УДК 004.853
DOI: 10.26102/2310-6018/2024.47.4.024

Концепция и архитектура парсинга и хранения единой базы патентов и научных журнальных публикаций

С.А. Козина, И.А. Кулинченко, Д.М. Коробкин✉, С.А. Фоменков

Волгоградский государственный технический университет, Волгоград, Российская Федерация

Резюме. Существующие на текущий момент методы автоматизированного сбора данных, хотя и облегчают данный процесс, но зачастую сталкиваются с проблемами низкой надежности, эффективности и скорости. Нестабильность соединений, блокировки IP-адресов и изменения структуры сайтов приводят к потере данных и необходимости постоянного контроля процесса парсинга, что увеличивает затраты на поддержание и эксплуатацию подобных систем. В связи с этим, разработка новых подходов и инструментов для парсинга необходимой информации является весьма актуальной задачей, способной трансформировать область интеллектуального анализа данных. В статье рассмотрен процесс разработки системы парсинга информации патентных систем и сайтов физико-технических журналов с использованием современных технологий и подходов, а также представлены результаты проверки его работоспособности. Данный инструмент может быть полезен патентным ведомствам, исследователям, студентам, инженерам, ученым, работающим в рассматриваемой предметной области. Использование такой системы позволит открыть новые возможности для интеллектуального анализа данных и принятия стратегических решений в области инновационного развития, а также для глубокого анализа технологических трендов, выявления перспективных разработок и построения стратегий инновационного развития.

Ключевые слова: патенты, физико-технические журналы, парсинг, масштабируемость, отказоустойчивость.

Благодарности: Исследование выполнено за счет гранта Российского научного фонда № 24-21-20140, https://rscf.ru/project/24-21-20140/, и Администрации Волгоградской области.

Для цитирования: Козина С.А., Кулинченко И.А., Коробкин Д.М., Фоменков С.А. Концепция и архитектура парсинга и хранения единой базы патентов и научных журнальных публикаций. *Моделирование, оптимизация и информационные технологии.* 2024;12(4). URL: https://moitvivt.ru/ru/journal/pdf?id=1740 DOI: 10.26102/2310-6018/2024.47.4.024

Concept and architecture of parsing and storing a unified database of patents and scientific journal publications

S.A. Kozina, I.A. Kulinchenko, D.M. Korobkin✉, S.A. Fomenkov

Volgograd State Technical University, Volgograd, the Russian Federation

Abstract. The currently existing methods of automated data collection, although they facilitate this process, often face problems of low reliability, efficiency and speed. Unstable connections, blocking IP addresses and changes in the structure of sites lead to data loss and the need for constant monitoring of the parsing process, which increases the cost of maintaining and operating such systems. In this regard, the development of new approaches and tools for parsing the necessary information is a very urgent task that can transform the field of data mining. The article discusses the process of developing a module for parsing information from patent systems and websites of physics and technology journals using modern technologies and approaches, and also presents the results of checking its operability. This tool can be useful for patent offices, researchers, students, engineers, and scientists working in the subject area under consideration. The use of such a module will open up new opportunities for data mining and strategic

decision-making in the field of innovative development, as well as for in-depth analysis of technological trends, identification of promising developments and building innovative development strategies.

Keywords: patents, physics and technology journals, parsing, scalability, fault tolerance.

Acknowledgements: The study was supported by the grant of the Russian Science Foundation No. 24-21-20140, https://rscf.ru/project/24-21-20140/, and the Administration of the Volgograd Region.

For citation: Kozina S.A., Kulinchenko I.A., Korobkin D.M., Fomenkov S.A. Concept and architecture of parsing and storing a unified database of patents and scientific journal publications. *Modeling, Optimization and Information Technology*. 2024;12(4). (In Russ.). URL: https://moitvivt.ru/ru/journal/pdf?id=1740 DOI: 10.26102/2310-6018/2024.47.4.024

Введение

В современном мире, где инновации движут прогрессом, патентование становится ключевым ресурсом для научного и технологического развития [1]. Анализ патентной информации позволяет выявлять тенденции и перспективные направления развития технологий, определять уровень конкуренции и технологического лидерства в различных отраслях [2].

Однако сбор патентных данных для анализа, особенно из различных источников, зачастую является сложным и длительным процессом, требующим значительных временных и трудовых ресурсов. Используя существующие методы, часто можно столкнуться с проблемами низкой надежности, эффективности и скорости. Ручной поиск и сбор патентов из разрозненных источников крайне трудоемкий и подвержен человеческим ошибкам, а использование устаревших или негибких парсеров приводит к неполным или некорректным данным из-за изменений в форматах и структуре патентной информации, а также ошибкам и разрывам соединения при парсинге из-за ограничений на частоту запросов.

Обладая актуальной и полной базой патентной информации, компании и организации могут проводить глубокий анализ, выявлять тенденции, оценивать риски и принимать взвешенные стратегические решения, опираясь на достоверные данные [3]. Это позволит им повысить конкурентоспособность, эффективность инновационной деятельности и обеспечить успешное развитие в быстро меняющихся технологических реалиях [4].

Целью данной работы является разработка системы парсинга, обеспечивающей эффективный и отказоустойчивый сбор информации патентных систем и сайтов физико-технических журналов для их последующей обработки и анализа.

Анализ предметной области

Патенты являются важным элементом интеллектуальной собственности, так как они правомерно защищают нововведения и технологические достижения [5]. Кроме того, патентная информация используется для анализа технологических трендов, оценки конкурентной среды и разработки новых продуктов, а с помощью парсинга – процесса извлечения данных из различных источников – данная информация собирается. Из-за необходимости обработки больших объемов данных, хранящихся в базах данных патентных систем, которые постоянно обновляются и дополняются новой информацией, эффективность и надежность инструментов парсинга становятся критически важной и особенно актуальной.

Однако процесс сбора патентных данных из различных источников сопряжен со значительными сложностями: патентная информация распределена по множеству баз данных и систем, каждый источник имеет свои особенности, структуры данных,

форматы и интерфейсы доступа [6]. А ручной поиск и извлечение необходимой информации из этих разрозненных систем требует огромных временных и трудовых затрат [7].

Анализ существующих решений для парсинга

В процессе выполнения работы был проведен анализ существующих решений для парсинга [8]. Рассмотренные решения предлагают различные возможности для парсинга веб-страниц [9], однако большинство из них имеют ограниченную функциональность и производительность, что важно для эффективного сбора информации патентных систем и сайтов физико-технических журналов. В Таблице 1 представлены результаты проведенного сравнительного анализа.

Наличие собственного инструмента парсинга для последующего создания собственной локальной базы данных, интегрирующей данные из патентных систем и научных журналов, имеет ряд преимуществ. Во-первых, это обеспечивает независимость от сторонних ресурсов и гарантирует бесперебойный доступ к информации. Во-вторых, создается возможность для дальнейшей обработки, анализа и сопоставления данных из разных источников, что открывает новые перспективы для выявления закономерностей, трендов и принятия обоснованных решений. Кроме того, наличие локальной базы данных повышает безопасность и конфиденциальность хранения информации, что особенно важно в условиях ограничений и санкций.

Таблица 1 – Результаты сравнения существующих решений
Table 1 – Results of comparing existing solutions

Критерий\Система	ParseHub	Chrome Scraper	Zyte	Dexi.io	Разрабаты-ваемая система
Производительность	Низкая	Средняя	Высокая	Высокая	Низкая
Масштабируемость	Низкая	Низкая	Высокая	Низкая	Низкая
Обход защиты	Нет	Нет	Есть	Нет	Нет
Стоимость	Бесплатно	Бесплатно	Высокая	Высокая	Бесплатно
Специализация	Универсаль-ный	Универсаль-ный	Универсаль-ный	Универсаль-ный	Только патенты
Доступность в России	Доступен	Доступен	Недоступен	Недоступен	Доступен

Реализация базы данных

В рамках разработки системы парсинга информации патентных систем была спроектирована и реализована гибридная архитектура базы данных (БД), состоящая из двух компонентов: ClickHouse и PostgreSQL.

ClickHouse используется в качестве основного хранилища для данных, полученных в результате парсинга с различных сайтов. Благодаря своей колоночной архитектуре и высокой производительности при выполнении аналитических запросов, ClickHouse идеально подходит для хранения и обработки больших объемов структурированных данных.

PostgreSQL используется для управления процессом парсинга и обработки данных. Ее основная задача – координировать распределение задач парсинга между различными репликами и обеспечивать согласованность данных. Ключевой особенностью PostgreSQL является наличие механизма пессимистичной блокировки (pessimistic locking). Данный механизм позволяет эффективно масштабировать процесс

парсинга как горизонтально (может быть запущено во множестве реплик), так и вертикально (параллельная обработка нескольких задач в каждой из реплик). Когда реплика получает задачу на парсинг, она выполняет запрос к соответствующей таблице в PostgreSQL. Это гарантирует, что другие реплики не смогут одновременно получить ту же самую задачу, предотвращая дубликаты и конфликты данных.

Таблица QueueElement представляет собой очередь для хранения элементов, которые должны быть обработаны. Каждый элемент в очереди имеет уникальный идентификатор (url), тип элемента (QueueElementTypeEnum), который может быть патентом Google, Яндекс-патентом, русскоязычной или англоязычной научной статьей. Также для каждого элемента хранится информация о времени начала обработки (startedAt), количестве попыток обработки (tries) и приоритете обработки (priority). Служебные поля (createdAt и updatedAt) используются для отслеживания времени создания и последнего обновления элемента в очереди.

Структура БД представлена на Рисунке 1.

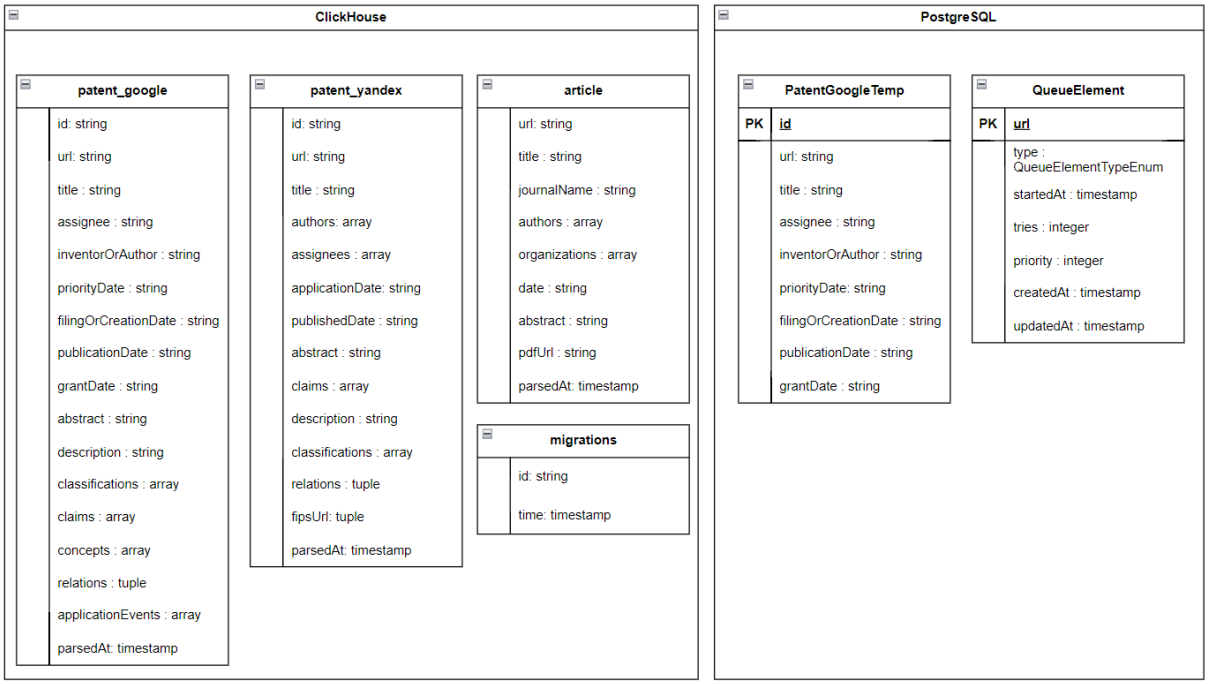

Рисунок 1 – Структура БД
Figure 1 – Database structure

Алгоритм обработки очереди

Для достижения поставленной цели была выполнена задача по разработке алгоритма обработки очереди. На Рисунке 2 представлены этапы данного алгоритма, выполняющегося автоматически сразу после инициализации и до момента остановки.

Алгоритм парсинга Google-патентов

Перед началом парсинга [10] патентов необходимо добавить их в очередь. На Рисунке 3 представлен алгоритм добавления Google-патентов в очередь. На вход алгоритма подаются параметры поиска патентов (ключевые слова, временной диапазон и т. д.). Далее осуществляется разбиение исходного промежутка дат на более мелкие для минимизации количества отказов при поиске, так как при нескольких запросах с анонимных IP-адресов Google резко повышается количество отказов. Для каждого

промежутка формируется массив ссылок (URL) на результаты поиска Google Patents по заданным параметрам. Производится цикл по полученным ссылкам на результаты поиска с постраничным переходом. Все объекты результатов поиска со страниц объединяются в один общий массив. Из общего массива отфильтровываются патенты, которые уже присутствуют в локальной базе данных, чтобы избежать дублирования. Оставшиеся новые ссылки на патенты добавляются в таблицу очереди в базе данных. Для добавленных ссылок создаются записи в таблице промежуточной информации о патентах с базовыми данными (номер, название и т. п.).

Алгоритм парсинга Google-патентов представлен на Рисунке 4.

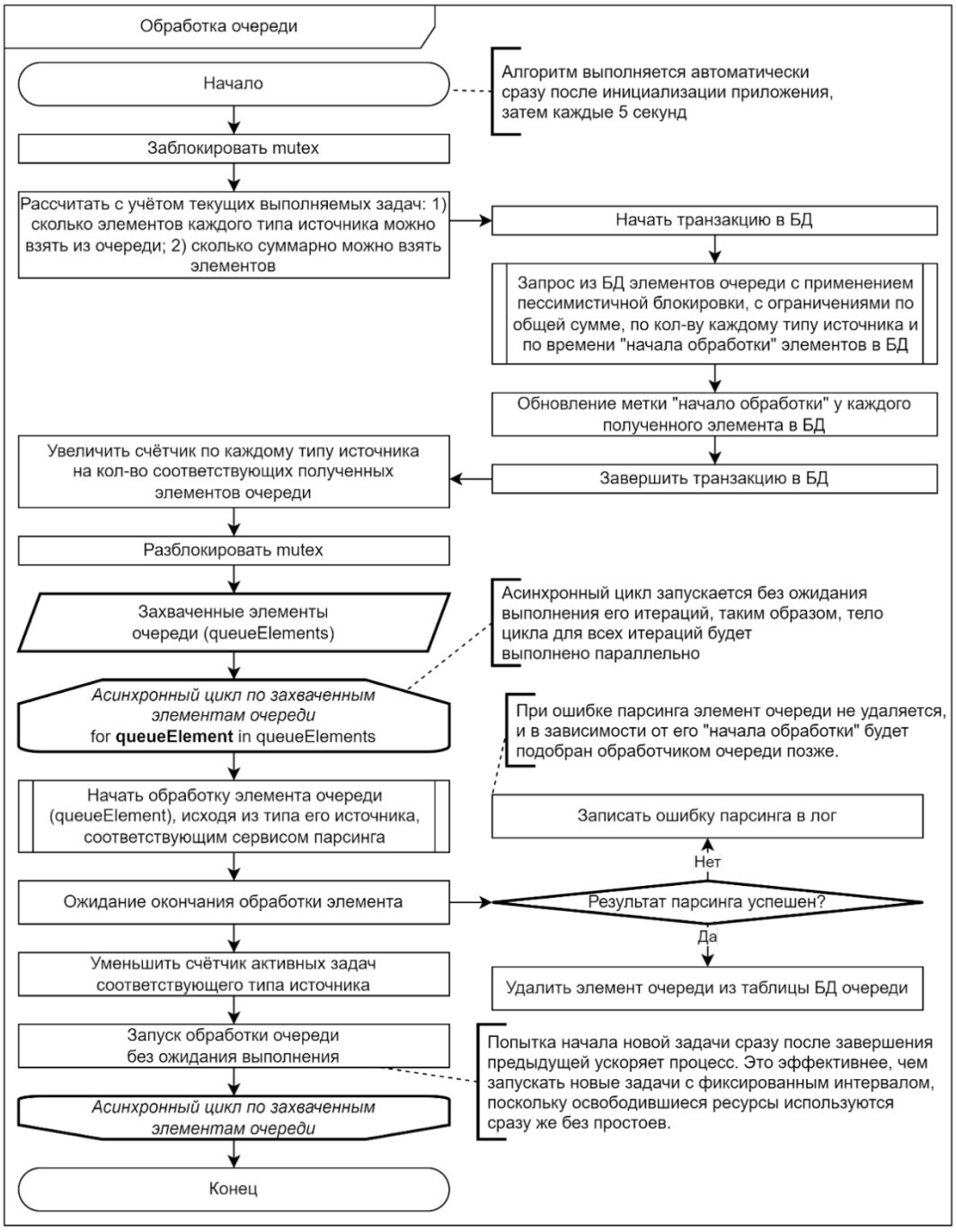

Рисунок 2 – Алгоритм обработки очереди
Figure 2 – Queue processing algorithm

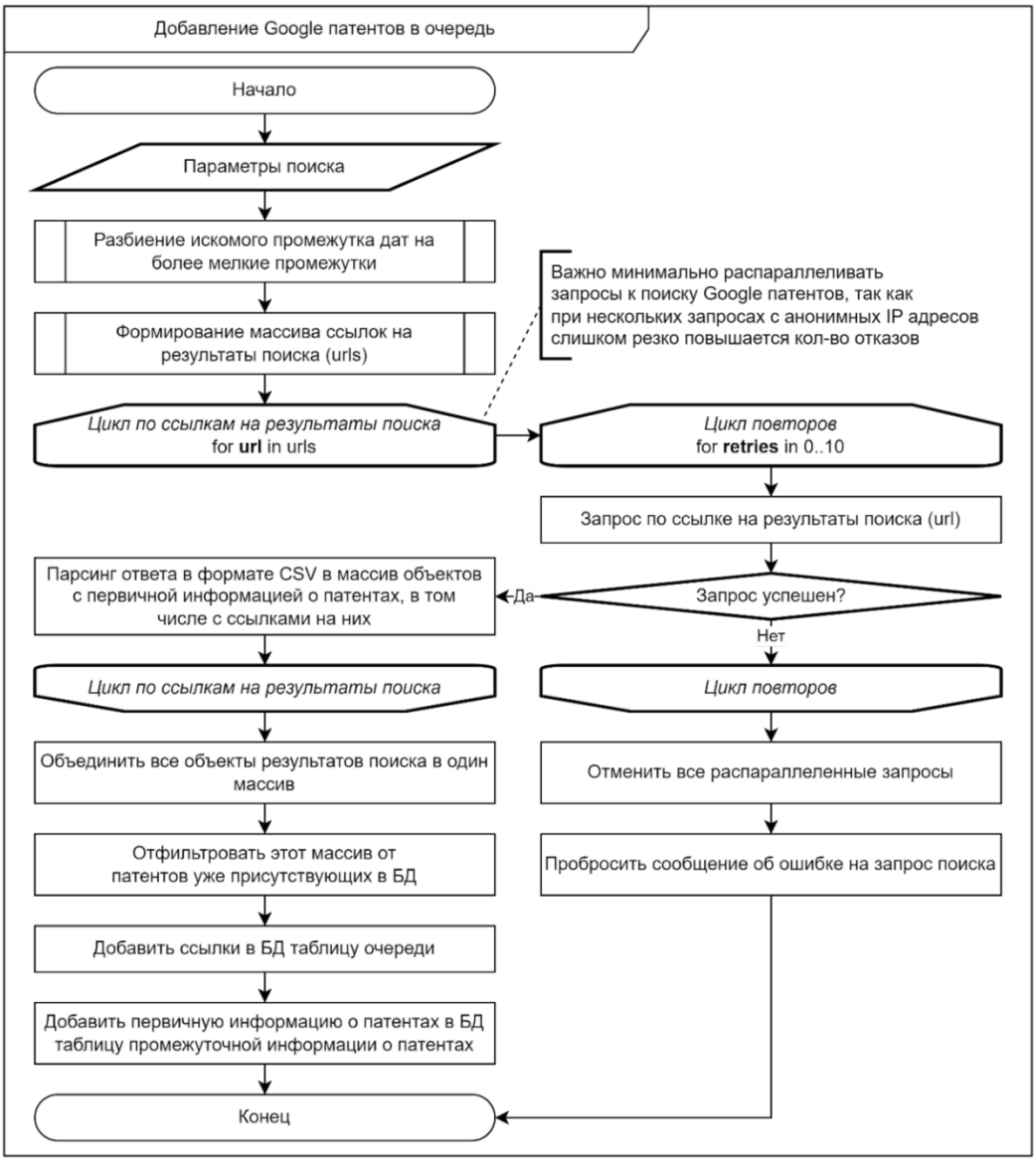

Рисунок 3 – Алгоритм добавления Google-патентов в очередь
Figure 3 – Algorithm for adding Google patents to the queue

Алгоритм парсинга Яндекс-патентов

Алгоритм добавления Яндекс-патентов в очередь представляет собой следующий процесс, представленный на Рисунке 5. Пользователь вводит параметры поиска (ключевые слова, временной интервал и т. д.) и настройки для перебора страниц результатов поиска. На основе введенных данных формируется URL-запрос для поиска патентов в Яндекс.Патенты. Для всех запросов к Яндекс-патентам необходимо использовать виртуальный headless-браузер. По сформированному URL выполняются HTTP-запросы и получаются HTML-страницы с результатами поиска. Из полученных HTML-страниц извлекаются ссылки на патенты, а также определяется общее количество найденных результатов. Извлеченные ссылки фильтруются: удаляются те, которые уже присутствуют в локальной базе данных патентов, оставшиеся новые ссылки на патенты добавляются в очередь задач на парсинг.

Алгоритм парсинга Яндекс-патентов представлен на Рисунке 6.

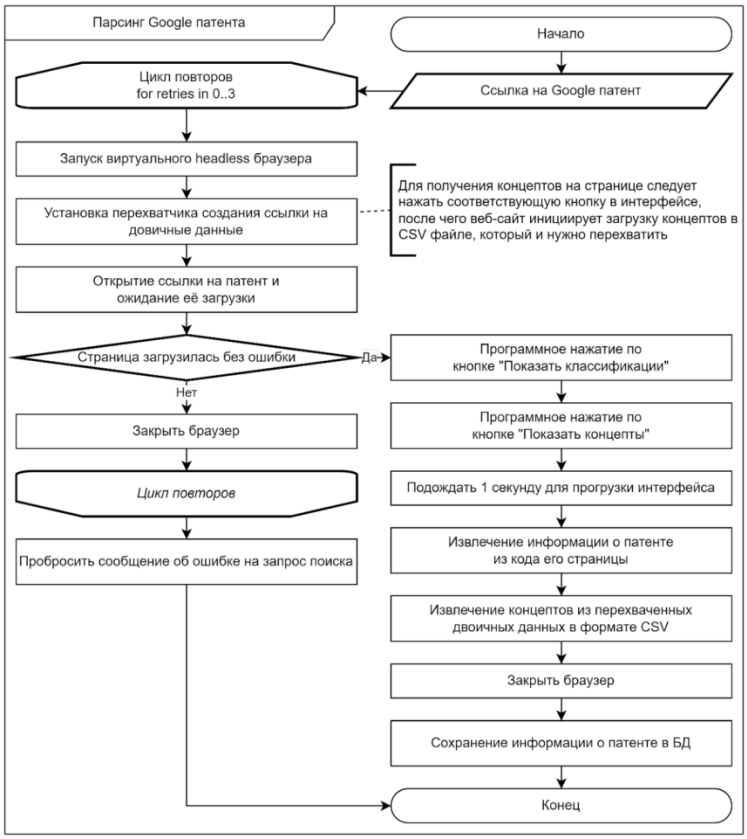

Рисунок 4 – Алгоритм парсинга Google-патентов
Figure 4 – Google Patent Parsing algorithm

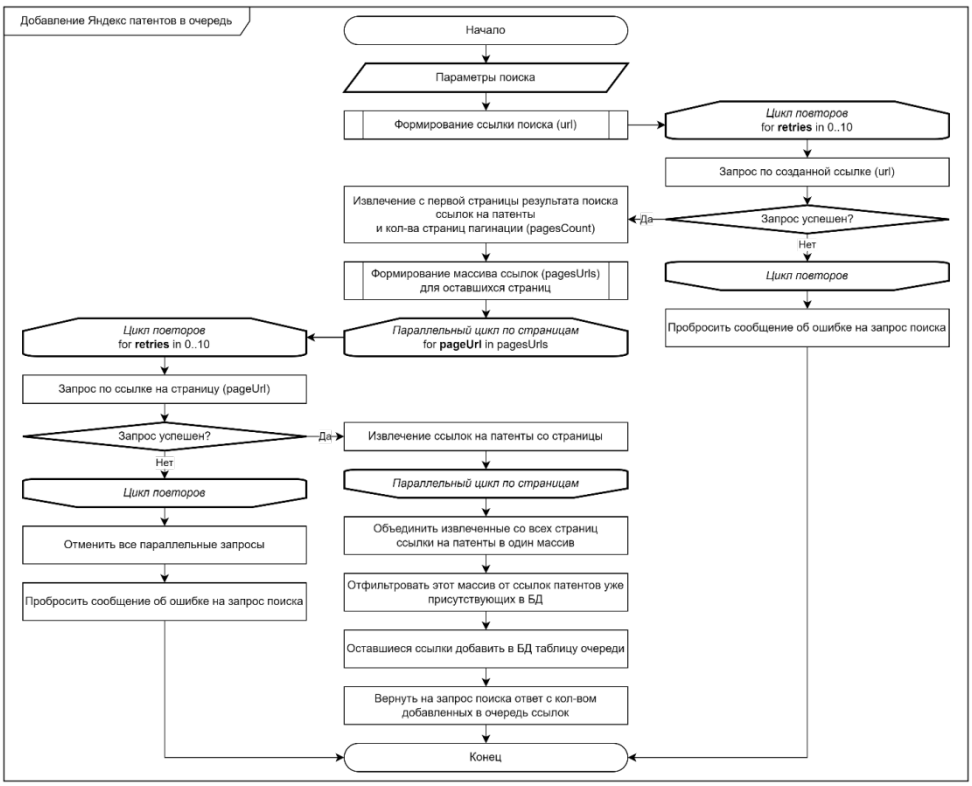

Рисунок 5 – Алгоритм добавления Яндекс-патентов в очередь
Figure 5 – Algorithm for adding Yandex patents to the queue

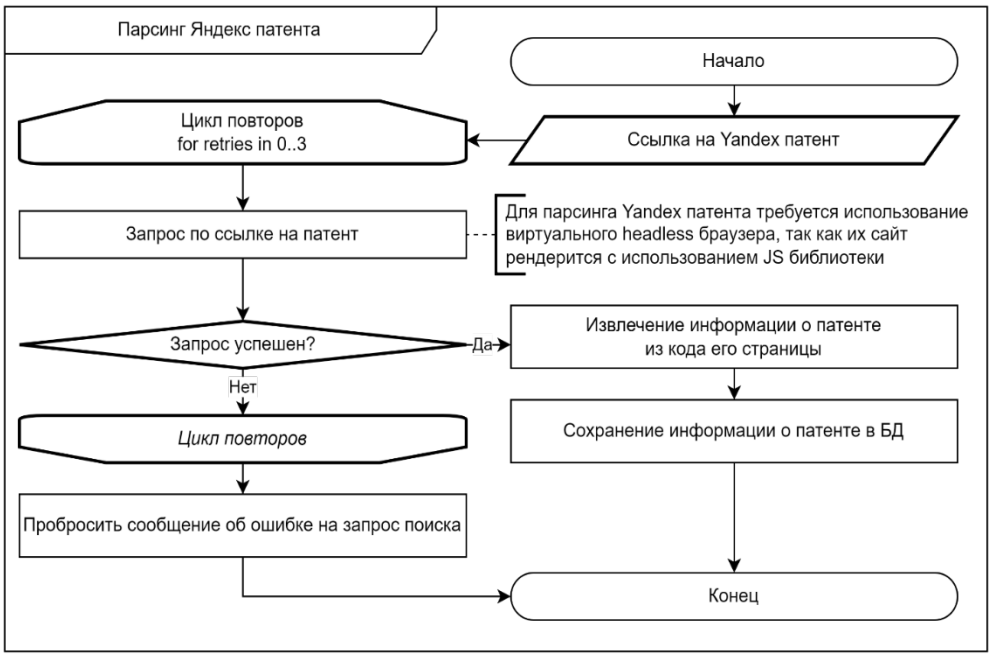

Рисунок 6 – Алгоритм парсинга Яндекс-патентов
Figure 6 – Yandex patent parsing algorithm

Алгоритм парсинга статей

В процессе выполнения работы анализировались англоязычные и русскоязычные статьи. Алгоритм добавления англоязычных статей с сайта rajpub.com в очередь представлен на Рисунке 7. Алгоритм добавления русскоязычных статей с сайта journals.ioffe.ru в очередь (Рисунок 8) аналогичен алгоритму добавления англоязычных статей в очередь и работает следующим образом: на вход подаются параметры поиска, формируется список исходных ссылок на журналы, запускается параллельный цикл по списку journalUrls: для каждого journalUrl извлекаются ссылки на выпуски журнала, после чего извлекаются ссылки на отдельные статьи выпуска, ссылки на статьи объединяются в один массив, отфильтровываются уже существующие в базе данных ссылки, остальные добавляются в таблицу очереди.

Парсинг статей (Рисунок 9) аналогичен парсингу Яндекс-патентов, за исключением того, что для парсинга статей выполняется один GET-запрос, а не запуск виртуального браузера.

Архитектура системы

На основе всех реализованных алгоритмов описывается архитектура системы, где основными компонентами является подмодуль поиска, анонимный прокси, очередь задач, обработчик очереди, подмодули парсинга и хранилище. Процесс работы системы выглядит следующим образом: пользователь инициирует процесс парсинга, задавая необходимые параметры поиска (ключевые слова, даты, источники и т. д.) и настройки. Подмодуль поиска формирует соответствующие URL-адреса для поиска на основе заданных параметров. Далее подмодуль обработки очереди запускает соответствующий модуль или модули парсинга, а они, в свою очередь, извлекают необходимые данные из полученных HTML-страниц. Извлеченная информация сохраняется в базе данных ClickHouse.

На Рисунке 10 представлена архитектура системы.

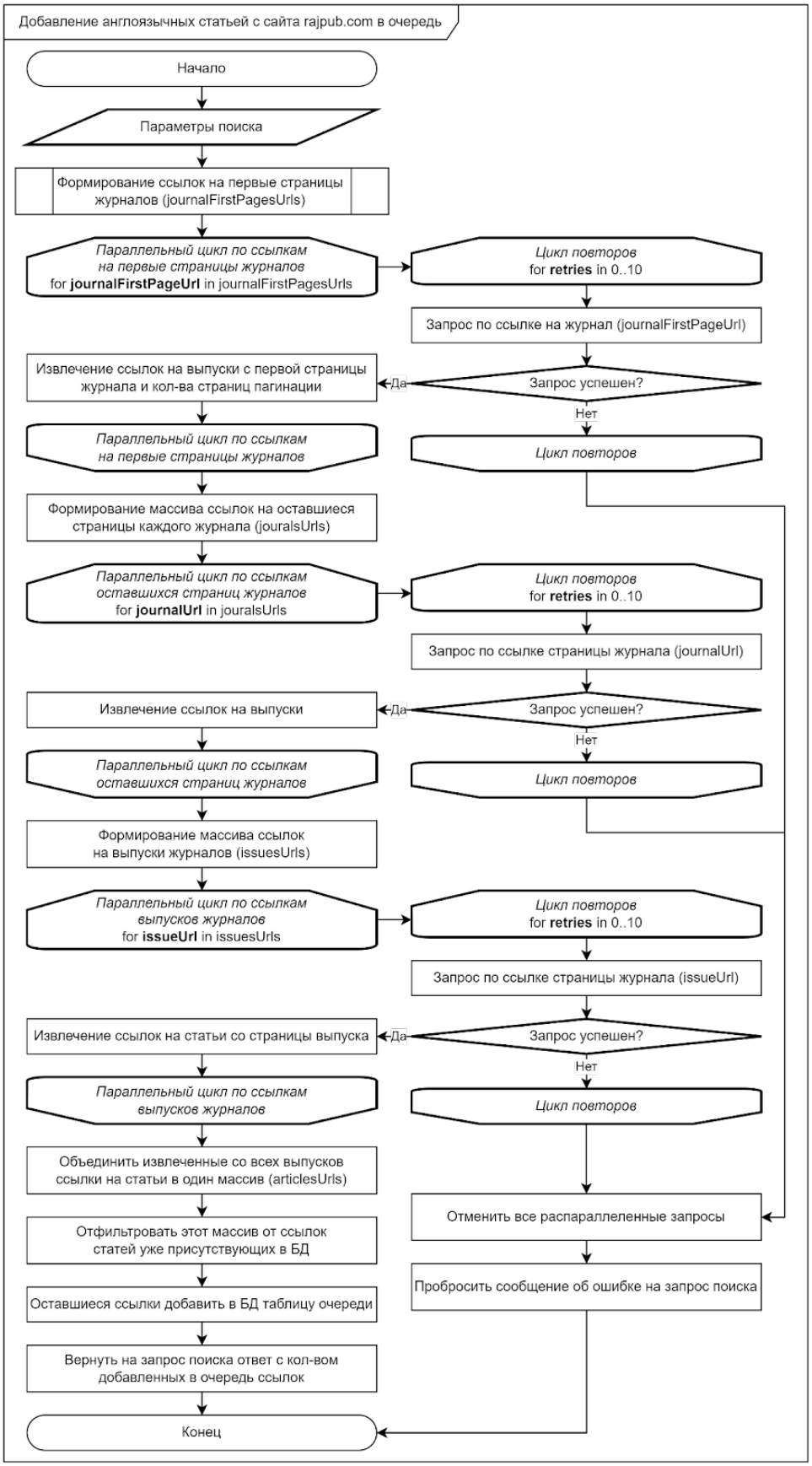

Рисунок 7 – Алгоритм добавления англоязычных статьей в очередь
Figure 7 – Algorithm for adding English-language articles to the queue

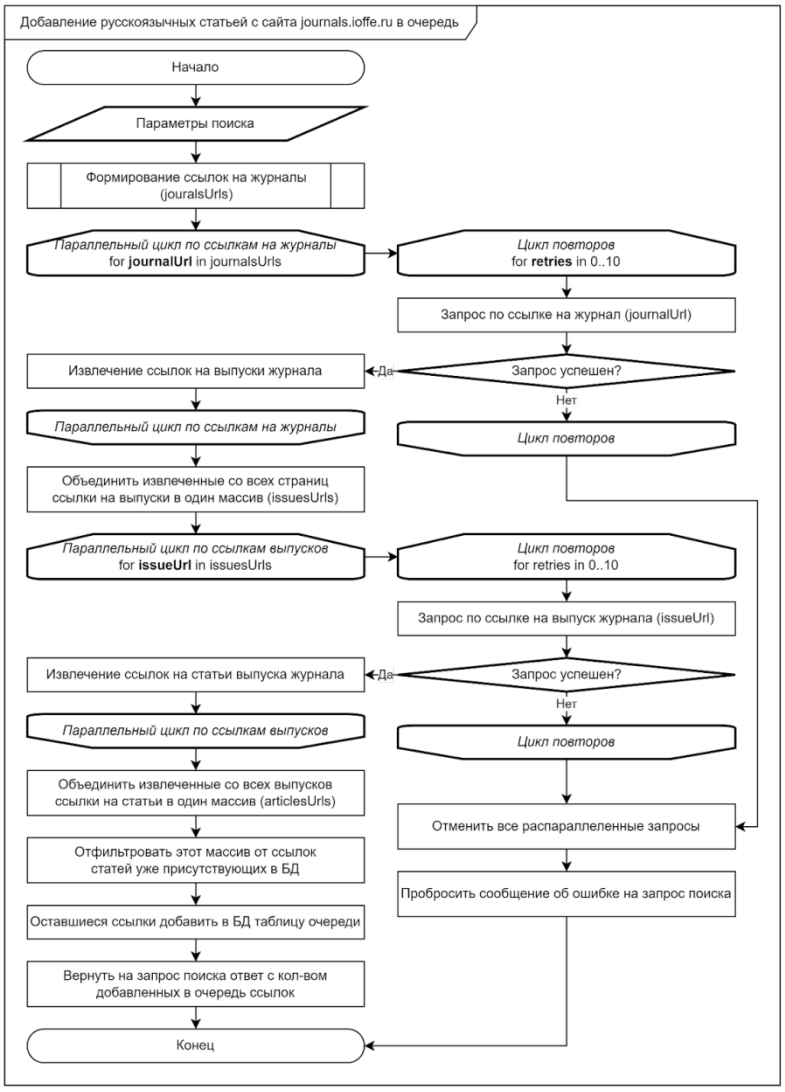

Рисунок 8 – Алгоритм добавления русскоязычных статьей в очередь
Figure 8 – Algorithm for adding Russian-language articles to the queue

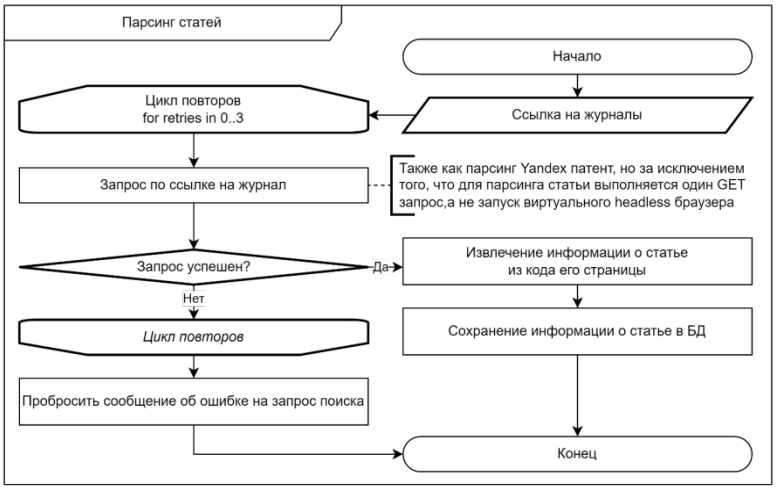

Рисунок 9 – Алгоритм парсинга статей
Figure 9 – Article parsing algorithm

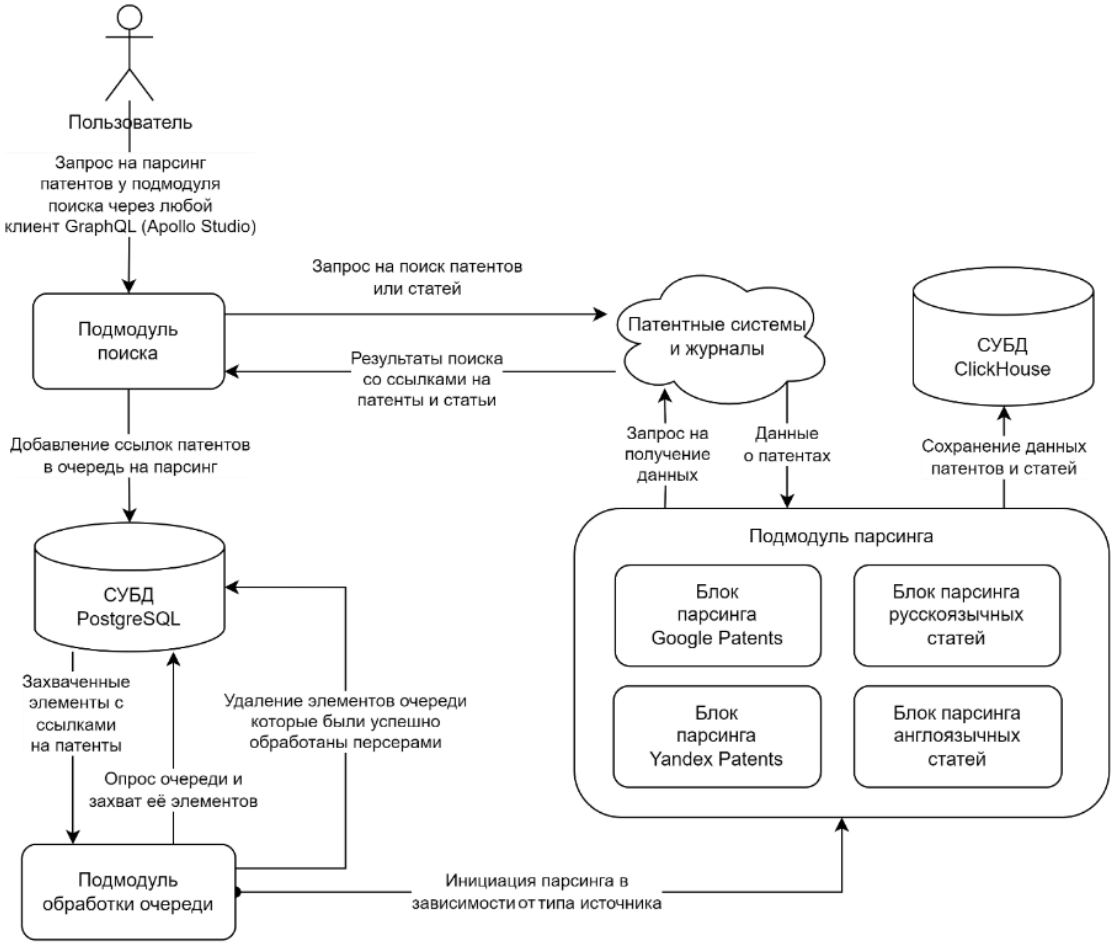

Рисунок 10 – Архитектура системы
Figure 10 – Architecture

Результаты

Для проверки работоспособности системы было проведено функциональное тестирование (Рисунки 11–14), а также тестирование отказоустойчивости.

При проведении тестирования отказоустойчивости был сделан запрос к Яндекс.Патенты на парсинг патентов с 1 по 14 января 2021 года. В результате в очередь было добавлено 533 патента. Изначально парсер был запущен параметром распараллеливания парсинга, равным 30, далее – 20, 15, 10, 5 и 1, после чего были произведены замеры времени. На Рисунке 15 отражена зависимость времени от количества параллельных задач.

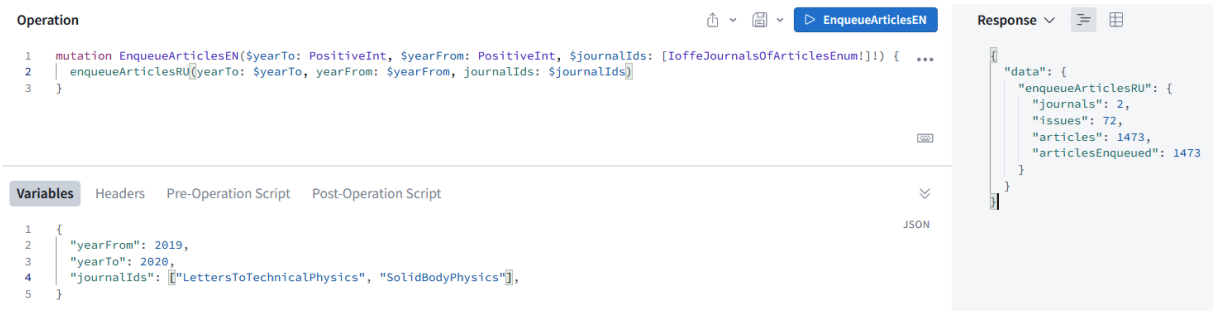

Рисунок 11 – Запрос на поиск статей
Figure 11 – Request to search for articles

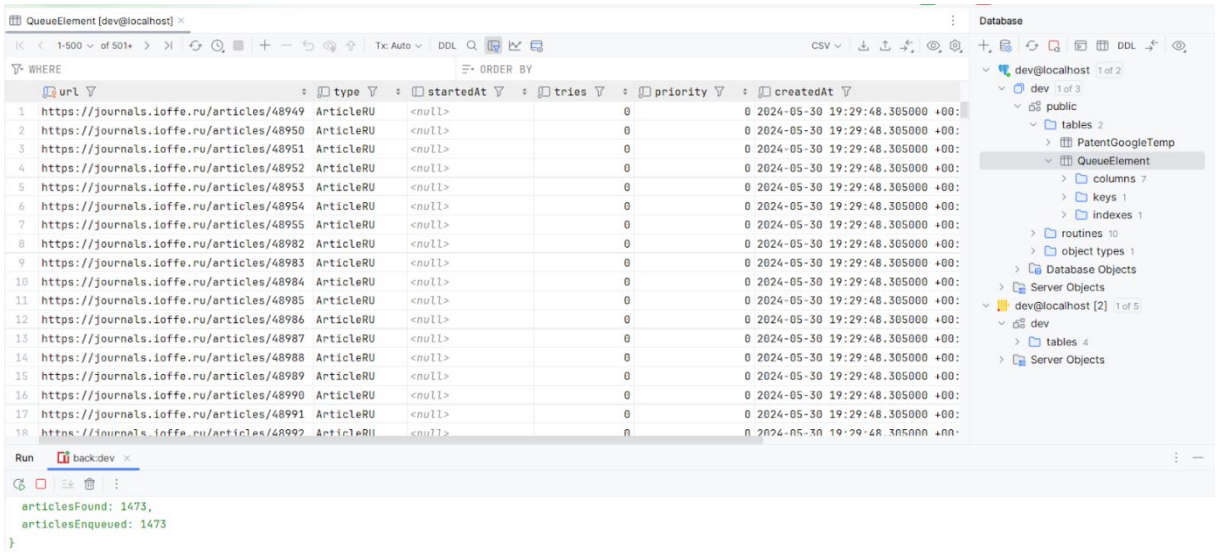

Рисунок 12 – Полученные ссылки на статьи в базе данных
Figure 12 – Received links to articles in the database

Рисунок 13 – Вывод в консоль системы парсинга
Figure 13 – Output of the parsing system results

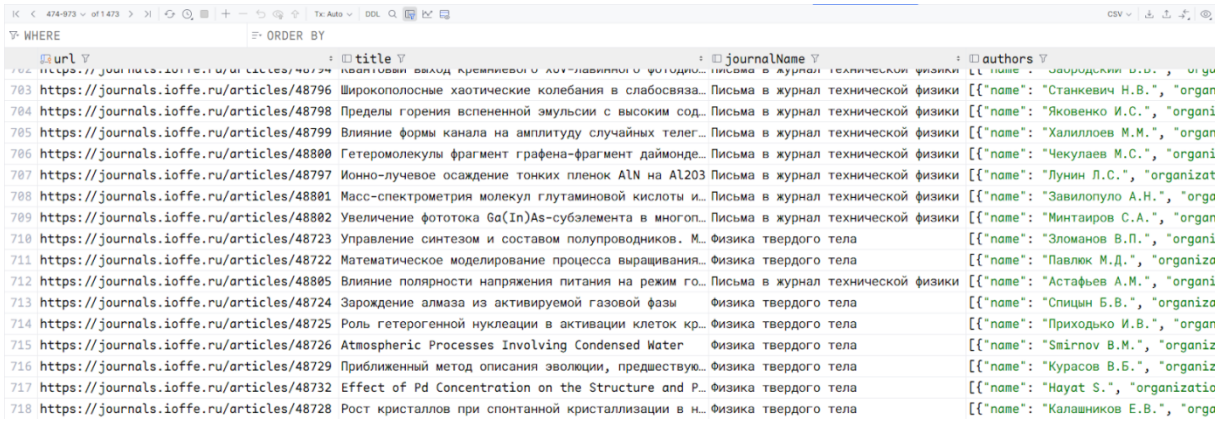

Рисунок 14 – Сохраненные записи в БД
Figure 14 – Saved records in the database

Рисунок 15 – Зависимость времени парсинга одного патента от времени
Figure 15 – Dependence of the parsing time of one patent on the time

Ключевые тренды графика:

− при небольшом количестве потоков (1–5), время загрузки значительно увеличивается. Это свидетельствует о неполной загрузке доступных ресурсов, основное время тратится на ожидание ответа от сервера Яндекс.Патенты, а вычислительные мощности и пропускная способность сети используются неэффективно;

− при увеличении количества потоков (10–30), время загрузки уменьшается, но незначительно. Это указывает на достижение точки насыщения, где производительность ограничивается внешними факторами, такими как пропускная способность сети, вычислительная способность процессора или недостаток оперативной памяти;

− оптимальное количество потоков для минимального времени загрузки находится в диапазоне от 10 до 20 потоков. В этом диапазоне достигается баланс между использованием ресурсов и ограничениями системы. Система эффективно использует доступные вычислительные мощности и пропускную способность сети, при этом не перегружая сервер Яндекс.Патенты.

В ходе тестирования не было зафиксировано ни одного случая потери соединения или отказа парсера. Система парсинга продемонстрировала высокую стабильность и отказоустойчивость, успешно обработав патенты без сбоев и ошибок.

Все тесты пройдены успешно.

Заключение

В результате проделанной работы была разработана концепция и архитектура хранения единой базы патентов и научных журнальных публикаций, создана система парсинга данной информации. Система позволяет обеспечивать автоматизированный сбор патентных данных и научных статей из различных источников, таких как Google Patents, Яндекс.Патенты, журнальный портал ФТИ им. А.Ф. Иоффе и Journal of advances in physics и использует современные технологии и архитектурные решения для достижения высокой производительности, отказоустойчивости и масштабируемости.

Система успешно извлекает все ключевые поля патентов и статей, сохраняя их в структурированном виде в базе данных для дальнейшего анализа. Обладая актуальной и полной базой патентной информации, компании и организации смогут принимать более взвешенные и обоснованные решения.

В целом, разработанная система парсинга информации патентных систем представляет собой эффективное и надежное решение для автоматизации сбора патентных данных и информации научных статей. Его применение открывает новые возможности для интеллектуального анализа данных и принятия стратегических решений в области инновационного развития.

СПИСОК ИСТОЧНИКОВ / REFERENCES

1. Загребельный М.С. Интеллектуальная собственность как ключевой ресурс в цифровой экономике. *Вестник науки*. 2024;1(6):502–511.
   Zagrebelny M.S. Intellectual property as key resource in the digital economy. *Vestnik nauki*. 2024;1(6):502–511. (In Russ.).
2. Горбашко Е.А., Карлик А.Е., Шепелев Р.Е. Патентная аналитика как элемент стратегического управления хозяйствующими структурами. *Известия Санкт-Петербургского государственного экономического университета.* 2023;(3–1):114–121.
   Gorbashko E.A., Karlik A.E., Shepelev R.E. Patent analytics as an element of strategic management of economic structures. *Izvestiya Sankt-Peterburgskogo gosudarstvennogo ekonomicheskogo universiteta.* 2023;(3–1):114–121. (In Russ.).
3. Николаев А.С. *Патентная аналитика.* Санкт-Петербург: Университет ИТМО; 2022. 98 с.
4. Никитенко С.М., Месяц М.А., Королев М.К. Патентная аналитика как инструмент формирования инновационных секторов экономики. *Экономика и управление инновациями*. 2022;(1):86–95. https://doi.org/10.26730/2587-5574-2022-1-86-95
   Nikitenko S.M., Mesyats M.A., Korolev M.K. Patent analytics as a tool of formation innovative sectors of the economy. *Economics and Innovation Management*. 2022;(1):86–95. (In Russ.). https://doi.org/10.26730/2587-5574-2022-1-86-95
5. Федорцова А.С. Объекты интеллектуальной собственности. *Российский экономический вестник*. 2021;4(2):287–290.
   Fedortsova A.S. Intellectual property objects. *Russian Economic Bulletin*. 2021;4(2):287–290. (In Russ.).
6. Мазаник А.А. Цели и основные методики патентно-информационного поиска в электронных базах данных. В сборнике: *Интеллектуальная собственность в современном мире: вызовы времени и перспективы развития: Материалы Международной научно-практической конференции: Часть 2, 20 октября 2021 года, Минск, Беларусь*. Минск: Альфа-книга; 2021. С. 7–13.
   Mazanik A.A. Goals and main methods of patent-information search in electronic databases. In: *Intellektual'naya sobstvennost' v sovremennom mire: vyzovy vremeni i perspektivy razvitiya: Materialy Mezhdunarodnoi nauchno-prakticheskoi konferentsii: Chast' 2, 20 October 2021, Minsk, Belarus.* Minsk: Al'fa-kniga; 2021. pp. 7–13. (In Russ.).
7. Меньшиков Я.С. Преимущества автоматического сбора данных в сети интернет над ручным сбором данных. *Universum: технические науки*. 2022;10(103). URL: https://7universum.com/ru/tech/archive/item/14383
   Menshikov Ya.S. Advantages of automatic data collection in the Internet over manual data collection. *Universum: tekhnicheskie nauki*. 2022;10(103). (In Russ.). URL: https://7universum.com/ru/tech/archive/item/14383
8. Козина С.А., Коробкин Д.М., Фоменков С.А. Система формирования единой базы данных по физической тематике. *Математические методы в технологиях и технике*. 2021;(8):89–92. https://doi.org/10.52348/2712-8873_MMTT_2021_8_89

Kozina S.A., Korobkin D.M., Fomenkov S.A. Formation of a unified database on physical subjects. *Mathematical Methods in Technologies and Technics*. 2021;(8):89–92. (In Russ.). https://doi.org/10.52348/2712-8873_MMTT_2021_8_89

9. Genin B.L., Zolkin D.S. Similarity search in patents databases. The evaluations of the search quality. *World Patent Information*. 2021;64. https://doi.org/10.1016/j.wpi.2021.102022
10. Feng Z. *Formal Analysis for Natural Language Processing: A Handbook*. Singapore: Springer; 2023. 796 p. https://doi.org/10.1007/978-981-16-5172-4

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Козина Светлана Александровна, инженер-исследователь, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: ksvetlan54@gmail.com
ORCID: 0000-0003-4049-620X

Svetlana A. Kozina, Research Engineer, Volgograd State Technical University, Volgograd, the Russian Federation.

Кулинченко Инна Александровна, магистрант, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: sallivan.klen@yandex.ru

Inna A. Kulinchenko, bachelor, Volgograd State Technical University, Volgograd, the Russian Federation.

Коробкин Дмитрий Михайлович, кандидат технических наук, доцент, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: dkorobkin80@mail.ru
ORCID: 0000-0002-4684-1011

Dmitry M. Korobkin, Candidate of Technical Sciences, Associate Professor, Volgograd State Technical University, Volgograd, the Russian Federation.

Фоменков Сергей Алексеевич, доктор технических наук, профессор, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: saf550@yandex.ru
ORCID: 0000-0001-9907-4488

Sergey A. Fomenkov, Doctor of Technical Sciences, Professor, Volgograd State Technical University, Volgograd, the Russian Federation.

Статья поступила в редакцию 13.11.2024; одобрена после рецензирования 25.11.2024; принята к публикации 27.11.2024.

The article was submitted 13.11.2024; approved after reviewing 25.11.2024; accepted for publication 27.11.2024.