучение не требует ручного ввода в алгоритм правил принятия решений — они автоматически определяются системой по эмпирическим данным.

Существует широкий спектр алгоритмов машинного обучения, подходящих для выполнения специализированного геологического анализа. Исходный мате-

риал для их обучения обычно либо уже имеется, либо может быть получен самостоятельно. Таким образом, машинное обучение можно использовать с целью выявления геологоразведочных объектов в условиях избытка данных (например, решения Goldspot Discoveries [8], SRK Consulting [9]), автоматического выявления геологических зон залегания полезных ископаемых (Maptek [10]), оценки твердости руды на основе результатов анализа (неопубликованные работы), распознавания частиц золота по фотоснимкам пробы ледниковых отложений (IOS Services Geoscientifiques [11]).

В последние десятилетия наблюдается стремительный рост интереса к нейронным сетям, которые успешно применяются в различных областях — бизнесе, медицине, технике, геологии, физике. Нейронные сети используются всюду, где нужно решать задачи прогнозирования, классификации, нелинейной регрессии или управления. Такой впечатляющий успех нейронных сетей определяется богатыми возможностями и простотой в использовании. Особенность работы нейросетей состоит в том, что такая сеть обучается на исторических данных, находит специфические паттерны, указывающие на зависимости внутри данных, и на их основе строит свой прогноз.

Одной из разновидностей нейронных сетей, предназначенной для обработки изображений и других точечных форматов, являются сверточные нейронные сети. В геологоразведке они применяются для выявления объектов (например, решения Orefox [12]), обработки и интерпретации сейсмических данных (Geolearn [13]), определения минералов-индикаторов в пробах ледниковых отложений (IOS Services Geoscientifiques [11]), а также количественного и качественного описания буровых кернов по их фотоснимкам (Geolearn [13]) или гиперспектральным данным (Solve Geosolutions [14]).

Последовательность входных данных анализируется с помощью такой разновидности нейронных сетей, как рекуррентные нейронные сети. Они адаптированы для анализа временных наборов данных, таких как временные последовательности или текстовая информация. В геологоразведке рекуррентные нейронные сети используют для выявления перспективных участков на основе отчетов, находящихся в свободном доступе (например, решения Goldspot Discoveries [8]),

или для геологического документирования данных бурения на основании измерений физических свойств пород (CGG).

Методы машинного обучения все чаще используются в горнодобывающей промышленности. Они эффективны в решении повторяющихся задач или задач с большим количеством многомерных данных (качественных и правильно обработанных) [1].

Объективность, продуктивность и адаптивность алгоритмов машинного обучения делают их идеальным решением широкого спектра проблем различного масштаба. Однако подготовка и внедрение таких технологий в разведке и добыче требуют немалого опыта. Моделирование — это комплексная работа, которой сопутствуют характерные сложности, и качество входных данных — не самая последняя из них.

Анализ изображений в геологии

В геологии большую часть времени работы исследователей занимают визуальная диагностика и описание горных пород. Состав породы, ее структура и текстура должны быть определены и описаны. Это занимает много времени, и естественно возникает желание автоматизировать эту работу. В работе [15] предложен новый подход к автоматической классификации пород при описании кернов. Описаны доступные методы автоматической классификации пород, предложен новый подход применения сверточной нейронной сети. Поскольку для тренировки нейронной сети требуется большое количество данных, были предприняты специальные усилия для генерации дополнительных изображений. Приведено описание и проведено сравнение разных архитектур нейронной сети. Была достигнута точность 72%. Система автоматической классификации натренирована на 20 000 изображений образцов из 10 нефтяных и газовых полей различных геологических условий. Описаны ограничения применения полученной модели.

Получить достаточное количество изображений для анализа не всегда возможно. Поэтому в работе [16] авторы воспользовались специальными методами обучения с нулевым результатом (zero-shot learning) и обучение с малым числом примеров (few-shot learning). В результате стало возможным предложить открытый критерий для распознавания необработанных минералов, которые отсут-

ствуют в обучающей выборке. Также предоставлены дополнительные наборы образцов для сегментации, определения размеров образцов и классификации с малым числом примеров. Для всех описанных задач компьютерного зрения предоставлены базовые алгоритмы. В статье показана важность унифицированных данных для корректной работы распознавания минералов. Созданные коллекции изображений минералов выложены в открытый доступ для использования всеми желающими.

Обработка естественного языка в геологии

Обработка естественного языка — это общее направление искусственного интеллекта и математической лингвистики. Оно изучает проблемы компьютерного анализа и синтеза текстов на естественном языке [17].

Рассмотрим некоторые определения из данной области: языковой моделью называется набор свойств и методов по построению распределения вероятностей последовательности слов [18]. Языковая модель создает вероятности в процессе обучения на корпусе текстов. Корпусом текстов называется набор текстов, подобранный и обработанный по определенным правилам. В лингвистике, откуда пришел этот термин, под корпусом текстов понимается коллекция текстов, размеченная с помощью специальных тэгов. В области машинного обучения под корпусом текстов часто понимается просто их коллекция, без специальной разметки. Векторное представление слова (word embeddings) – назначение слову числового вектора на основе его анализа языковой моделью.

В области наук о Земле с помощью методов обработки естественного языка (NLP) решаются следующие задачи: выделение геологических и географических именованных сущностей (NER), извлечение пространственных и временных взаимосвязей, классификация, кластеризация, реферирование геологических отчетов и публикаций, ответы на вопросы.

Большие языковые модели, такие как BERT, ChatGPT и GPT-4, достигли большого успеха в применении к общеупотребительным языковым областям, таким как новостные, социокультурные, медийные. В области наук о Земле их применение не столь впечатляющее, в первую очередь, из-за отсутствия в их арсенале спе-

цифической геологической терминологии, поскольку она отсутствует в тех языковых корпусах текста, на которых обучались данные модели [19]. Авторами была создана большая языковая модель для наук о Земле – K2 [20] – на основе обучения GPT подобной языковой модели LLaMA. Для целей тренировки модели создан большой набор текстов, содержащий более 6 млн. записей, включающих статьи, метаданные статей и данные из Wikipedia. K2 – генеративная модель. Она может создавать тексты по теме наук о Земле и отвечать на соответствующие вопросы.

Предсказательные возможности методов обработки естественного языка основаны на статистическом языковом моделировании, имеющем в своей основе векторное представление слов (word embeddings). Векторное представление каждого слова создается на основе частот взаимного расположения других слов, находящихся вблизи выбранного слова в обучающем корпусе текстов. Таким образом создается контекстно-зависимое представление слов, что очень полезно для слов с множественными значениями (полисемия). Например, такие геологические термины, как щит, плита, кора, пояс, узел, чехол, трубка, осадки, мел (сокращенное от «меловой период»), мантия, свита, имеют значения, совсем отличные от общеупотребительных. Понять их геологический смысл можно только по контексту. Появление таких контекстно-зависимых моделей, как BERT, позволило значительно улучшить возможности этого класса моделей в обработке естественного языка.

Указанная контекстная зависимость современного поколения языковых моделей имеет свою обратную сторону. Чтобы получать адекватные результаты, необходимо тренировать такие модели на текстах, специфичных для каждой предметной области. В [21] дано сравнение результатов выполнения разными языковыми моделями тестов со специфической геологической терминологией. Модели, обученные на корпусах геологических текстов, превзошли модели, обученные на значительно больших корпусах общих текстов.

Указанная работа [21] делает упор на обучении двух моделей: GloVe – не контекстно зависимой и BERT – контекстно зависимой, на неразмеченных коллекциях текстов (геологические отчеты и научные публикации, доступные в свободном доступе). Созданы внутренние критерии оценки моделей (аналогии, образо-

вание кластеров, родство и ближайшее окружение) взамен внешних тестов. Показана возможность извлечения, с помощью указанных моделей, данных о геохимических и минеральных ассоциациях из необработанных текстовых данных геологической направленности.

В работе [22] выполнены извлечение ключевых слов и их визуализация (облако тэгов, индексы центральности) из геологических отчетов. Извлечение ключевых слов выполнялось на основе модифицированного алгоритма TF-IDF.

В работе [23] создана языковая модель для области наук о Земле (GeoVec), обученная на 280 000 научных статей из данной области. Для внутренней оценки модели были проведены тесты по созданию аналогий, на определение родственных терминов и деление терминов на категории. Было показано превосходство данной модели для области наук о Земле над стандартными моделями, которые обучались на общемедийных текстах.

Обработка естественного языка была использована для извлечения текстовых данных из описательной части геологических карт [24]. Выбирались данные описания пород, геологического возраста, литостратиграфические описания. Извлеченные данные преобразовывались в векторную форму, и с использованием статистических методов находились семантические связи между типами пород. Кроме того, с помощью тех же методов выполнялось предсказание территорий, перспективных на Zn-Pb оруденение.

В работе [25] обработка естественного языка была использована для классификации и 3-х мерного литологического картирования. Были использованы текстовые данные материалов бурения. Векторные представления слов были получены из заранее обученной модели GloVe [23]. Текстовые описания были размечены экспертами. Это позволило выполнить классификацию текстовых данных с помощью нейронной сети и посредством интерполяции создать приемлемые 3-х мерные литологические карты исследованного района Австралии.

Большой объем неструктурированной геологической информации содержится в геологических отчетах и научных статьях. Невозможно простым прочтением охватить эту лавину информации. Отсюда вытекает задача извлечения краткого резюме из имеющейся информации для быстрого первичного анализа имеющихся источников [26]. Геологические тексты обладают большой спецификой

из-за использования большого количества специфических терминов и стоящих за ними взаимосвязей и геологических концепций. Авторы предложили последовательный подход по извлечению таких терминов в виде геологических именованных сущностей на неразмеченных текстах геологических отчетов.

Задачи бинарной классификации

Из приведенного выше краткого обзора работ по применению методов обработки естественного языка в области наук о Земле видно, что для работы с этими методами необходимо иметь обученную языковую модель. И не просто обученную, а обученную на текстах из интересующей нас предметной области. Русскоязычные языковые модели существуют, но нам не удалось найти ни одной, обученной на текстах геологической направленности. Таким образом, для начала работы в этой области необходимо получить коллекцию геологических текстов.

Эта проблема оказалась решаемой, т. к. в Государственном геологическом музее им. В.И. Вернадского имеются два текстовых ресурса: архив научных публикаций с тематикой «Науки о Земле» (https://repository.geologyscience.ru/) [27] и wiki-Геология России (http://wiki.geologyscience.ru) [28]. Текстовые данные этих ресурсов находятся в SQL-базах, что позволяет достаточно легко их извлекать.

Первая задача, которая нам показалась интересной не столько с практической, но и с методологической точки зрения, – это бинарная классификация. Эта задача подробно и многократно описана в интернете. Мы выбрали следующее описание [29], как очень подробное и предоставляющее возможность скачать все обсуждаемые примеры с GitHub [30]. В упомянутой статье представлено 8 моделей для классификации. Последние две модели из списка мы не использовали, поскольку там используется заранее натренированная англоязычная модель. Естественно, мы не стали использовать предоставляемые данные, поскольку они не нашей тематики и к тому же англоязычные. Нами были выбраны 50 научных статей из архива публикаций (https://repository.geologyscience.ru/), касающихся месторождений золота и железа. PDF-файлы этих статей были конвертированы в текст с помощью пакета PDFReader [31]. Дополнительно тексты были очищены от всех некириллических символов и цифр с помощью регулярных выражений и удалены стоп-слова (слова типа предлогов, местоимений, которые не несут смысловой нагрузки).

Для дальнейшей обработки текст необходимо его токенизировать, т. е. разбить на отдельные слова и привести их в нормальную форму (единственное число, именительный падеж, мужской род). Для этого мы использовали две русскоязычные модели: 1-я модель – ru_core_news_lg [32]– автор Александр Кукушкин, 2-я модель – spacy-stanza [33], предыдущее название – StanfordNLP. Анализ текстов после токенизации показал, что модель spacy-stanza работает в нашем случае лучше, но с ошибками, коверкает некоторые слова, например, вместо «кристаллический» вставляет «каллический».

В соответствии с рекомендациями по подготовке исходных данных [34] все тексты после токенизации были вычитаны и отредактированы. Расчет по 6-ти указанным выше моделям оказался неудовлетворительным – низкая точность, высокие потери (Рис. 2).

```
Epoch 1/5

1/1 [==============================] - ETA: 0s - loss: 0.6754 - accuracy: 0.6800
1/1 [==============================] - 2s 2s/step - loss: 0.6754 - accuracy: 0.6800 - val_loss: 0.6609 - val_accuracy: 0.6667
Epoch 2/5

1/1 [==============================] - ETA: 0s - loss: 0.6322 - accuracy: 0.7600
1/1 [==============================] - 0s 172ms/step - loss: 0.6322 - accuracy: 0.7600 - val_loss: 0.6537 - val_accuracy: 0.6667
Epoch 3/5

1/1 [==============================] - ETA: 0s - loss: 0.5961 - accuracy: 1.0000
1/1 [==============================] - 0s 156ms/step - loss: 0.5961 - accuracy: 1.0000 - val_loss: 0.6470 - val_accuracy: 0.6667
Epoch 4/5

1/1 [==============================] - ETA: 0s - loss: 0.5652 - accuracy: 1.0000
1/1 [==============================] - 0s 156ms/step - loss: 0.5652 - accuracy: 1.0000 - val_loss: 0.6401 - val_accuracy: 0.6667
Epoch 5/5

1/1 [==============================] - ETA: 0s - loss: 0.5366 - accuracy: 1.0000
1/1 [==============================] - 0s 172ms/step - loss: 0.5366 - accuracy: 1.0000 - val_loss: 0.6330 - val_accuracy: 0.6667
```

Рис. 2. Результат обучения сверточной нейронной сети на малом числе примеров

После этого мы изменили свой подход к получению исходных данных: вместо полнотекстовых статей из архива публикаций (https://repository.geologyscience.ru) были выбраны абстракты статей, касающихся описания месторождений золота и железа. Всего было выбрано 1750 записей (1150 – про золото и 600 – про железо). Средняя длина строк текста – 120 слов. На этих данных мы получили более вдохновляющие результаты – точность достигла 92% (Рис. 3). Наилучшие результаты показала модель одномерной сверточной нейронной сети (1D Convolutional Neural Network). При этом мы не очищали эти тексты и не удаляли стоп-слова. Дополнительная чистка текстов и удаление стоп слов незначительно повысили точность (менее чем на 1%). Завершающий

этап – проверка полученной модели. На этом этапе на вход модели подается текст, который она раньше «не видела». Ее задача – отнести этот текст к одной из двух категорий, на которых она обучалась. Модель в основном успешно классифицировала тексты, как описывающие месторождения золота или железа. Этот результат с очевидностью показал, что при обучении языковых моделей количество (1750 против 50) имеет первостепенное значение.

```
28/50 [===============>..............] - ETA: 7s - loss: 0.0262 - accuracy: 1.0000
29/50 [================>.............] - ETA: 6s - loss: 0.0259 - accuracy: 1.0000
30/50 [================>.............] - ETA: 6s - loss: 0.0260 - accuracy: 1.0000
31/50 [=================>............] - ETA: 6s - loss: 0.0257 - accuracy: 1.0000
32/50 [=================>............] - ETA: 5s - loss: 0.0257 - accuracy: 1.0000
33/50 [=================>............] - ETA: 5s - loss: 0.0256 - accuracy: 1.0000
34/50 [==================>...........] - ETA: 5s - loss: 0.0256 - accuracy: 1.0000
35/50 [==================>...........] - ETA: 4s - loss: 0.0253 - accuracy: 1.0000
36/50 [===================>..........] - ETA: 4s - loss: 0.0253 - accuracy: 1.0000
37/50 [===================>..........] - ETA: 4s - loss: 0.0253 - accuracy: 1.0000
38/50 [===================>..........] - ETA: 3s - loss: 0.0253 - accuracy: 1.0000
39/50 [====================>.........] - ETA: 3s - loss: 0.0251 - accuracy: 1.0000
40/50 [====================>.........] - ETA: 3s - loss: 0.0250 - accuracy: 1.0000
41/50 [=====================>........] - ETA: 2s - loss: 0.0250 - accuracy: 1.0000
42/50 [=====================>........] - ETA: 2s - loss: 0.0250 - accuracy: 1.0000
43/50 [=====================>........] - ETA: 2s - loss: 0.0249 - accuracy: 1.0000
44/50 [======================>.......] - ETA: 1s - loss: 0.0248 - accuracy: 1.0000
45/50 [======================>.......] - ETA: 1s - loss: 0.0247 - accuracy: 1.0000
46/50 [======================>.......] - ETA: 1s - loss: 0.0245 - accuracy: 1.0000
47/50 [=======================>......] - ETA: 0s - loss: 0.0244 - accuracy: 1.0000
48/50 [=======================>......] - ETA: 0s - loss: 0.0241 - accuracy: 1.0000
49/50 [========================>.....] - ETA: 0s - loss: 0.0240 - accuracy: 1.0000
50/50 [========================>.....] - ETA: 0s - loss: 0.0240 - accuracy: 1.0000
50/50 [==============================] - 17s 335ms/step - loss: 0.0240 - accuracy: 1.0000 - val_loss: 0.1498 - val_accuracy: 0.9261
```

Рис. 3. Результат обучения сверточной нейронной сети на большом числе примеров

Далее мы несколько изменили исходные данные: вместо абстрактов с описанием месторождений железа были выбраны абстракты с общим описанием геологических массивов. Таких набралось 740. В этом случае мы выполнили бинарную классификацию – месторождения золота – геологические описания, не содержащие месторождений. Нам также успешно удалось выполнить тренировку этой модели, с такими же показателями точности. Но на этапе тестирования модели нас ждало разочарование. При выполнении классификации текстов с описанием объектов, на которых она обучалась (месторождения золота и описания геологических массивов), модель, как и в предыдущем случае, уверенно выполняла классификацию текстов. Но при попытке классификации текстов с описанием месторождений железа, которых не было в этой обучающей выборке, модель все их распознавала как описание месторождений золота. Это так называемая проблема обучения с нулевым результатом (zero-shot) [35], т. е. мы предложили язы-

ковой модели классифицировать объект, который она не видела на этапе обучения. Поэтому требуются дополнительное дообучение и корректировка модели или использование другого класса моделей (zero-shot learning), чтобы корректно обрабатывать такие ситуации. Это задача на будущее.

Задача выделения ключевых слов

Следующая задача в области обработки естественного языка, которую мы попытались решить, – это выделение ключевых слов из текстов геологической тематики. Извлечение ключевых слов (фраз) является высокоуровневым реферированием, позволяющим сжать большой документ до уровня емких коротких определений. В приведенном выше обзоре литературы, по крайней мере, две статьи решают эту задачу в области наук о Земле различными методами. Кроме того, мы можем отослать читателя к более подробным общим статьям по данной теме [36, 37].

Для решения этой задачи мы решили воспользоваться русскоязычной моделью T5 [38] и алгоритмом тренировки модели [39]. Эта модель относится к классу моделей генерирующего реферирования. Она распознает входной текст и создает новый текст на основе материала, на котором она была обучена. С ее помощью можно выполнять перевод текста, перефразирование, заполнение пропусков, восстановление, упрощение, ответы на вопросы по тексту, генерацию заголовков.

Как всегда, при обучении новой модели первым делом необходимо подготовить обучающую выборку. Мы воспользовались нашим архивом публикаций (https://repository.geologyscience.ru) и выбрали абстракты статей, которые сопровождаются ключевыми словами. Всего удалось получить 9320 записей.

Имеются две русскоязычные модели T5 – base и large. Как было отмечено [39], модель base не всегда корректно производит выделение ключевых слов, поэтому мы не стали экспериментировать с этой моделью, а сразу выбрали large модель. К сожалению, из-за размеров модели обучать ее на локальном компьютере оказалось проблематичным – не хватает ресурсов. Поэтому мы выбрали облачный сервис Яндекса – DataSphere [40]. Этот сервис обладает масштабируемой архитектурой и гибкой тарифной политикой. Мы использовали конфигурацию g2.4

(112 vCPU, 4 GPU A100). Обучение длилось около 20 минут.

Результаты тестирования модели

Абстракт:

«Изучены космоструктуры заангарский части Енисейского кряжа по материалам мультиспектральных космических систем Modis и Landsat ETM+. Выделены четыре системы кольцевых структур первого порядка, интерпретируемые как глубинные очаги гранитизации. Показаны закономерности размещения золотого оруденения в космогеологических структурах. Выделенные разноранговые космогеологические структуры находят отражение в аномальных структурах геохимических полей»

Реальные ключевые слова: золоторудные объекты; гранитизация; линейные структуры; кольцевые структуры; енисейский кряж;

Предложенные моделью: золото; енисейский кряж; енисейский кряж

Абстракт:

«Рассмотрены минералого-геохимические особенности каолинитовых прослоев (тонштейнов) Азейского месторождения Иркутского бассейна. Выявлена редкометалльная геохимическая специализация тонштейнов месторождения на РЗЭ, Y, Zr, Hf, U, Th, Ta, Sn, Ga, Cu, Pb, Se и Te. Проведен комплексный анализ возможных источников накопления первичного вещества тонштейнов. Приведено обоснование аэрогенного механизма накопления первичного материала каолинитовых прослоев. Обоснована вулканогенная модель их образования за счет пирокластического вещества кислого (липаритового) состава»

Реальные ключевые слова: иркутский угольный бассейн; азейское месторождение; минералогия; геохимия; уголь; тонштейны;

Предложенные моделью: иркутский бассейн; каолинитовые прослои;

Абстракт:

«Проведен литологический анализ верхнеюрско-нижнемеловой черносланцевой баженовской свиты двух районов северной части Хантейской гемиантеклизы и Межовского мегамыса. Первый из них расположен в пределах Пурпейско-Васюганского фациального района и отвечает глубоководной части палеобассейна, второй находится в Сильгинском фациальном районе и значи-

тельно более приближен к береговой линии. Разрезы баженовской свиты в районе Хантейской гемиантеклизы отличаются повышенной карбонатностью и кремнистостью, а также широким развитием аутигенного барита, что связывается с более интенсивным развитием фауны в этом районе. Баженовская свита на Межовском мегамысе характеризуется относительно низким содержанием карбонатных минералов и повышенным содержанием глинистого материала в породах»

Реальные ключевые слова: литология; верхняя юра; меловая;

Предложенные моделью: юрская; межовский мегамыс; баженовская свита;

Из приведенного примера видно, что обученная модель в целом неплохо справилась с выделением ключевых слов – нет явных промахов с неактуальными ключевыми словами. Но и недостатки налицо, в частности, это повторение ключевых слов и меньшее их количество. Главное в приведенном тесте, что все ключевые слова, предложенные моделью, хорошо соотносятся с текстом.

ЗАКЛЮЧЕНИЕ

Важный первый шаг в применении методов искусственного интеллекта для обработки текстов – это получение обучающей коллекции текстов. Во многих работах, связанных с обработкой текстов на русском языке, отмечаются проблемы с получением таких коллекций, особенно коллекций тематических. Поэтому еще раз отметим важность и своевременность создания ними архива публикаций с тематикой «Науки о Земле» (https://repository.geologyscience.ru/) [27], который позволил получить такие коллекции.

Бурное развитие в последние годы методов искусственного интеллекта, связанное с обработкой и генерацией текстов, открыло замечательные возможности по извлечению новых знаний из потока научной информации, которую очень трудно, а зачастую и невозможно переработать традиционным методом чтения. Приведенный краткий обзор литературы показывает, что применение этих методов в области наук о Земле находится в начальной фазе. Необходимо ознакомиться с данными методами, попробовать на простых задачах, выяснить области применения, их достоинства и недостатки. Наш первый опыт показал, что

методы обработки естественного языка действительно работают в области наук о Земле: можно получать непротиворечивые результаты. Необходимо дальнейшее изучение этих методов, чтобы можно было решать действительно серьезные задачи.

Работы выполняются в рамках Государственного задания ГГМ РАН по теме № 0140-2019-0005 «Разработка информационной среды интеграции данных естественнонаучных музеев и сервисов их обработки для наук о Земле», а также темы государственного задания № 1021061009468-8-1.5.1 «Цифровая платформа интеграции и анализа геологических и музейных данных».

СПИСОК ЛИТЕРАТУРЫ

1. *Kate A.* Машинное обучение и искусственный интеллект в геологии // Золотодобыча, №257, Апрель, 2020, пер.

2. *Kaplan A., Haenlein M.* Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence // Business Horizons. 2019. V. 62, No. 1. P. 15–25.

3. PROSPECTOR // URL: http://www.computing.surrey.ac.uk/ai/PROFILE /prospector.html (дата обращения 18.09.2023)

4. ESRI // URL: https://www.esri.com/en-us/home (дата обращения 18.09.2023)

5. USGS // URL: https://www.usgs.gov/ (дата обращения 18.09.2023)

6. *Родионов С.М., Сыркин В.К.* Экспертная прогнозирующая система «Олово» // Тихоокеанская геология. 1995. Т. 14, №5. С. 63–71.

URL: http://itig.as.khb.ru/POG/archive/1995/N5_1995.pdf

7. SOLSA Expert System // URL: https://solsa-dem-up.eu/en (дата обращения 17.09.2023)

8. GoldSpot // URL: https://www.alsglobal.com/en/consulting-and-analytics (дата обращения 18.09.2023)

9. SRK Consulting // URL: https://www.srk.com/ru/ (дата обращения 18.09.2023)

10. Maptek // URL: https://www.maptek.com/ (дата обращения 18.09.2023)

11. IOS Services Geoscientifiques // URL: https://www.iosgeo.com/en/ (дата обращения 18.09.2023)

12. Orefox // URL: https://orefox.com/ (дата обращения 18.09.2023)

13. Geolearn // URL: https://www.geolearn.ai/ (дата обращения 18.09.2023)

14. Datarock // URL: https://datarock.com.au/platform/ (дата обращения 18.09.2023)

15. *Baraboshkin E.E., Ismailova L.S., Orlov D.M., Zhukovskaya E.A., Kalmykov G.A., Khotylev O.V., Baraboshkin E.Yu., Koroteev D.A.* Deep Convolutions for In-Depth Automated Rock Typing // Computers & Geosciences. 2020. V. 135. https://doi.org/10.1016/j.cageo.2019.104330

16. *Nesteruk S., Agafonova J., Pavlov I., Gerasimov M., Latyshev N., Dimitrov D., Kuznetsov A., Kadurin A., Plechov P.* MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description // Computers & Geosciences. 2023. V. 178. https://doi.org/10.1016/j.cageo.2019.104330

17. Обработка естественного языка // URL: https://ru.wikipedia.org/wiki/Обработка_естественного_языка (дата обращения 18.09.2023)

18. *Jurafsky D., Martin J.H.* N-gram Language Models // Speech and Language Processing 3rd. 2021.

19. *Deng C., Zhang T., He Z., Chen Q., Shi Y., Zhou L., Fu L., Zhang W., Wang X., Zhou C., Lin Z., He J.* Learning Foundation Language Models for Geoscience Knowledge Understanding and Utilization // arXiv:2306.05064, 2023. URL: https://arxiv.org/abs/2306.05064v1

20. K2 model // URL: https://github.com/davendw49/k2?ysclid=lmswxywt6i750905070 (дата обращения 18.09.2023)

21. *Lawley C.J.M., Raimondo S., Chen T., Brin L., Zakharov A., Kur D., Hui J., Newton G., Burgoyne S.L., Marquis G.* Geoscience language models and their intrinsic evaluation // Applied Computing and Geosciences. 2022. V. 14, 100084. P. 1–10.

22. *Wang B., Ma K., Wu L., Qiu Q., Xie Z., Tao L.* Visual analytics and information extraction of geological content for text-based mineral exploration reports // Ore Geology Reviews. 2022. V. 144, 104818. P. 1–12.

23. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. V. 5. P. 177–187.

24. *Lawley C.J.M., Gadd M.G., Parsa M., Lederer G.W., Graham G.E., Ford A.* Applications of Natural Language Processing to Geoscience Text Data and Prospectivity Modeling // Natural Resources Research. 2023. V. 32, No. 4. P. 1503–1527.

25. *Fuentes I., Padarian J., Iwanaga T., Vervoort R.W.* 3D lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. V. 141, 104516.

26. *Qiu Qinjun, Xie Zhong, Wu Liang, Li Wenjia.* Geoscience keyphrase extraction algorithm using enhanced word embedding // Expert Systems with Applications. 2019. V. 125. P. 157–169.

27. *Патук М.И., Наумова В.В., Ерёменко В.С.* Цифровой репозиторий "geologyscience.ru": открытый доступ к научным публикациям по геологии России. // Электронные библиотеки. 2020. Т. 23, № 6. С. 1324–1338. https://doi.org/10.26907/1562-5419-2020-23-6-1324-1338

28. *Патук М.И., Наумова В.В.* Построение цифровой системы управления геологическими знаниями для поддержки научных исследований. // Электронные библиотеки. 2022. Т. 25, № 2. С. 148–158. https://doi.org/10.26907/1562-5419-2022-25-2-148-158

29. *Bourke D.* 08. Natural Language Processing with TensorFlow. URL: https://dev.mrdbourke.com/tensorflow-deep-learning/08_introduction_to_nlp_in_tensorflow/ (дата обращения 18.09.2023)

30. mrdbourke / tensorflow-deep-learning. URL: https://github.com/mrdbourke/tensorflow-deep-learning (дата обращения 18.09.2023)

31. Pdfreader 0.1.12. URL: https://pypi.org/project/pdfreader/ (дата обращения 18.09.2023)

32. spaCy URL: https://spacy.io/models/ru (дата обращения 18.09.2023)

33. Spacy-stanza. URL: https://spacy.io/universe/project/spacy-stanza (дата обращения 18.09.2023)

34. SberDevice. Как мы анализируем предпочтения пользователей виртуальных ассистентов Салют. URL: https://habr.com/ru/companies/ sberdevices/articles/547568/ (дата обращения 18.09.2023)

35. Zero-shot learning.
URL: https://en.wikipedia.org/wiki/Zero-shot_learning (дата обращения 18.09.2023)

36. *Ванюшкин А.С., Гращенко Л.А.* Методы и алгоритмы извлечения ключевых слов // Новые информационные технологии в автоматизированных системах. 2016. С. 85–93.

37. *Pay T., Lucci F., Cox J.L.* An Ensemble of Automatic Keyword Extractors: TextRank, RAKE and TAKE // Computación y Sistemas. 2019. V. 23, No. 3. P. 703–710. https://doi.org/10.13053/CyS-23-3-3234

38. *Дале Д.* Многозадачная модель T5 для русского языка.
URL: https://habr.com/ru/articles/581932/ (дата обращения 18.09.2023)

39. *Данил*, keyT5 или генерация ключевых слов из текста.
URL: https://habr.com/ru/articles/599715/ (дата обращения 18.09.2023)

40. Yandex DataSphere. URL: https://datasphere.yandex.ru/?yc-skip-auth=1 (дата обращения 18.09.2023)

ARTIFICIAL INTELLIGENCE METHODS FOR SCIENTIFIC RESEARCH IN GEOLOGY

Mikhail I. Patuk[1] [0000-0003-3036-2275], Vera V. Naumova[2] [0000-0002-3001-1638]

[1, 2]*State Geological Museum named after Vladimir Vernadsky of RAS, Moscow*

[1]*patuk@mail.ru;* [2]*Naumova_new@mail.ru*

Abstract

A brief overview of some methods of artificial intelligence in the field of Earth sciences is given. The prospects of using these methods to obtain new knowledge are noted. The results of the authors' first attempts to apply natural language processing methods for processing scientific articles on geology are presented. The possibilities of developing work in this direction are discussed.

Keywords*:* Artificial intelligence, machine learning, natural language processing, geology.

REFERENCES

1. *Caté A.* Machine Learning and Artificial Intelligence for Mining Geoscience // Geological Association of Canada, 2019.

2. *Kaplan A., Haenlein M.* Siri, Siri in my Hand, who's the Fairest in the Land? On the Interpretations, Illustrations and Implications of Artificial Intelligence // Business Horizons. 2019. V. 62. No. 1. P. 15–25.

3. PROSPECTOR // URL: http://www.computing.surrey.ac.uk/ai/PROFILE /prospector.html (дата обращения 18.09.2023) (date of access 18.09.2023)

4. ESRI // URL: https://www.esri.com/en-us/home (date of access 18.09.2023)

5. USGS // URL: https://www.usgs.gov/ (date of access 18.09.2023)

6. *Rodionov S.M., Syrkin V.K.* Expert forecasting system "Olovo" // Geology of the Pasofic Ocean. 1995. V. 14, No. 5. P. 63–71.
URL: http://itig.as.khb.ru/POG/archive/1995/N5_1995.pdf

7. SOLSA Expert System // URL: https://solsa-dem-up.eu/en (date of access 18.09.2023)

8. GoldSpot // URL: https://www.alsglobal.com/en/consulting-and-analytics (date of access 18.09.2023)

9. SRK Consulting // URL: https://www.srk.com/ (date of access 18.09.2023)

10. Maptek // URL: https://www.maptek.com/ (date of access 18.09.2023)

11. IOS Services Geoscientifiques // URL: https://www.iosgeo.com/en/ (date of access 18.09.2023)

12. Orefox // URL: https://orefox.com/ (date of access 18.09.2023)

13. Geolearn // URL: https://www.geolearn.ai/ (date of access 18.09.2023)

14. Datarock // URL: https://datarock.com.au/platform/ (date of access 18.09.2023)

15. *Baraboshkin E.E., Ismailova L.S., Orlov D.M., Zhukovskaya E.A., Kalmykov G.A., Khotylev O.V., Baraboshkin E.Yu., Koroteev D.A.* Deep Convolutions for In-Depth Automated Rock Typing // Computers & Geosciences. 2020. V. 135. https://doi.org/10.1016/j.cageo.2019.104330

16. *Nesteruk S., Agafonova J., Pavlov I., Gerasimov M., Latyshev N., Dimitrov D., Kuznetsov A., Kadurin A., Plechov P.* MineralImage5k: A benchmark for zero-shot raw mineral visual recognition and description // Computers & Geosciences. 2023. V. 178. https://doi.org/10.1016/j.cageo.2019.104330

17. Natural language processing URL: https://en.wikipedia.org/wiki/ Natural_language_processing (date of access 18.09.2023)

18. *Jurafsky D., Martin J.H.* N-gram Language Models // Speech and Language Processing 3rd. 2021.

19. *Deng C., Zhang T., He Z., Chen Q., Shi Y., Zhou L., Fu L., Zhang W., Wang X., Zhou C., Lin Z., He J.* Learning Foundation Language Models for Geoscience Knowledge Understanding and Utilization // arXiv:2306.05064, 2023. URL: https://arxiv.org/abs/2306.05064v1

20. K2 model // URL: https://github.com/davendw49/k2?ysclid=lmswxywt6i750905070 (date of access 18.09.2023)

21. *Lawley C.J.M., Raimondo S., Chen T., Brin L., Zakharov A., Kur D., Hui J., Newton G., Burgoyne S.L., Marquis G.* Geoscience language models and their intrinsic evaluation // Applied Computing and Geosciences. 2022. V. 14, 100084. P. 1–10.

22. *Wang B., Ma K., Wu L., Qiu Q., Xie Z., Tao L.* Visual analytics and information extraction of geological content for text-based mineral exploration reports // Ore Geology Reviews. 2022. V. 144, 104818. P. 1–12.

23. *Padarian J., Fuentes I.* Word embeddings for application in geosciences: development, evaluation, and examples of soil-related concepts // SOIL. 2019. V. 5. P. 177–187.

24. *Lawley C.J.M., Gadd M.G., Parsa M., Lederer G.W., Graham G.E., Ford A.* Applications of Natural Language Processing to Geoscience Text Data and Prospectivity Modeling // Natural Resources Research. 2023. V. 32, No. 4. P. 1503–1527.

25. *Fuentes I., Padarian J., Iwanaga T., Vervoort R.W.* 3D lithological mapping of borehole descriptions using word embeddings // Computers & Geosciences. 2020. V. 141, 104516.

26. *Qiu Qinjun, Xie Zhong, Wu Liang, Li Wenjia.* Geoscience keyphrase extraction algorithm using enhanced word embedding // Expert Systems with Applications. 2019. V. 125. P. 157–169.

27. *Patuk M.I., Naumova V.V., Eryomenko V.S.* Digital repository "geology-science.ru": open access to scientific publications on russian geology // Russian Digital Library Journal. 2020. V. 23, No. 6. P. 1324–1338. https://doi.org/10.26907/1562-5419-2020-23-6-1324-1338

28. *Patuk M.I., Naumova V.V.* Building a digital geological knowledge management system to support scientific research // Russian Digital Library Journal. 2022. V. 25, No. 2. P. 148–158. https://doi.org/10.26907/1562-5419-2022-25-2-148-158

29. *Bourke D.* 08. Natural Language Processing with TensorFlow, URL: https://dev.mrdbourke.com/tensorflow-deep-learning/08_introduction_to_nlp_in_tensorflow/ (date of access 18.09.2023)

30. mrdbourke / tensorflow-deep-learning URL: https://github.com/mrdbourke/ tensorflow-deep-learning (date of access 18.09.2023)

31. Pdfreader 0.1.12. URL: https://pypi.org/project/pdfreader/ (date of access 18.09.2023)

32. spaCy. URL: https://spacy.io/models/ru (date of access 18.09.2023)

33. Spacy-stanza. URL: https://spacy.io/universe/project/spacy-stanza (date of access 18.09.2023)

34. SberDevice, How do we analyze the preferences of users of virtual assistants Salute. URL: https://habr.com/ru/companies/sberdevices/articles/547568/ (date of access 18.09.2023)

35. Zero-shot learning. URL: https://en.wikipedia.org/wiki/Zero-shot_learning (date of access 18.09.2023)

36. *Vanushkin A.S., Graschenko L.A.* Methods and algorithms for keyword extraction // New information technologies in automated systems. 2016. P. 85–93.

37. *Pay T., Lucci F., Cox J.L.* An Ensemble of Automatic Keyword Extractors: TextRank, RAKE and TAKE // Computación y Sistemas. 2019. V. 23, No. 3. P. 703–710. https://doi.org/10.13053/CyS-23-3-3234

38. *Dale D.* Multitasking model T5 for Russian. URL: https://habr.com/ru/articles/581932/ (date of access 18.09.2023)

39. *Danil*, keyT5 or generating keywords from text. URL: https://habr.com/ru/ articles/599715/ (date of access 18.09.2023)

40. Yandex DataSphere URL: https://datasphere.yandex.ru/?yc-skip-auth=1 (date of access 18.09.2023)

СВЕДЕНИЯ ОБ АВТОРАХ

ПАТУК Михаил Иванович – к. г.-м. н., и. о. н. с., научный отдел Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Michail I. PATUK – PhD, scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: patuk@mail.ru

ORCID: 0000-0003-3036-2275

НАУМОВА Вера Викторовна – д. г.-м. н., г. н. с., зав. Научным отделом Государственного геологического музея им. В.И. Вернадского РАН, Москва, Россия.

Vera V. NAUMOVA – Prof., head of SGM scientific department, Vernadsky State Geological Museum RAS, Moscow (Russia).

Email: naumova_new@mail.ru

ORCID: 0000-0002-3001-1638

Материал поступил в редакцию 22 сентября 2023 года

УДК 004.414.3

РАЗРАБОТКА МЕТОДОВ И ПРОГРАММНЫХ ИНСТРУМЕНТОВ ФОРМИРОВАНИЯ ЦИФРОВОГО ПОРТРЕТА УЧАЩИХСЯ

М. А. Солнцев[1] [0009-0002-4106-3035], М. М. Абрамский[2] [0000-0003-3063-8948]

[1, 2]*Институт информационных технологий и интеллектуальных систем, Казанский (Приволжский) федеральный университет.*

[1]mrt.solncev@gmail.com, [2]ma@it.kfu.ru

Аннотация

Рассмотрены вопросы возможности использования данных об обучающихся, представленных в электронном виде, для построения цифрового портрета. Предложен набор характеристик, необходимых для его построения, обозначена модель данных. Реализованы инструменты сбора данных об обучающихся из социальных сетей и других интернет-ресурсов. Предложены алгоритмы построения цифрового портрета. Проиллюстрировано применение алгоритмов машинного обучения для этих задач. Приведены примеры использования цифрового портрета в образовании.

Ключевые слова: социальные сети, сбор данных, портрет пользователя, образование

ВВЕДЕНИЕ

Согласно Постановлению Правительства Российской Федерации «О проведении эксперимента по внедрению цифровой образовательной среды» на территории отдельных субъектов РФ будут организованы мероприятия по внедрению ЦОС [1]. В рамках этого эксперимента планируется проверка применения возможности формирования «цифрового профиля обучающегося». Цифровой профиль станет обязательным, его будут регистрировать при первом обращении за образовательной услугой, например, при зачислении в детский сад или первый класс школы. В связи с этим становится актуальным вопрос формирования цифрового портрета обучающегося.

В настоящей работе представлены методы и программные инструменты

формирования цифрового портрета обучающихся. Для построения цифрового портрета могут быть использованы информация о социальной активности, а также цифровой след.

Решением схожих задач занимаются системы, созданные для поиска целевой аудитории в социальных сетях. В последнее время их количество стремительно растет, к наиболее популярным системам такого характера можно отнести «Pepper.ninja», «Segmento Target», «Target Hunter» и «Церебро Таргет» [2–5].

В настоящей работе рассмотрено использование текстовой, графической и медиа информации.

Статья построена следующим образом.

В разделе 1 выделены основные источники данных и характеристики, используемые для построения цифрового портрета, отмечены особенности их хранения.

В разделе 2 описаны способы извлечения и алгоритмы обработки данных, участвующих в построении.

Третий раздел посвящен методам и алгоритмам анализа данных, используемых для построения цифрового портрета, рассмотрены способы их применения.

1. ИСТОЧНИКИ ДАННЫХ ДЛЯ ЦИФРОВОГО ПОРТРЕТА

Задачи, связанные со сбором и анализом данных с целью последующего нахождения в них полезной информации, принято относить к типу *Data Mining* задач. Для решения такого класса задач в основном используются технологии *краулинга* и *скрейпинга*, а также методы *машинного обучения* [6–8].

В задачах формирования цифрового портрета личности большой популярностью пользуются алгоритмы классификации и кластерного анализа, а также ассоциативные правила.

В современном мире популярность социальных сетей растет с каждым годом всё динамичнее, во многом по этой причине их количество увеличивается. Вместе с этим люди оставляют всё больший цифровой след, который может точно описать их личность: характер, взгляды и интересы. Результаты статистического исследования ресурса statista.com, приведенные на рис. 1, показывают, что в настоящее время жители России активно используют порядка 15 социальных се-

тей каждый день [9]. Эти результаты свидетельствуют, что наибольшей популярностью пользуются такие социальные сети, как YouTube и ВКонтакте. Ниже рассмотрена социальная сеть ВКонтакте, так как в ней пользователи помимо подписок на различные тематические сообщества и каналы могут публиковать собственные медиа и текстовые публикации – это дает возможность применить множество методов анализа данных для получения полезной информации. Например, текстовая информация может быть использована для математической лингвистики с целью классификации публикуемых текстов.

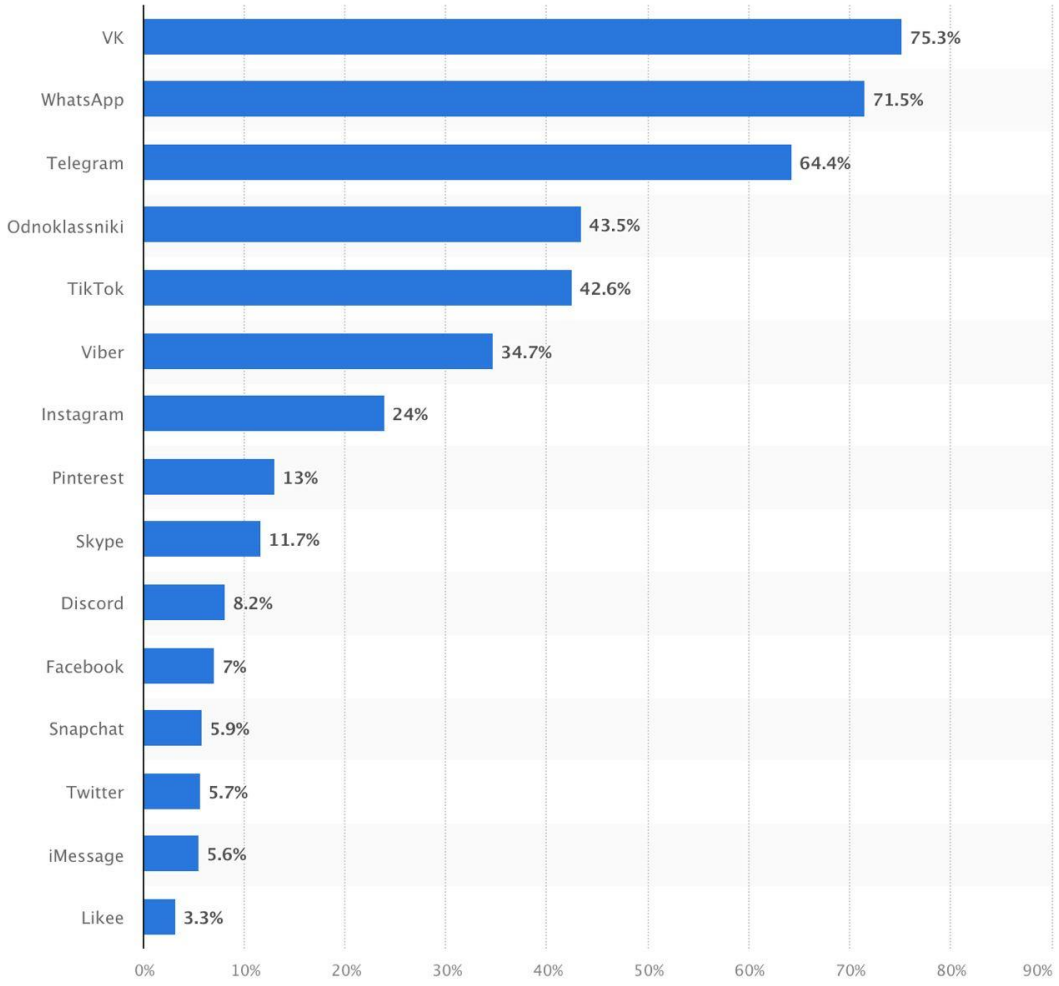

Рис. 1. Использование социальных сетей в России в 2022 году [9]

Помимо этого, при построении цифрового портрета необходимо учитывать информацию о получаемом образовании и курсах дополнительного образования. Успешность дополнительной образовательной и спортивной деятельности может быть подтверждена результатами выступлений на различных олимпиадах,

конкурсах и соревнованиях. Для этого могут быть использованы финальные протоколы участия, которые могут быть получены у организаторов такого рода мероприятий. Протоколы участия в большинстве случаев предоставляются в формате Excel и имеют вид, представленный на рис. 2.

	Фамилия уч	Имя	Отчество		Район об	Образовате	Педагог		пи	у	и	Статус
2	Гайнуллин	Әмир	Илнурович	2	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	8	32	40	Победитель
3	Имамиев	Камиль	Ильназович	2	Московский	МБОУ Татарска	Нагимулли	Оли	8	33	41	Победитель
4	Маулиханова	Ралина	Ранисовна	4	Тетюшский	МБОУ "Тетюшская СОШ №		Оли	6	29	35	Победитель
5	Гарипова	Исламия	Марселевна	1	Авиастроите	МБОУ "Гимназ	Закирова Л	Оли	7	16	23	Призер
6	Кашапова	Азалия	Айратовна	3	Кировский	МБОУ "СОШ"5	Шакирова	Оли	10	25	35	Призер
7	Низамова	Джамиля	Дамировна	1	Приволжски	МАОУ "Гимназ	Галаветдин	Оли	8	19	27	Призер
8	Рамазанова	Амина	Ренатовна	2	Кировский	МБОУ"Лицей 1	Мингалиев	Оли	7	23	30	Призер
9	Хамзина	Ранелия	Радиковна	2	Кировский	МБОУ "Полите	Мингалиев	Оли	7	30	37	Призер
10	Шагитова	Камиля	Ильфатовна	4	Лаишевский	МБОУ Пелевск	Шагитова Л	Оли	6	22	28	Призер
11	Шарифуллина	Диляра	Айратовна	4	Приволжски	МБОУ "Школа	Яковлева Р	Оли	10	20	30	Призер
12	Шафикова	Дина	Альбертовна	2	Авиастроите	МБОУ "Школа	Муктат Флё	Оли	4	29	33	Призер
13	Абдуллина	Самира	Наилевна	4	Зеленодоль	МБОУ "Гимназ	Зиганшина	Оли	9	14	23	Участник
14	Абдуллина	Алина	Рустамовна	4	Кировский р	МБОУ Политех	Гибадулли	Оли	4	12	16	Участник
15	Байбалаева	Самира	Баходуровна	4	Зеленодоль	МБОУ "Гимназ	Зиганшина	Оли	5	16	21	Участник
16	Билалова	Руфина	Рифатовна	3	Кировский р	МБОУ Политех	Гибадулли	Оли	4	28	32	Участник
17	Габдрахманова	Гульназ	Ильгамовна	4	Зеленодоль	МБОУ "Гимназ	Зиганшина	Оли	5	19	24	Участник
18	Галимов	Карим	Тауфийкович	1	Ново-Савин	Гимназия №155		Оли	7	11	18	Участник
19	Гарифуллина	Самира	Ильгамовна	4	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	2	13	15	Участник
20	Гилманова	Амелия	Марселевна	3	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	4	21	25	Участник
21	Гимадиев	Самир	Русланович	3	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	4	24	28	Участник
22	Ибрагимова	Газиза	Айтугановна	3	Кировский	МБОУ"Лицей 1	Мингалиев	Оли	8	7	15	Участник
23	Идиатов	Камил	Ринатович	2	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	3	14	17	Участник
24	Исмагилова	Зилә	Ленаровна	1	Казань,Сове	МБОУ "Гимназ	Насибулли	Оли	4	13	17	Участник
25	Касимова	Джамиля	Дамировна	2	Ново-Савин	МБОУ "Школа	Галимова Л	Оли	3	19	22	Участник
26	Латыпова	Диляра	Ильшатовна	4	Кировский р	МБОУ Политех	Гибадулли	Оли	3	21	24	Участник
27	Лотфуллина	Самира	Алмазовна	3	Кировский р	МБОУ Политех	Гибадулли	Оли	5	19	24	Участник
28	Любицкая	Сафия	Вадимовна	4	Советский	МБОУ "Гимназ	Насибулли	Оли	3	14	17	Участник
29	Масгутова	Алина	Фанилевна	4	Кировский р	МБОУ Политех	Гибадулли	Оли	6	15	21	Участник

Рис. 2. Пример финального протокола олимпиады

В Таблице 1 представлена модель данных, используемая при построении цифрового портрета.

Таблица 1. Модель данных для цифрового портрета

Хранимый объект	**Описание и источник**
ФИО	
Дата рождения	
Пол	
Текущее место обучения	

Посещаемые курсы дополнительного образования	
Результаты участия в различных конкурсах и олимпиадах	Информация из протоколов мероприятий и олимпиад
Образование	Указанная информация об образовании
Идентификаторы в социальных сетях	Идентификаторы в ВКонтакте
Информация в разделе «О себе»	Содержимое поля «О себе» из профилей в ВКонтакте
Интересы	Содержимое поля «Интересы» в ВКонтакте
Родной город	Родной город, указанный в ВКонтакте
Город проживания	Город, указанный в ВКонтакте
Знание языков	Содержимое поля «Языки» в ВКонтакте
Опыт работы	Указанные места работы и стаж работы
Любимые книги	Содержимое поля «Любимые книги»
Любимые фильмы	Содержимое поля «Любимые фильмы»
Список понравившихся публикаций	Список публикаций, которые пользователь пометил отметкой «Мне нравится» или разместил на своей странице
Список сообществ	Список сообществ, на которые подписан пользователь в ВКонтакте
Список медиа публикаций	Список медиа публикаций в ВКонтакте и их метаинформация (локация, хештеги, содержимое изображения и т. д.)

2. ИСПОЛЬЗУЕМЫЕ МЕТОДЫ СБОРА И ОБРАБОТКИ ДАННЫХ

Информация, выбранная для построения цифрового портрета, находится в разных источниках информации, имеет разные структуру и формат. Поэтому для

её получения необходимо поддерживать различные методы, используя:

- *API* внешнего ресурса;
- сканирование страниц с информацией;
- обработку документов в цифровом формате;
- ручной ввод и корректировку данных.

В случае, когда внешний источник информации предоставляет открытый *API*, информация может быть получена путем отправки соответствующего HTTP *REST* запроса. Ответ в таком случае будет получен в формате *JSON* и может быть использован без преобразований.

В случае, когда внешний источник информации не обладает открытым *API*, необходимо использовать специальные техники получения данным путем *краулинга* и *скрейпинга*. В таком случае необходимо обрабатывать статические HTML-страницы для получения интересующей информации в более понятном формате (например, JSON).

Еще одним рассматриваемым источником данных являются документы, представленные в цифровом или бумажном форматах. Бумажные документы для автоматической обработки необходимо предварительно оцифровать путем сканирования. Оцифрованные документы могут быть обработаны с использованием различных библиотек, независимо от формата документа (Excel, PDF и т. д.).

Для получения информации из профилей социальных сетей используется несколько методов. Для анализа извлеченной информации применяются методы машинного обучения: так, для анализа медиа контента из ВКонтакте применяется нейронная сеть. Схема преобразований данных изображена на рис. 3.

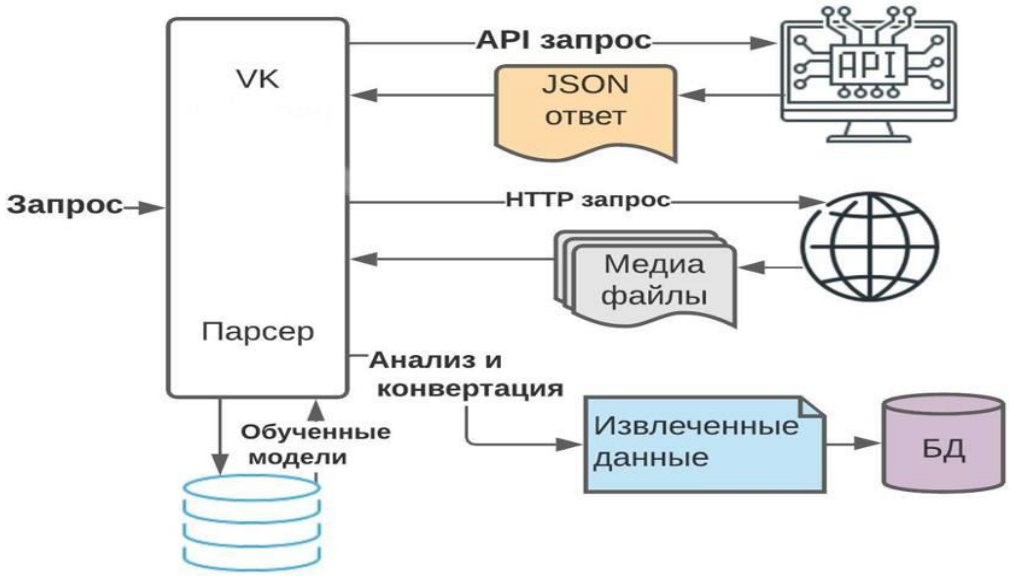

Рис. 3. Схема обработки данных из социальных сетей

Для обработки данных об учебной и творческой успешности используется информация о результатах участия в мероприятиях. Так как организаторы таких мероприятий, как правило, не имеют открытого API, необходимо применять скрейпинг их сайта. Такой метод позволяет извлечь из HTML-страниц необходимые URL-адреса файлов с финальными протоколами. После этого файлы, находящиеся в формате Excel, преобразуются в JSON. Соответствующая схема действий изображена рис. 4.

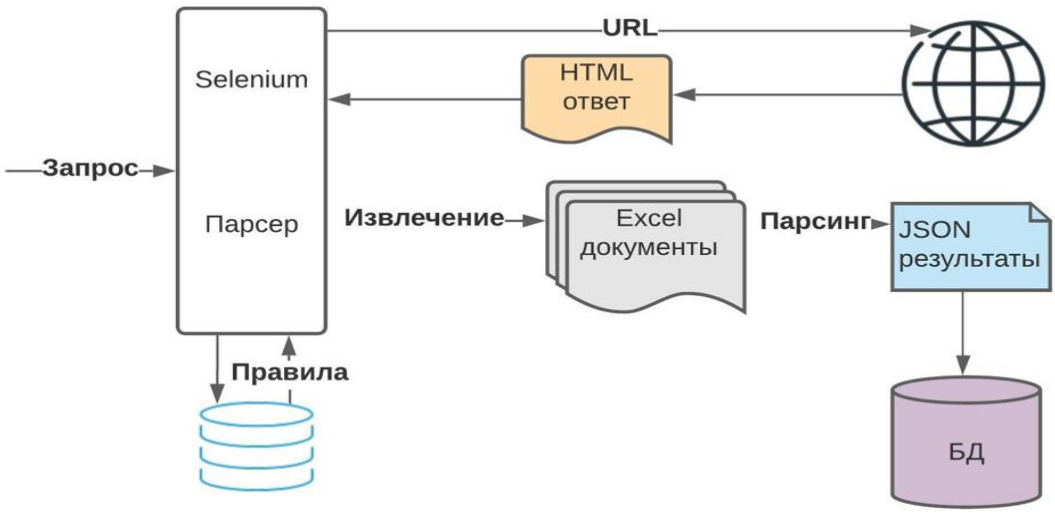

Рис. 4. Схема извлечения данных о достижениях

3. АНАЛИЗ ДАННЫХ ДЛЯ ЦИФРОВОГО ПОРТРЕТА

Анализ данных социальных сетей: данный алгоритм позволяет собирать и анализировать данные из профилей социальных сетей. Сервис состоит из трех модулей:

- анализа социальной активности – отвечает за получение и обработку данных из профилей «ВКонтакте» (рис. 5);
- анализа медиа контента – отвечает за анализ медиа информации из профилей (рис. 6);
- анализа тональности текстовых публикаций.

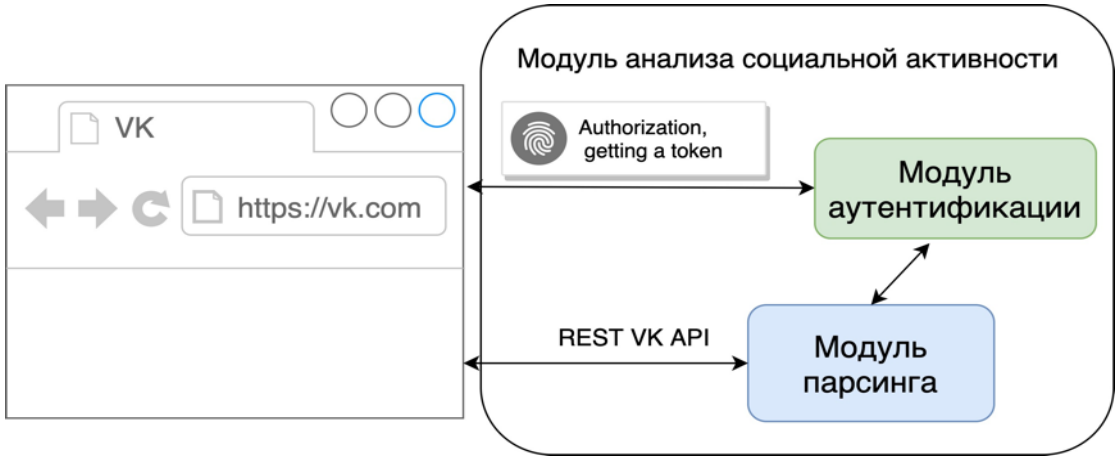

Рис. 5. Архитектура модуля анализа социальной активности

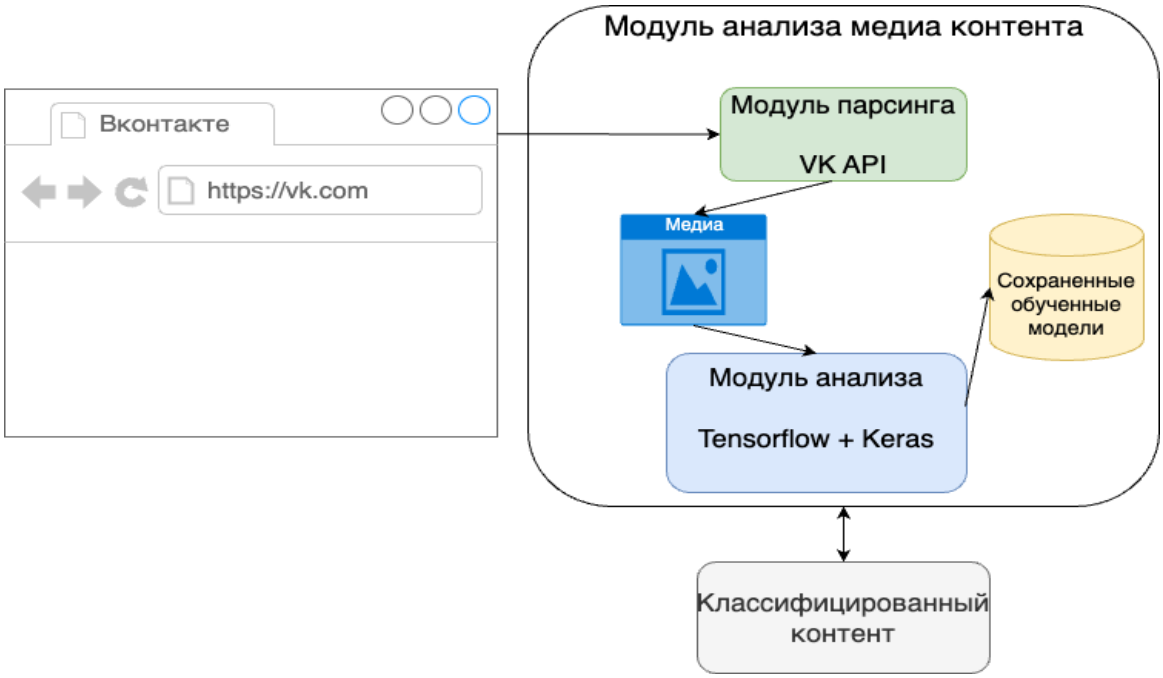

Рис. 6. Модуль анализа медиа контента

Сервис взаимодействует с внешними сервисами («ВКонтакте»), используя предоставляемый ими открытый API.

Для начала работы с VK API [10] необходимо зарегистрировать сервисное приложение в «ВКонтакте», от его лица совершаются все запросы для получения данных. Приложение имеет ряд настроек, которые в дальнейшем могут быть изменены. Зарегистрированному приложению присваиваются уникальный идентификатор, защищенный ключ и сервисный ключ доступа (рис. 7), которые используются в системе клиентом VK SDK [11].

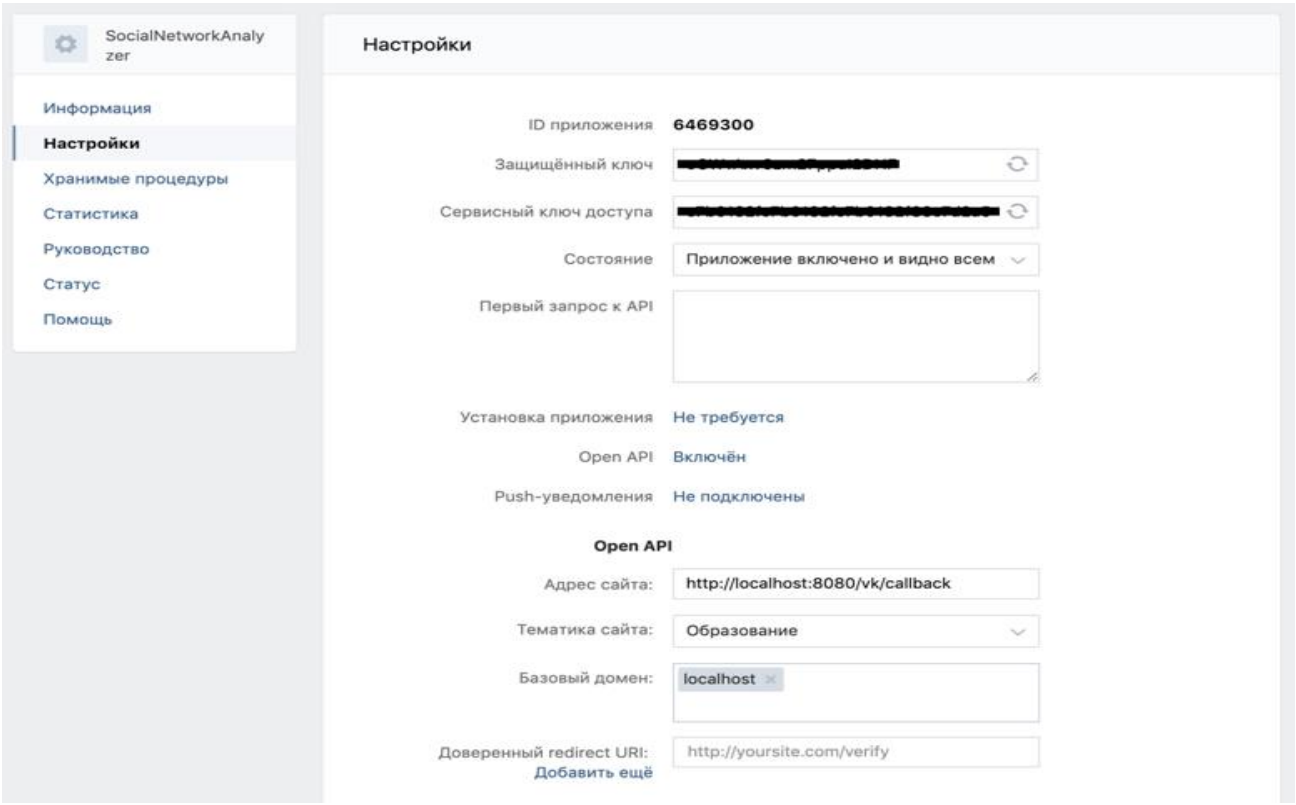

Рис. 7. Настройки параметров приложения ВКонтакте

Для начала выполнения запросов клиенту, выполняющему запросы к VK API, необходимо авторизоваться. Для этого используется OAuth-аутентификация (рис. 8).

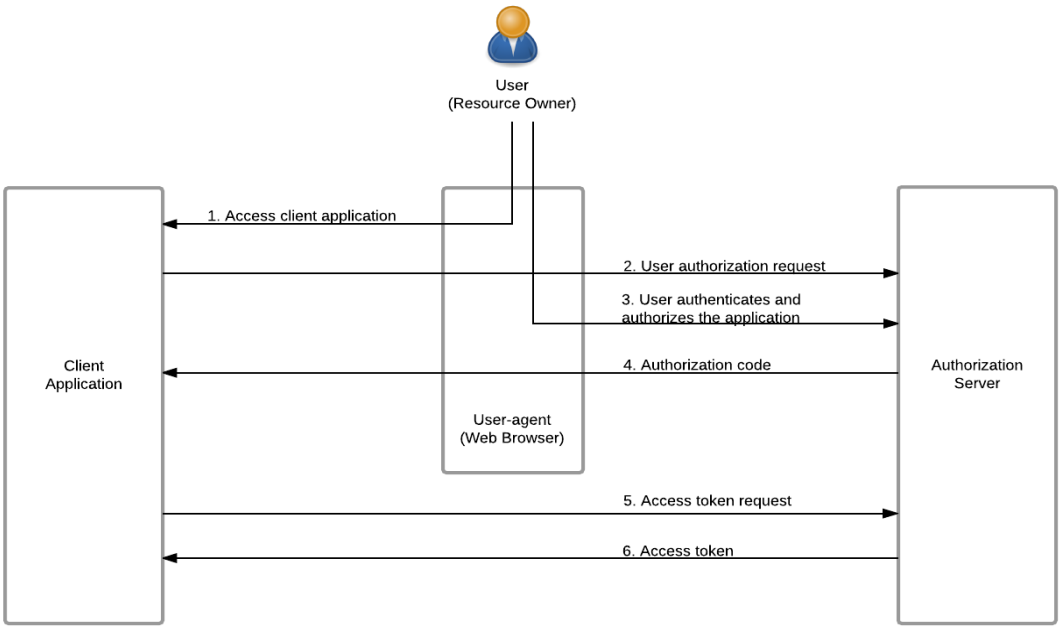

Рис. 8. Auth Code Flow для получения ключа доступа

Для более удобного процесса авторизации приложения используется ScribeJava [12] – клиент для работы с OAuth-авторизацией [13]. Для того чтобы клиент смог получить токен доступа, ему необходимо передать: идентификатор приложения; список разделов, к которым будет иметь доступ система при выполнении запросов; секретный ключ доступа и Callback URL. После получения токена система может выполнять запросы на получения данных. Пример ответа VK API изображен на рис. 9. VK API придерживается архитектуры REST [14], поэтому возвращает данные в формате JSON [15].

```
{
  "response": [{
    "first_name": "Lindsey",
    "id": 210700286,
    "last_name": "Stirling",
    "can_access_closed": true,
    "is_closed": false,
    "photo_50": "https://sun7-9.us...8,641,641&ava=1",
    "verified": 1,
    "city": {
      "id": 5331,
      "title": "Los Angeles"
    },
    "interests": "Family, Friends, Dancing, Music",
    "about": "http://www.lindse...com/LindseyStirling",
    "career": [],
    "university": 0,
    "university_name": "",
    "faculty": 0,
    "faculty_name": "",
    "graduation": 0
  }]
}
```


Рис. 9. Пример ответа VK API

Это позволяет проанализировать группы и сообщества группы пользователей ВКонтакте. Для этого собираются данные о подписках пользователей на различные страницы, определяются их название, тематика и описание. Для каждого пользователя подсчитывается количество групп, принадлежащих к заранее определенном наборам тематик. В данной системе были выделены две основные группы тематик: группы с развлекательным характером и группы, связанные с образованием, личностным и профессиональным ростом. К первой группе были отнесены сообщества с тематиками: 'Покупки', 'Туризм, путешествия', 'Развлечения', 'СМИ', 'Спорт', 'Юмор', 'Шоу, передача', 'Игры', 'Стиль, одежда, обувь', 'Музыка', 'Кино', 'Веб-сайт', 'Обмен музыкой', 'Интернет-СМИ', 'Городское сообщество', 'Искусство и развлечения', 'Молодёжное движение', 'Музыкальная группа'; ко второй – 'Искусство', 'IT', 'Наука', 'Образование', 'Саморазвитие', 'Техника', 'Экономика', 'Языки', 'Бизнес', 'Дизайн и графика', 'История', 'Финансы', 'Культура',

'Философия', 'Обучающие курсы', 'Литература', 'Творчество', 'Фотография', 'Культурный центр', 'Программное обеспечение', 'Образовательное учреждение', 'Программирование'. Данные списки в дальнейшем могут быть скорректированы, также можно добавить большее количество подгрупп тематик. После того, как выделены группы тематик, подсчитывается количество сообществ пользователя, принадлежащих к той или иной группе. На основе данных о процентном содержании сообществ всех подгрупп составляется характеристический вектор студента. Размерность вектора определяется количеством подгрупп сообществ, значения – долей сообществ соответствующей группы.

Для получения сведений о заинтересованности пользователей в определенных областях, нахождения тенденций и связей интересов пользователей используется кластеризация. Система применяет метод К-средних, который является одним из самых популярных и простых в реализации. Данный метод позволяет разделить множество объектов, имеющих определенные свойства, на количество кластеров, равное К [16]. Величина К является параметром и может задаваться вручную, но в системе определяется автоматически на основании мощности кластеризуемого множества. В реализации модуля анализа данных была использована реализация метода К-средних Scikit-Learn [17]. На вход подаются характеристические векторы пользователей, построенные на основании данных о подписках, на выходе – список принадлежностей объекта кластеризации (пользователя) к определенному кластеру (рис. 10).

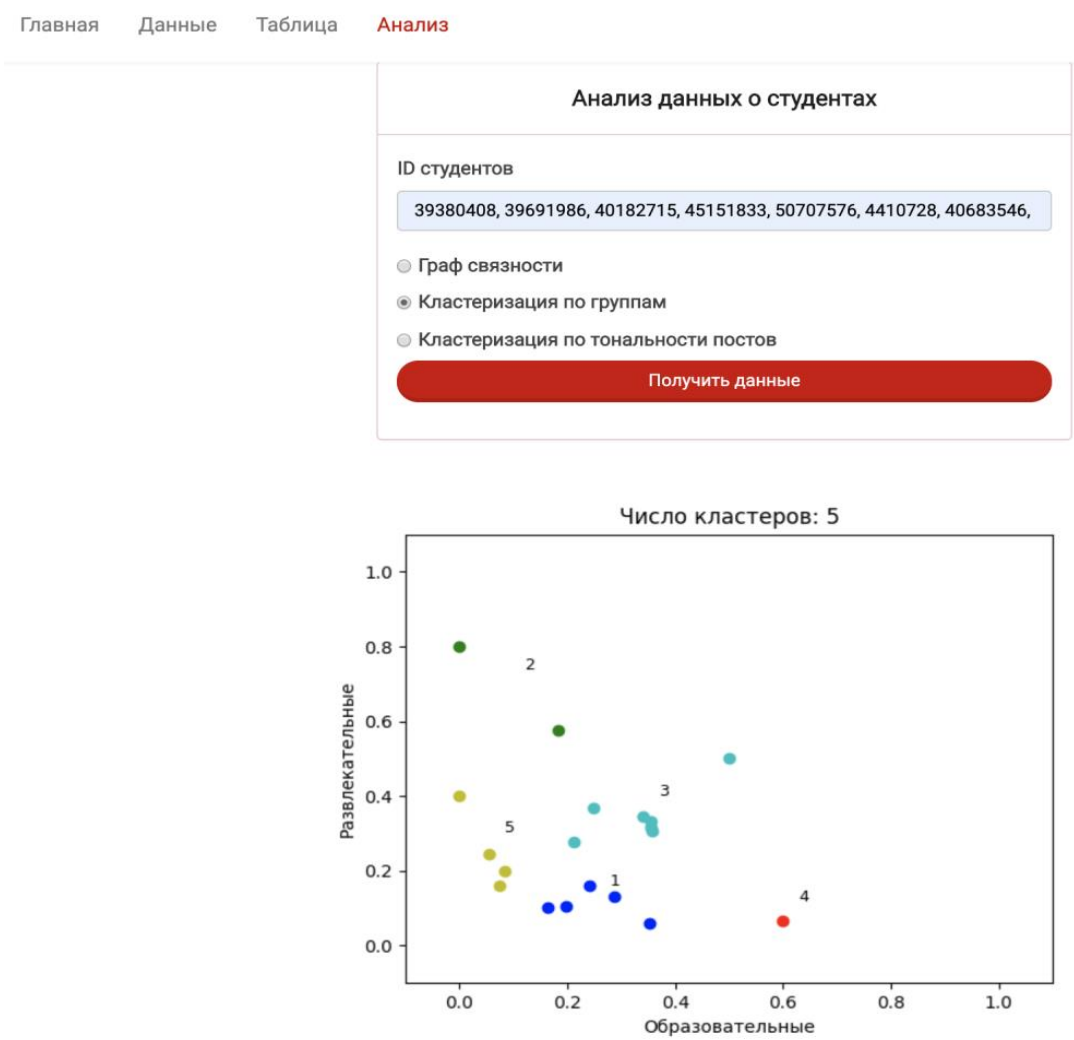

Рис. 10. Кластерный анализ по тематикам групп

Таким же способом получаются ссылки на медиаконтент пользователя. После того, как медиа файлы получены, они передаются в модуль анализа информации. Для классификации фото в нем используются библиотеки Numpy и Keras [18, 19]. В системе реализована многослойная сверточная сеть, обученная на датасете *cifar-100*, представленным Kaggle [20]. Этот датасет содержит 100 классов, модуль анализа определяет принадлежность медиа контента к одному из этих классов.

Для анализа текстовых публикаций используется модель SocialNetworkModel, которая поставляется в библиотеке с открытым исходным кодом Dostoevsky, предназначенной для анализа русскоязычных текстов [21]. Данная модель обучается на наборе текстов, оставленных в социальной сети ВКонтакте. Система использует эту модель для определения тональности текстов

публикаций группы пользователей. На вход подается список идентификаторов пользователей, затем из базы данных достаются тексты публикаций, выложенные этими пользователями. Тексты проходят обработку – они токенизируются, из полученного набора токенов удаляются стоп-слова, затем из оставшихся токенов выделяются леммы. По набору лемм обученная модель определяет тональность текста. Для каждого пользователя подсчитывается процентное соотношение текстов, имеющих позитивную, негативную и нейтральную тональности. Полученные сведения могут быть прочитаны в текстовом формате, а также могут быть отображены на графике. Осями графиков служат две выбранные тональности, каждая точка на графике – пользователь, координаты которой определяются количеством текстов соответствующей тональности.

Информация о тональности публикаций применяется для дальнейшей кластеризации группы пользователей. Для кластеризации пользователей по тональности публикаций также используется метод К-средних (рис. 11).

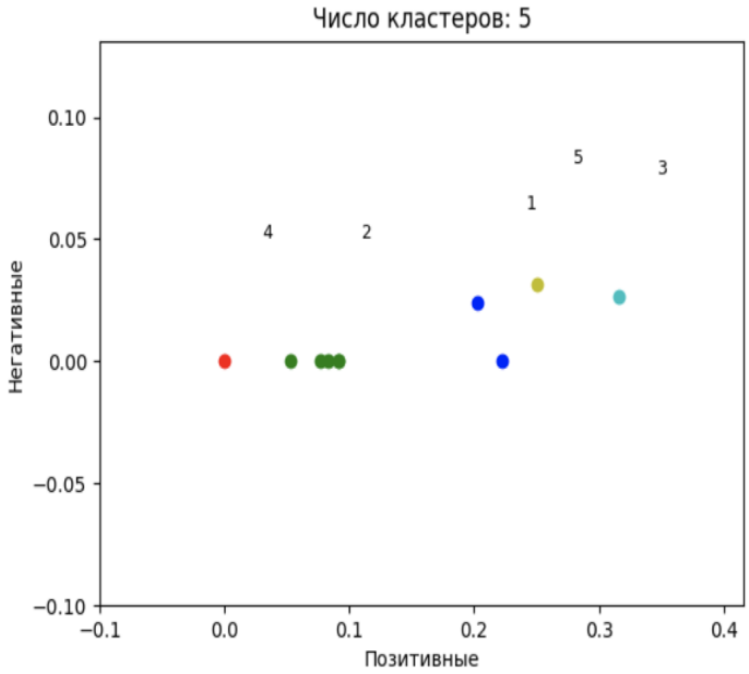

Рис. 11. Анализ тональности публикаций

Анализ результатов участия в мероприятиях

Данный алгоритм позволяет получать информацию о результатах выступлений на различных олимпиадах и конкурсах. Он состоит из двух модулей:

- парсинга содержимого страниц сайтов организаторов мероприятий;

- обработки результатов и протоколов, представленных в формате Excel.

Схема соответствующего алгоритма представлена на рис. 12.

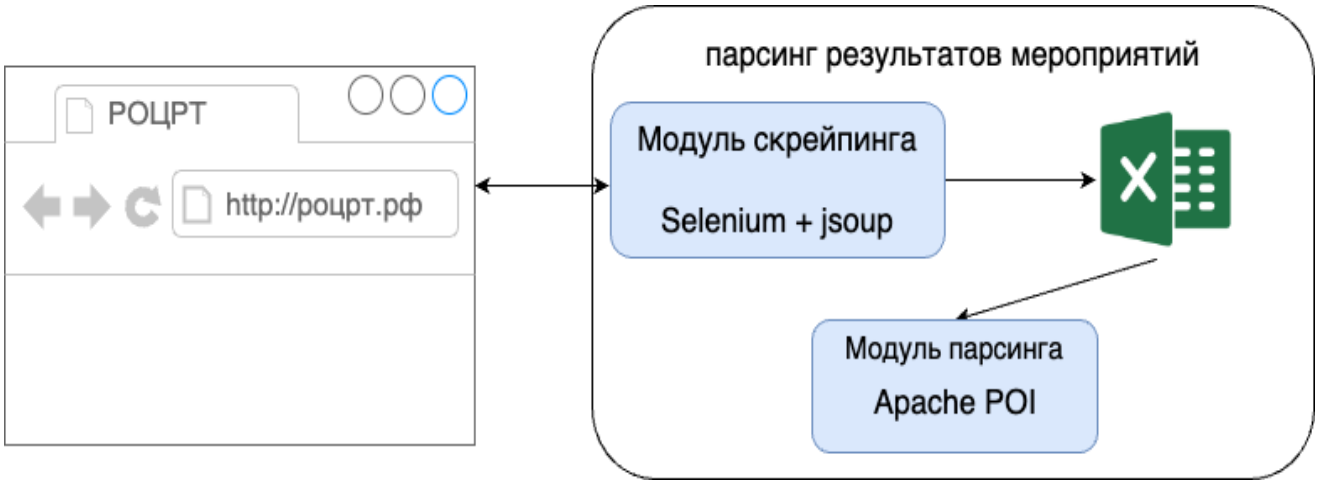

Рис. 12. Алгоритм парсинга результатов

Модуль парсинга страниц реализует нахождение URL необходимых файлов с результатами. Для прохождения по содержимому сайтов организаторов используются Selenium, а также библиотека jSoup [22, 23]. Selenium дает возможность выполнять JavaScript код для получения доступа к определённому полю, в то время как более легковесная библиотека Jsoup дает возможность получения объектов HTML-страниц путем обращения по селектору. Selenium – это веб-драйвер, который использует для работы браузер, поэтому для его запуска необходимо больше ресурсов. Поэтому в системе для получения данных из объектов разметки в первую очередь используется jsoup, и только в случае необходимости выполняются JavaScript скрипты. В случае увеличения количества рассматриваемых сайтов организаторов олимпиад и других мероприятий можно использовать специальные хранилища для правил обработки, в данной же системе для хранения правил используется только память сервера.

Для каждого из организаторов мероприятий в базе данных хранится набор правил для обработки страниц их сайтов. Правила включают в себя селекторы и JavaScript код, который необходимо запустить для получения ссылок на файлы с результатами. После применения правил модуль парсинга предоставляет список ссылок на файлы с результатами.

После того, как ссылки получены, модуль обработки результатов скачивает

файлы с результатами, используя HTTP-клиент. Дальше Excel файлы трансформируются в формат, который может быть использован для хранения и дальнейшего использования. Пример содержимого файла с результатами представлен на рис. 2. После этого содержимое файла фильтруется и считывается в JSON-строки, которые затем записываются в базу данных. Чтение Excel-таблиц в Json происходит с помощью библиотеки Apache POI [24].

Собранная информация позволяет группировать пользователей по заинтересованности в определенных тематиках, а также по успешности выступлений. Согласно работе [25] плотность взаимодействия может трактоваться как сплоченность группы. В системе реализована функциональность, позволяющая, используя информацию о взаимодействии группы пользователей, визуализировать степень социальной сплоченности группы пользователей.

ЗАКЛЮЧЕНИЕ

Рассмотрены источники информации, необходимые для построения цифрового портрета учащегося. Был выделен перечень используемых характеристик. Разработаны методы получения, обработки, и анализа рассматриваемых данных.

Созданы инструменты, оценивающие результаты выступлений учащегося в различных мероприятиях и анализирующие данные из профилей социальных сетей для построения цифрового портрета. Реализованы алгоритмы, позволяющие при построении цифрового портрета оценивать тематику и тональность публикуемых учащимся текстов. Предложены варианты использования цифрового портрета в образовательных целях.

СПИСОК ЛИТЕРАТУРЫ

1. Постановление Правительства Российской Федерации от 07.12.2020 № 2040 «О проведении эксперимента по внедрению цифровой образовательной среды». URL: https://open.edu.gov.ru/files/faq/subjects.pdf (дата обращения: 28.10.2023).

2. Pepper.ninja [Электронный ресурс]. URL: https://pepper.ninja/ (дата обращения: 28.10.2023).

3. Segmento Target [Электронный ресурс].

URL: https://segmento-target.ru/ (дата обращения: 28.10.2023).

4. TargetHunter [Электронный ресурс]. URL: https://targethunter.ru (дата обращения: 28.10.2023).

5. Церебро Таргет [Электронный ресурс]. URL: https://церебро.рф (дата обращения: 28.10.2023).

6. Top 50 open-source web crawlers for data mining [Электронный ресурс]. URL: https://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining (дата обращения: 28.10.2023).

7. 8 Best Web Scraping Tools [Электронный ресурс]. URL: https://hevodata.com/learn/8-best-web-scraping-tools/ (дата обращения: 28.10.2023).

8. Обзор алгоритмов Data Mining [Электронный ресурс]. URL: https://www.intuit.ru/studies/courses/6/6/info (дата обращения: 28.10.2023).

9. Статистический портал «Statista» [Электронный ресурс]. URL: https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/ (дата обращения: 28.10.2023).

10. VK API [Электронный ресурс]. URL: https://vk.com/apiclub (дата обращения: 28.10.2023).

11. VK Java SDK [Электронный ресурс]. URL: https://vk.com/dev/Java_SDK (дата обращения: 28.10.2023).

12. ScribeJava. Simple OAuth library for Java [Электронный ресурс]. URL: https://github.com/scribejava/scribejava (дата обращения: 28.10.2023).

13. OAuth authorization framework [Электронный ресурс]. URL: https://oauth.net (дата обращения: 28.10.2023).

14. REST. Representational State Transfer [Электронный ресурс]. URL: https://restfulapi.net/ (дата обращения: 28.10.2023).

15. JSON. JavaScript Object Notation [Электронный ресурс]. URL: https://www.json.org/ (дата обращения: 28.10.2023).

16. *Черезов Д.С., Тюкачев Н.А.* Обзор основных методов классификации и кластеризации данных // Вестник Воронежского государственного университета. Серия: Системный анализ и информационные технологии. 2009. №. 2. С. 25–29.

17. Scikit-Learn. Machine Learning in Python [Электронный ресурс].

URL: https://scikit-learn.org/stable (дата обращения: 28.10.2023).

18. Numpy. The fundamental package for scientific computing with Python [Электронный ресурс]. URL: https://numpy.org/ (дата обращения: 28.10.2023).

19. Keras. Python deep learning API [Электронный ресурс]. URL: https://keras.io/ (дата обращения: 28.10.2023).

20. Kaggle. the world's largest data science community [Электронный ресурс]. URL: https://keras.io/ (дата обращения: 28.10.2023).

21. Dostoevsky. Sentiment analysis library for Russian language [Электронный ресурс]. URL: https://github.com/bureaucratic-labs/dostoevsky (дата обращения: 28.10.2023).

22. Selenium. Automates browsers [Электронный ресурс]. URL: https://www.selenium.dev/ (дата обращения: 28.10.2023).

23. Jsoup. Java HTML Parser [Электронный ресурс]. URL: https://jsoup.org/ (дата обращения: 28.10.2023).

24. Apache POI. Java API for Microsoft Documents [Электронный ресурс]. URL: https://poi.apache.org/ (дата обращения: 28.10.2023).

25. *Печенкин В.В., Ярская-Смирнова Е.Р.* Сетевые подходы в анализе социальной сплоченности // Вестник Саратовского государственного технического университета. 2014. Т. 4. № 1 (77).

DEVELOPMENT OF METHODS AND SOFTWARE TOOLS FOR THE FORMATION OF A DIGITAL PORTRAIT OF STUDENTS

M. A. Solncev[1] [0009-0002-4106-3035], M. M. Abramskiy[2] [0000-0003-3063-8948]

[1, 2] *Institute of Information Technology and Intelligent Systems of Kazan Federal University*

[1]mrt.solncev@gmail.com, [2]ma@it.kfu.ru

Abstract

This paper considers the questions about the possibility of using data about the students presented in electronic form to build their digital portraits. A set of characteristics necessary for its construction is proposed, a data model is designated.

Implemented tools for collecting data about students from social networks and other Internet resources. Algorithms for constructing a digital portrait are proposed. The application of machine learning algorithms for these tasks is illustrated. Examples of the use of digital portraits in education are given.

Keywords: social networks, data retrieval, personal portrait of user, education

REFERENCES

1. Resolution of the Government of the Russian Federation dated 07.12.2020 No. 2040 "On conducting an experiment on the introduction of a digital educational environment". URL: https://open.edu.gov.ru/files/faq/subjects.pdf (date of access: 28.10.2023).

2. Project Pepper.ninja [Electronic resource]. URL: https://pepper.ninja/ (date of access: 28.10.2023).

3. Project Segmento Target [Electronic resource]. URL: https://segmento-target.ru/(date of access: 28.10.2023).

4. Project TargetHunter [Electronic resource]. URL: https://targethunter.ru (date of access: 28.10.2023).

5. Project Cerebro Target [Electronic resource]. URL: https://церебро.рф (date of access: 28.10.2023).

6. Top 50 open-source web crawlers for data mining [Electronic resource] // bigdata-madesimple.com.
URL: https://bigdata-madesimple.com/top-50-open-source-web-crawlers-for-data-mining (date of access: 28.10.2023).

7. 8 Best Web Scraping Tools [Electronic resource] // hevodata.com URL: https://hevodata.com/learn/8-best-web-scraping-tools/(date of access: 28.10.2023).

8. Data Mining Algorithms Overview [Electronic resource] // intuit.ru: URL: https://www.intuit.ru/studies/courses/6/6/info (date of access: 28.10.2023).

9. Leading social media platforms in Russia in 3rd quarter 2022, by monthly penetration rate [Electronic resource] // statista.com.
URL: https://www.statista.com/statistics/867549/top-active-social-media-platforms-in-russia/ (date of access: 28.10.2023).

10. Project VK API [Electronic resource] // vk.com.
URL: https://vk.com/apiclub (date of access: 28.10.2023).

11. VK Java SDK Library [Electronic resource] // vk.com. URL: https://vk.com/dev/Java_SDK (date of access: 28.10.2023).

12. ScribeJava. Simple OAuth library for Java [Electronic resource] // github.com. URL: https://github.com/scribejava/scribejava (date of access: 28.10.2023).

13. OAuth authorization framework [Electronic resource] // oauth.net. URL: https://oauth.net (date of access: 28.10.2023).

14. REST. Representational State Transfer [Electronic resource] // restfulapi.net. URL: https://restfulapi.net/ (date of access: 28.10.2023).

15. JSON. JavaScript Object Notation [Electronic resource] // json.org. URL: https://www.json.org/ (date of access: 28.10.2023).

16. *Cherezov D.S., Tyukachev N.A* Overview of the main methods of data classification and clustering // Bulletin of the Voronezh State University. Series: System Analysis and Information Technologies. 2009. No. 2. P. 25–29.

17. Scikit-Learn. Machine Learning in Python [Electronic resource] // scikit-learn.org. URL: https://scikit-learn.org/stable (date of access: 28.10.2023).

18. Numpy. The fundamental package for scientific computing with Python [Electronic resource] // numpy.org. URL: https://numpy.org/ (date of access: 28.10.2023).

19. Keras. Python deep learning API [Electronic resource] // keras.io. URL: https://keras.io/ (date of access: 28.10.2023).

20. Kaggle. the world's largest data science community [Electronic resource] // kaggle.com. URL: https://keras.io/ (date of access: 28.10.2023).

21. Dostoevsky. Sentiment analysis library for Russian language [Electronic resource] // github.com. URL: https://github.com/bureaucratic-labs/dostoevsky (date of access: 28.10.2023).

22. Selenium. Automates browsers [Electronic resource] // selenium.dev. URL: https://www.selenium.dev/ (date of access: 28.10.2023).

23. Jsoup. Java HTML Parser [Electronic resource] // jsoup.org. URL: https://jsoup.org (date of access: 28.10.2023).

24. Apache POI. Java API for Microsoft Documents [Electronic resource] // apache.org. URL: https://poi.apache.org/ (date of access: 28.10.2023).

25. *Pechenkin V.V., Yarskaya-Smirnova E.R.* Network approaches in the analysis of social cohesion // Bulletin of the Saratov State Technical University. 2014. No. 1 P. 77.

СВЕДЕНИЯ ОБ АВТОРАХ

СОЛНЦЕВ Марат Альбертович – аспирант Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета

Marat Albertovich SOLNTSEV – Master, post graduate (Institute of Information Technology and Intelligent Systems, Kazan Federal University).

email: mrt.solncev@gmail.com

ORCID: 0009-0002-4106-3035

АБРАМСКИЙ Михаил Михайлович – директор Института информационных технологий и интеллектуальных систем Казанского (Приволжского) федерального университета, кандидат технических наук.

Mikhail Mikhailovich ABRAMSKIY – director of the Institute of Information Technology and Intelligent Systems, Kazan Federal University, PhD (Cand Sci. – Tech.)

email: mabramsk@kpfu.ru

ORCID: 0000-0003-3063-8948

Материал поступил в редакцию 30 октября 2023 года