УДК 004.853
DOI: 10.26102/2310-6018/2023.42.3.023

Разработка системы информационного поиска для сопоставления с уровнем техники

А.В. Бобунов, Д.М. Коробкин✉, С.А. Фоменков

Волгоградский государственный технический университет, Волгоград, Российская Федерация

Резюме. Актуальность данного исследования заключается в повышении эффективности извлечения ключевых фраз и слов из русскоязычного патентного массива. В настоящее время эксперты патентного ведомства вынуждены проводить анализ текста патентных заявок вручную, чтобы определить ключевые фразы и слова, которые затем используются для поиска патентов-аналогов. Этот процесс требует значительных временных затрат и может быть подвержен ошибкам. Другая проблема заключается в отсутствии системы, аналогичной Google Patents, но для русскоязычных патентов. В настоящее время не существует надежного и эффективного инструмента для автоматического определения ключевых патентных фраз и слов в русскоязычных патентах. Это ограничивает возможности экспертов при поиске и анализе патентных аналогов, а также при принятии решений о патентовании. Повышение эффективности извлечения ключевых фраз и слов из русскоязычного патентного массива имеет большое практическое значение. Это позволит сократить временные затраты на анализ патентных заявок, повысить точность поиска патентов-аналогов и обеспечить более надежные решения о патентовании. Такой инструмент будет полезен для патентных ведомств, юридических консультантов, инженеров и исследователей, которые работают с русскоязычными патентами. В целом данное исследование обусловлено необходимостью совершенствования и автоматизации процесса анализа патентных заявок, что приведет к повышению эффективности и точности работы с русскоязычным патентным массивом и сделает его более доступным и удобным для пользователей.

Ключевые слова: патенты, патентный поиск, извлечение ключевых фраз и слов, полнотекстовый поиск, HDFS, Apache Solr, Django, keyT5.

Благодарности: исследование выполнено за счет гранта Российского научного фонда № 23-21-00464, https://rscf.ru/project/23-21-00464/.

Для цитирования: Бобунов А.В., Коробкин Д.М., Фоменков С.А. Разработка системы информационного поиска для сопоставления с уровнем техники. *Моделирование, оптимизация и информационные технологии.* 2023;11(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1413 DOI: 10.26102/2310-6018/2023.42.3.023

The development of the information retrieval system for state of art assessment

A.V. Bobunov, D.M. Korobkin✉, S.A. Fomenkov

Volgograd State Technical University, Volgograd, the Russian Federation

Abstract. The relevance of this study is due to the need to improve the efficiency of extracting key phrases and words from the Russian-language patent array. Currently, patent office experts have to analyze texts of patent applications manually in order to identify key phrases and words that are then used to search for patent counterparts. This process is time-consuming and can be error-prone. Another problem is the lack of a system similar to Google Patents but for Russian-language patents. Currently, there is no reliable and effective tool for automatic identification of key patent phrases and words in Russian-language patents. This limits the ability of experts to search and analyze patent analogues as

well as to make decisions on patenting. Improving the efficiency of extracting key phrases and words from the Russian-language patent array is of great practical importance. This will reduce the time spent on the analysis of patent applications, improve the accuracy of the search for similar patents and provide more reliable patenting solutions. Such a tool will be useful for patent offices, legal consultants, engineers and researchers who work with Russian-language patents. In general, this study is conditioned by the need to improve and automate the process of analyzing patent applications, which will lead to an increase in the efficiency and accuracy of managing the Russian-language patent array and make it more accessible and user-friendly.

Keywords: patents, patent search, keyword extraction, full-text search, HDFS, Apache Solr, Django, keyT5.

Acknowledgements: the research was funded by a grant from Russian Science Foundation (No. 23-21-00464, https://rscf.ru/project/23-21-00464/).

For citation: Bobunov A.V., Korobkin D.M., Fomenkov S.A. The development of the information retrieval system for state of art assessment. *Modeling, Optimization and Information Technology*. 2023;11(3). URL: https://moitvivt.ru/ru/journal/pdf?id=1413 DOI: 10.26102/2310-6018/2023.42.3.023 (In Russ.).

Введение

В современном информационном обществе каждый день создается и патентуется огромное количество новых технических решений и изобретений. Патентные данные содержат ценную информацию о новых технологиях, разработках и научных достижениях, которые могут быть использованы для определения уровня техники в определенной области. Однако с ростом объема патентных данных становится все сложнее и труднее осуществлять поиск и анализ этой информации вручную.

Для эффективного изучения патентных данных и определения уровня техники необходимы специализированные системы информационного поиска. Разработка такой системы имеет актуальность в свете возрастающего интереса к инновациям и необходимости обеспечения доступа к актуальным исследованиям в различных областях науки и техники.

Цели и задачи данного исследования связаны с повышением эффективности информационного поиска для определения уровня техники в патентных данных. Объектом исследования являются патентные данные, содержащие информацию о новых технологиях, разработках и научных достижениях. Предметом исследования выступает система информационного поиска, которая позволит осуществлять эффективный поиск, анализ и сравнение патентной информации с целью определения текущего уровня техники и выявления трендов в различных областях науки и техники. Критериями повышения эффективности данного исследования являются полнота и точность, а также скорость извлечения ключевой информации из текста.

Такая система может предоставлять возможность осуществлять поиск патентных данных по ключевым словам, классификациям, авторам и другим параметрам, а также производить анализ и сравнение технических решений в различных патентах. Она может помочь исследователям, инженерам, юристам и другим специалистам в определении текущего уровня техники, выявлении трендов и разработке новых инноваций.

Анализ предметной области

Патент – это правовой документ, который выдается правительством для защиты изобретения или инновации. Это документ, который дает его обладателю исключительное право на использование и продажу изобретения в течение определенного периода времени, обычно составляющего от 10 до 20 лет. В обмен на это

право обладатель патента обязуется предоставлять обществу полную информацию о своем изобретении.

Патент выдается для защиты новых идей, процессов, устройств, композиций материалов, машин и многого другого. Чтобы получить патент, необходимо подать заявку с описанием изобретения, его применения и преимуществ. Заявитель также должен представить доказательства, что его изобретение является новым и не является очевидным для эксперта в соответствующей области знаний.

Один из ключевых аспектов патента – это его исключительность. Обладатель патента имеет право запретить другим лицам применение, производство, продажу или импорт продукта, которые используют его изобретение без его разрешения. Это обеспечивает обладателя патента защитой от конкурентов и дает ему возможность получить экономические выгоды от своего изобретения.

Патенты играют важную роль в инновационной индустрии и науке, поскольку они способствуют стимулированию исследований и разработок новых технологий и продуктов. Патенты также являются важным инструментом защиты интеллектуальной собственности и могут помочь компаниям обезопасить свои продукты и технологии от копирования или незаконного использования другими организациями [1].

Патентный поиск – это процесс поиска и анализа патентной информации с целью выявления новых технических решений, которые уже были описаны в существующих патентах. Он имеет несколько целей и применений.

Во-первых, патентный поиск позволяет проверить оригинальность и новизну изобретения перед его подачей на патентование. Это позволяет убедиться в том, что предлагаемое изобретение действительно новое и не нарушает патентные права других компаний или изобретателей.

Во-вторых, патентный поиск помогает компаниям и изобретателям ориентироваться в существующей технической области и выявлять сильные и слабые стороны своих конкурентов. Это позволяет улучшать свои продукты и технологии и быть на шаг впереди конкурентов.

В-третьих, патентный поиск может использоваться для выявления новых направлений и тенденций в технической области. Изучение патентов позволяет выявлять существующие и новые технологические решения, их преимущества и недостатки, что может привести к разработке новых продуктов и технологий.

Кроме того, патентный поиск полезен при судебных разбирательствах и защите патентных прав. Если компания уверена в том, что ее права были нарушены другими компаниями или изобретателями, патентный поиск может помочь найти доказательства и убедительные аргументы для суда.

Патентный поиск является важным инструментом для инновационных компаний и изобретателей, которые хотят создавать новые продукты и технологии и защищать свои патентные права. Благодаря патентному поиску компании могут убедиться в том, что их изобретения действительно новые и оригинальные, а также получить ценную информацию о существующей технической области и конкурентах [2].

Поиск уровня техники в патентном массиве – это процесс поиска патентных документов, которые описывают технологические решения, разработанные ранее. Такой поиск обычно проводится с целью оценки уровня инноваций и существующих патентных прав на технологические решения. Для проведения поиска уровня техники в патентном массиве необходимо выполнить следующие шаги:

1. Сначала необходимо определить технологическую область, в которой требуется провести поиск патентных документов. Для этого используются ключевые слова и фразы, которые характеризуют технологию.
2. Затем осуществляется поиск в базе данных патентов, которая содержит

информацию о зарегистрированных патентах. Такая база данных может быть платной или бесплатной, но в любом случае в нее необходимо вводить ключевые слова и фразы, которые были определены на первом этапе.

3. После завершения поиска необходимо проанализировать полученные результаты. Для этого следует отсортировать патенты по дате выдачи, а также оценить соответствие найденных документов искомым технологическим решениям.

4. После анализа результатов поиска необходимо определить уровень техники, который описывает найденные патенты. Для этого используются критерии, такие как инновационность, технологическая сложность и степень развития данной области технологий.

Поиск уровня техники в патентном массиве несет в себе сложный и трудоемкий процесс, который требует знаний в области технологий и патентного законодательства, а также использования специальных баз данных и инструментов для поиска и анализа патентов.

Анализ существующих решений информационного поиска

Сравнивая системы информационного поиска патентов, такие как Google Patents, «Яндекс Патенты» и «Роспатент», можно выделить следующие особенности и характеристики:

1. Количество патентов: все три системы предлагают доступ к высокому объему патентных данных, что обеспечивает широкий охват и возможности для исследования.

2. Релевантность поиска: Google Patents имеет высокую релевантность поиска, что означает, что результаты поиска обычно соответствуют заданным критериям. В то время как «Яндекс Патенты» и «Роспатент» имеют среднюю релевантность поиска, что может потребовать более точного формулирования запросов для получения наиболее соответствующих результатов.

3. Перевод документов: все три системы предоставляют возможность перевода патентных документов. Это позволяет пользователям получать информацию на разных языках и удобно работать с патентами на иностранных языках.

4. Стоимость использования: все три системы предоставляют бесплатный доступ к базам данных патентов. Это делает их доступными для широкого круга пользователей без необходимости платить за доступ к информации.

5. Доступность данных: Google Patents и «Яндекс Патенты» предоставляют открытый доступ ко всем патентным данным. Однако у «Роспатента» ограниченный доступ к некоторым данным в связи с конфиденциальностью или другими ограничениями.

6. Качество данных: все три системы обладают высоким качеством данных, что обеспечивает достоверность и точность информации, содержащейся в патентных документах.

7. Удобство использования: Google Patents и «Яндекс Патенты» предлагают высокое удобство использования, предоставляя интуитивно понятные пользовательские интерфейсы и удобные функции поиска и фильтрации. «Роспатент» имеет средний уровень удобства использования, требуется некоторое привыкание к интерфейсу системы.

Таблица 1 – Существующие системы информационного поиска
Table 1 – Existing information retrieval systems

Характеристики	Google Patents	«Яндекс Патенты»	«Роспатент»
Русские патенты	+	+	+
Англоязычные патенты	+	+	–
Извлечение ключевых слов и фраз	+ (патент на русском языке, ключевые слова и фразы на английском языке)	–	–

В Таблице 1 приведен сравнительный анализ систем по извлечению ключевых фраз и слов. В этом аспекте Google Patents предоставляет возможность извлечения ключевых слов и фраз как для русских, так и для англоязычных патентов. «Яндекс Патенты» и «Роспатент» такую функциональность ограничивают: они не предоставляет возможности извлечения ключевых слов и фраз вообще.

В результате проведенного анализа можно сделать вывод, что функционал извлечения ключевых фраз и слов отсутствует в сервисах «Яндекс Патенты» и «Роспатент», что может негативно сказываться на эффективности поиска уровня техники в данных системах.

Анализ существующих моделей для извлечения ключевых фраз и слов

Извлечение ключевых фраз и слов – это процесс автоматического выделения наиболее значимых слов и фраз из текста, которые наиболее полно описывают его содержание. Это важная задача для многих приложений обработки естественного языка, включая автоматическую классификацию текстов, поиск похожих документов, анализ тональности и другие задачи [3].

Существует несколько подходов к извлечению ключевых фраз и слов, но наиболее популярными среди них являются методы на основе частотности и методы на основе машинного обучения. Различные модели языковых моделей, такие как KeyBERT, keyPhrasetransformer и keyT5-base/large, могут использоваться для обработки текста.

Методы на основе частотности обычно основаны на подсчете частоты встречаемости слов в тексте и выборе наиболее часто встречающихся слов или фраз. Например, можно использовать алгоритм TF-IDF, который учитывает, насколько часто слово встречается в тексте и во всем корпусе текстов, чтобы выделить наиболее важные слова. Эти методы просты в реализации и эффективны для обработки больших объемов текста, но не всегда учитывают контекст и семантические связи между словами [4].

Методы на основе машинного обучения используют модели глубокого обучения, такие как KeyBERT, keyPhrasetransformer и keyT5-base/large, чтобы изучить семантические связи между словами и определить, какие слова и фразы наиболее важны для описания содержания текста. Эти методы учитывают контекст и семантические связи между словами и могут применяться для различных языковых задач, включая извлечение ключевых фраз и слов. Кроме того, модели на основе машинного обучения могут учитывать множество факторов, такие как частотность, контекст, длина фразы и т. д., что делает их более точными в определении наиболее важных слов и фраз в тексте.

Сравнительный анализ трех моделей, KeyBERT, keyPhrasetransformer и keyT5-base/large, был проведен на базе данных текстов на русском и английском языке.

KeyBERT – это модель, основанная на архитектуре BERT, предназначенная для

извлечения ключевых фраз из текста на основе семантической схожести.

keyPhrasetransformer – это модель, основанная на архитектуре T5 Transformer, разработанная для быстрого извлечения ключевых фраз и тематик из текстовых документов.

keyT5-base/large – это модель, основанная на архитектуре T5 Transformer, обученная для обработки текста и генерации ключевых фраз, доступна в двух размерах: базовая (base) и большая (large).

Каждая из этих моделей имеет свои преимущества и недостатки. KeyBERT и keyPhrasetransformer основаны на архитектуре трансформера и учитывают контекст и семантические связи в тексте. Они работают с полнотекстовым описанием документов и позволяют извлекать информативные ключевые фразы. Однако требуется предварительное обучение на специализированных данных, и время обработки текста может быть значительным.

В Таблице 2 приведено сравнение моделей для извлечения ключевых фраз и слов.

Таблица 2 – Сравнение моделей извлечения ключевых фраз и слов
Table 2 – Comparison of keyword and word extraction models

Модель	Описание	Преимущества	Недостатки
KeyBERT	Модель на основе BERT, обученная для извлечения ключевых фраз из текста.	Учитывает контекст и семантические связи. Гибкость и адаптивность. Высокая точность.	Требуется предварительное обучение на специализированных данных. Время обработки текста.
keyPhrasetransformer	Модель на основе T5 Transformer, предназначенная для извлечения ключевых фраз из текста.	Работает с полнотекстовым описанием документов. Учитывает контекст и семантические связи.	Требуется обучение на специализированных данных. Время обработки текста.
keyT5-base/large	Модель на основе ruT5, обученная для обработки текста и генерации ключевых фраз.	Учитывает контекст и семантические связи. Работает с полнотекстовым описанием документов.	Высокие вычислительные требования. Необходимость дообучения для конкретных задач.

Модель keyT5-base/large также использует трансформерную архитектуру, но имеет более широкий набор предобученных моделей. Она может быть использована для обработки текста и генерации ключевых фраз. Модель также работает с полнотекстовым описанием документов, но требует больших вычислительных ресурсов. Для конкретных задач может быть необходимо дообучение.

Для решения задачи извлечения ключевых фраз и слов на русском языке больше всего подходит модель keyT5-large, основанная на архитектуре модели ruT5-large [5], которая лучше всего выполняет разные варианты задач, как показано на Рисунке 1.

В целом все три модели KeyBERT, keyPhrasetransformer и keyT5-large предназначены для извлечения ключевых фраз и слов и могут быть использованы для повышения эффективности информационного поиска в патентном массиве. Однако с

учетом особенностей данной задачи модель keyT5-large представляется наиболее подходящей, поскольку она обладает большей мощностью и гибкостью, но требует дообучения на патентных данных для достижения оптимальной производительности и точности.

Rank	Name	Team	Link	Score	LiDiRus	RCB	PARus	MuSeRC	TERRa	RUSSE	RWSD	DaNetQA	RuCoS
1	HUMAN BENCHMARK	AGI NLP		**0.811**	0.626	0.68 / 0.702	0.982	0.806 / 0.42	0.92	0.805	0.84	0.915	0.93 / 0.89
2	Golden Transformer	Avengers Ensemble		**0.679**	0.0	0.406 / 0.546	0.908	0.941 / 0.819	0.871	0.587	0.545	0.917	0.92 / 0.924
3	ruT5-large-finetune	sberdevices		**0.634**	0.32	0.306 / 0.498	0.66	0.815 / 0.537	0.747	0.735	0.669	0.711	0.81 / 0.764
4	ruRoberta-large finetune	sberdevices		**0.624**	0.339	0.357 / 0.518	0.508	0.83 / 0.561	0.801	0.715	0.571	0.82	0.73 / 0.716
5	ruT5-base-finetune	sberdevices		**0.596**	0.267	0.307 / 0.468	0.554	0.769 / 0.446	0.692	0.682	0.669	0.732	0.79 / 0.752
6	ruBert-large finetune	sberdevices		**0.583**	0.235	0.356 / 0.5	0.492	0.76 / 0.427	0.704	0.682	0.669	0.773	0.68 / 0.658
7	ruBert-base finetune	sberdevices		**0.578**	0.224	0.333 / 0.509	0.476	0.742 / 0.399	0.703	0.706	0.669	0.712	0.74 / 0.716
8	YaLM 1.0B few-shot	Yandex		**0.577**	0.124	0.408 / 0.447	0.766	0.673 / 0.364	0.605	0.587	0.669	0.637	0.86 / 0.859
9	RuGPT3XL few-shot	sberdevices		**0.535**	0.096	0.302 / 0.418	0.676	0.74 / 0.546	0.573	0.565	0.649	0.59	0.67 / 0.665
10	MT5 Large	AGI NLP		**0.528**	0.061	0.366 / 0.454	0.504	0.844 / 0.543	0.561	0.633	0.669	0.657	0.57 / 0.562
11	RuBERT plain	DeepPavlov		**0.521**	0.191	0.367 / 0.463	0.574	0.711 / 0.324	0.642	0.726	0.669	0.639	0.32 / 0.314

Рисунок 1 – Сравнение моделей
Figure 1 – Comparison of models

Разработка алгоритма парсинга патентов

Алгоритм начинается с указания пути к папке, где содержатся XML-файлы патентов. Затем создается пустой DataFrame с заданными столбцами для записи данных патентов. Происходит итерация по файлам в папке, где каждый патентный файл извлекается с помощью библиотеки BeautifulSoup. Из различных тегов XML-файла извлекаются необходимые данные, такие как номер патента, название, дата, страна, авторы, классификация, аннотация, формула изобретения и описание. Извлеченные данные записываются в виде словаря, который затем добавляется в DataFrame. Наконец, DataFrame сохраняется в файл Excel формата (xlsx) для сохранения полученных данных, как показано на Рисунке 2.

Разработка алгоритма формирования обучающей выборки

Алгоритм, как показано на Рисунке 3, состоит из следующих этапов. Сначала загружаются необходимые библиотеки, такие как multiprocessing, requests, pandas, nltk, keyphrasetransformer, time и joblib. Затем определяется функция translate_text, которая используется для перевода текста с одного языка на другой с помощью Google Translate. Она принимает текст, исходный язык и целевой язык в качестве аргументов и возвращает переведенный текст.

Далее определяется функция process_row, которая обрабатывает каждую строку патента в обучающей выборке. Внутри этой функции определяются списки столбцов, содержащих текстовые данные патента, а также списки для хранения переводов и ключевых фраз. Для каждого столбца происходит следующий процесс: извлечение текста из столбца, разделение текста на предложения с помощью nltk.sent_tokenize, формирование пакета предложений для перевода и их объединение в одну строку, перевод пакета предложений с помощью translate_text, разделение переведенного текста на предложения и добавление их в список переводов, извлечение ключевых фраз из каждого предложения с помощью kp.get_key_phrases, перевод ключевых фраз на русский язык с помощью translate_text, обновление значения соответствующего столбца

ключевых фраз патента.

В основной части происходит загрузка файла с патентными данными, считывание файла Excel с помощью pd.read_excel и сохранение его в объект DataFrame. Затем создается экземпляр KeyPhraseTransformer, используемый для извлечения ключевых фраз. Создается пул процессов с использованием Pool, который позволяет параллельно обрабатывать строки патентов. Далее итерируется по каждой строке в DataFrame, и каждая строка передается в функцию process_row с помощью Parallel и delayed. Обновленные строки сохраняются, после чего пул процессов закрывается. Обновленные значения в DataFrame заменяют исходные значения. Сохраняются только столбцы с ключевыми фразами в новый файл Excel, как показано на Рисунке 4.

A	B	C	D	E	F	G	H	I	J	K
X	Y									
Изобретение относится к области сельского хозяйства, в частности к технолог	рассада ивы', 'ивы', 'рассада', 'болотистая почва'									
Высевающий аппарат включает бункер для семян с горизонтальным высевающ	посевной', 'вращающийся сменный ячеистый диск', 'агрегат', 'посевной агрегат', 'семенной бункер'									
Изобретение относится к сельскохозяйственному машиностроению, в частно	Агротехника', 'комбайны для сбора корней'									
Изобретение относится к сельскохозяйственному машиностроению и может	сельскохозяйственная', 'техника', 'сельскохозяйственная техника', 'уборка и транспортировка урожая'									
Способ относится к области сельского хозяйства и растениеводства. Способ в	сельское хозяйство', 'растениеводство', 'метод', 'хозяйство', 'сельское'									
Изобретение относится к области технических средств, предназначенных для	облака', 'переохлажденные облака', 'переохлажденные', 'град', 'искусственные осадки'									
Способ включает обработку полос поверхности орошаемого участка, покрыти	пленка', 'синтетическая пленка', 'поливная вода', 'синтетическая'									
Способ включает обработку полос поверхности орошаемого участка, покрыти	пленка', 'синтетическая пленка', 'поливная вода', 'синтетическая'									
Способ полива включает обработку полос поверхности орошаемого участка, п	метод', 'орошения', 'орошаемая площадь', 'метод орошения'									
Способ включает обработку полос поверхности орошаемого участка, покрыти	пленка', 'синтетическая пленка', 'поливная вода', 'синтетическая'									
Способ включает обработку полос поверхности орошаемого участка, покрыти	пленка', 'синтетическая пленка', 'поливная вода', 'синтетическая'									
Изобретение относится к сельскому хозяйству, в частности к устройствам для	молочное производство', 'производство', 'молочное', 'контроль качества'									
Изобретение относится к области сельского хозяйства, в частности к устройст	корма', 'раздача корма', 'раздача'									
Изобретение относится к животноводству, в частности к птицеводству. Для сн	животноводство', 'птицеводство'									
Рыбацкое грузило-отцепляйка содержит грузило, водорастворимый материал	рыболовное грузило-разцепитель', 'грузило-разцепитель', 'водорастворимый материал', 'рыболовное'									

Рисунок 4 – Результаты формирования обучающей выборки
Figure 4 – Results of training sample creation

Обучение модели KeyT5-large

Вначале происходит импорт необходимых библиотек, таких как pandas, numpy, torch, transformers, random и tqdm. Затем данные загружаются из файла Excel в объект DataFrame. Далее происходит разделение данных на обучающую и тестовую выборки.

Модель keyT5-large и токенизатор инициализируются, и задается оптимизатор для обновления параметров модели. Далее происходит процесс обучения модели, который включает итерацию по эпохам, перемешивание пар значений данных, итерацию по пакетам данных, токенизацию и преобразование входных и выходных данных в формат PyTorch tensor, вычисление потери модели и выполнение обратного распространения градиента для обновления параметров модели. Значения потерь сохраняются для отслеживания процесса обучения.

После обучения модель проверяется на тестовой выборке. Модель переводится в режим оценки, и определяется функция для генерации ответов модели на новых данных. Затем модель применяется на обучающей и тестовой выборках, и генерируются ответы на вопросы.

В конце процесса обучения модель и токенизатор сохраняются в новую папку с названием "keyT5-large" с помощью методов model.save_pretrained() и tokenizer.save_pretrained(). Этот алгоритм предоставляет основу для обучения модели извлечения ключевых фраз и слов на основе модели keyT5-large.

Разработка алгоритма формирования ключевых фраз и слов на основе обученной модели keyT5-large

Алгоритм формирования ключевых фраз и слов на основе обученной модели keyT5-large используется для автоматического извлечения и определения важных терминов, фраз и слов из полнотекстового описания патента. Процесс включает чтение файла XLSX, извлечение значений из нужных столбцов, обработку этих значений с

помощью модели keyT5-large для извлечения ключевых слов и фраз, запись результатов обратно в файл XLSX и конвертацию файла XLSX в формат JSON, как показано на Рисунке 5. Затем этот файл JSON загружается в HDFS.

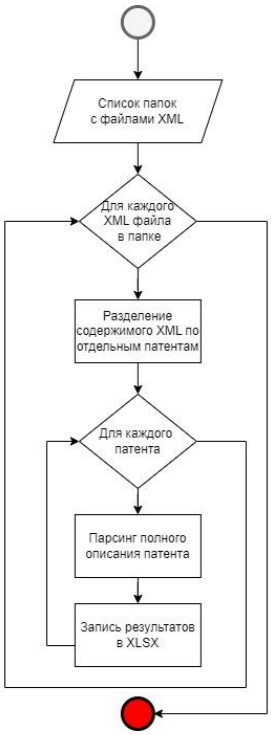


Рисунок 2 – Алгоритм парсинга патентов
Figure 2 – Patent parsing algorithm

Рисунок 3 – Алгоритм формирования обучающей выборки
Figure 3 – Algorithm for creating a training sample

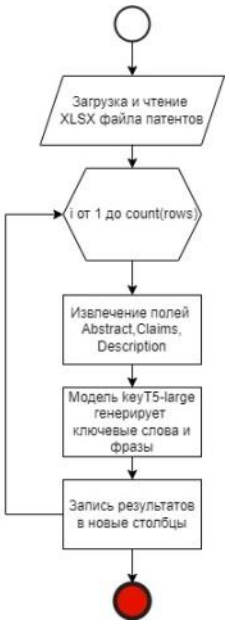


Рисунок 5 – Алгоритм извлечения ключевых фраз и слов на обученной модели
Figure 5 – Algorithm for extracting keywords and words using a trained model

Архитектура системы

На основе всех реализованных алгоритмов описывается архитектура системы, где входные данные представляют собой XML-файлы патентов, которые парсятся и сохраняются в xlsx-файле, содержащем все поля патентов, включая полнотекстовые поля. С помощью алгоритма формирования обучающей выборки для каждой строки патента в xlsx-файле поля Abstract, Claims и Description переводятся на английский язык. Переведенные тексты на английском языке подаются на вход модели извлечения ключевых фраз и слов keyPhrasesTransformer, которая извлекает ключевые фразы и слова на английском языке. Извлеченные ключевые фразы и слова на английском языке переводятся на русский язык. В результате получается файл, где в первом столбце содержится полнотекстовое описание каждого патента, а во втором – ключевые слова и фразы на русском языке. На основе этих данных обучается модель keyT5-large. Затем используется обученная модель для распарсенного xlsx-файла, и для каждой строки извлекаются ключевые слова и фразы на русском языке, результаты записываются в файл JSON. Файл JSON сохраняется в HDFS. С помощью поисковой системы Apache Solr данные из HDFS индексируются. После этого можно осуществлять полнотекстовый поиск по всем полям патента, а также по полям патента с извлеченными ключевыми фразами и словами, которые предоставила модель KeyT5-large. В результате полученное решение интегрируется с веб-системой на основе фреймворка Django.

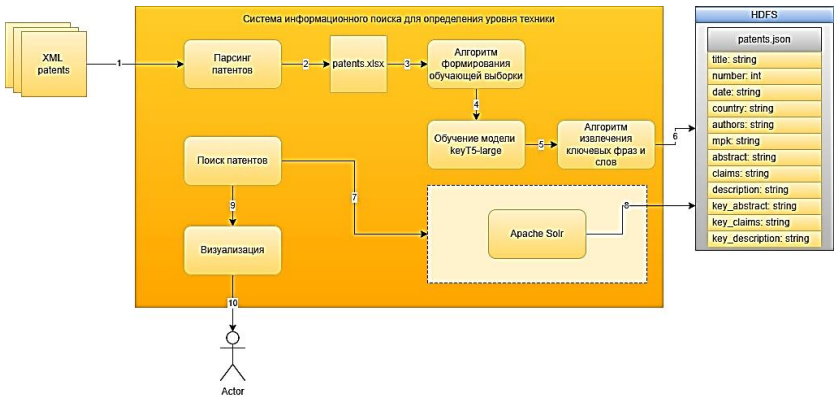


Рисунок 6 – Архитектура системы
Figure 6 – System architecture

Применение системы Apache Solr для полнотекстового поиска и индексации данных из HDFS

Hadoop HDFS и Apache Solr хорошо сочетаются в контексте обработки и поиска больших объемов данных. Объединение HDFS и Solr позволяет эффективно хранить и индексировать данные, а также выполнять сложные операции поиска и анализа.

HDFS, распределенная файловая система Hadoop, предоставляет надежное и масштабируемое хранилище для данных. Она разбивает большие файлы на блоки и распределяет их на различные узлы кластера. HDFS также обеспечивает отказоустойчивость и высокую доступность данных, создавая несколько копий каждого блока. Это идеально подходит для хранения больших объемов данных, таких как логи, медиафайлы и др.

Solr, в свою очередь, является мощной платформой поиска и аналитики, основанной на Apache Lucene. Он предоставляет возможность индексации и поиска данных с использованием широкого спектра функций и возможностей. Solr позволяет создавать полнотекстовые индексы данных, выполнять сложные запросы и фильтры, а также проводить аналитику и агрегацию данных.

HDFS и Solr могут интегрироваться путем индексации данных из HDFS в Solr. Это позволяет быстро и эффективно искать и анализировать данные, хранящиеся в HDFS, с использованием мощных поисковых возможностей Solr. Данные из HDFS могут быть периодически синхронизированы с Solr, чтобы отражать обновления и изменения в данных [6].

Такая интеграция между HDFS и Solr позволяет использовать преимущества обеих систем: надежность и масштабируемость HDFS для хранения данных и мощные возможности поиска и аналитики Solr для эффективного извлечения информации [7]. Это особенно полезно при обработке и анализе больших объемов данных, когда требуется высокая производительность и точность поиска.

Разработка серверной части системы информационного поиска

Для создания серверной части приложения используется Django, популярный фреймворк Python. С помощью Django определяются основные компоненты, такие как отображение, чтение и поиск в индексированных данных, хранящихся в HDFS, с использованием Solr [8]. Django облегчает работу с Solr, маршрутизацией запросов и обработкой данных.

Пользователь имеет возможность выполнять поиск с учетом всех атрибутов патента, включая полное описание, ключевые фразы и слова. Система обеспечивает

соответствующий механизм поиска для выполнения запросов и получения соответствующих результатов, как показано на Рисунке 7.

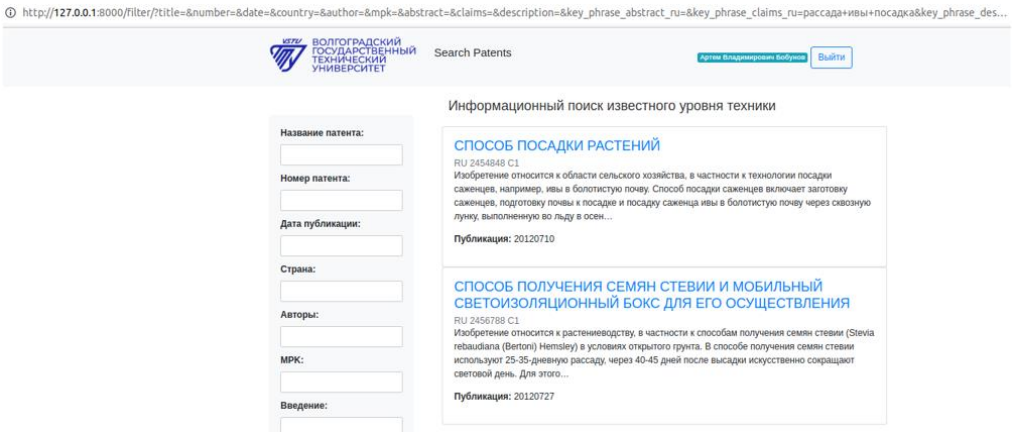
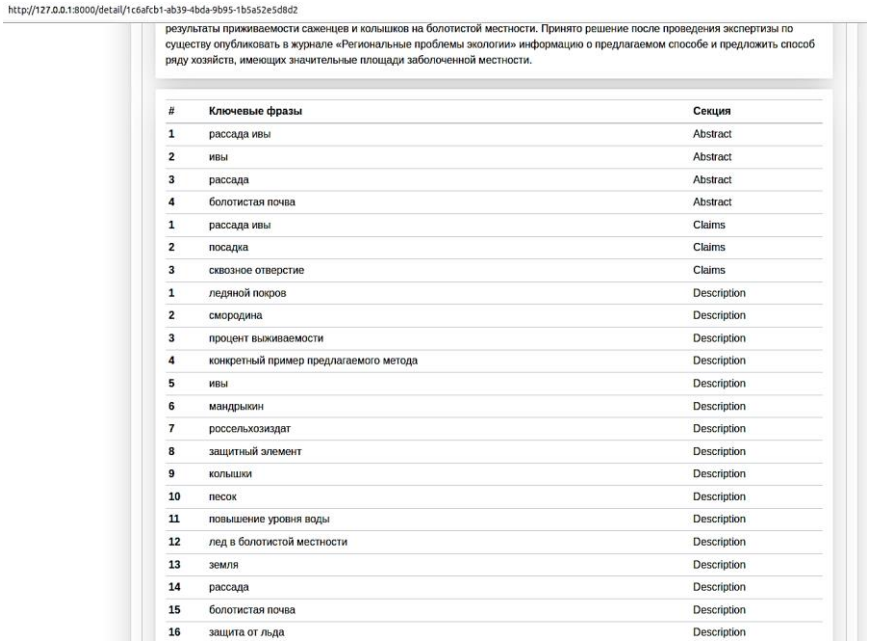


Рисунок 7 – Поиск по ключевым фразам и словам
Figure 7 – Search by key phrases and words

Пользователь имеет доступ к полному описанию патента, включая все атрибуты и извлеченные ключевые фразы и слова. Система обеспечивает правильное отображение и доступность полной информации о патенте для дальнейшего изучения и анализа.

Результаты поиска, представленные пользователю, включают извлеченные ключевые фразы и слова для каждого патента. Это позволяет пользователям быстро оценить релевантность и содержание найденных патентов, как показано на Рисунке 8.



#	Ключевые фразы	Секция
1	рассада ивы	Abstract
2	ивы	Abstract
3	рассада	Abstract
4	болотистая почва	Abstract
1	рассада ивы	Claims
2	посадка	Claims
3	сквозное отверстие	Claims
1	ледяной покров	Description
2	смородина	Description
3	процент выживаемости	Description
4	конкретный пример предлагаемого метода	Description
5	ивы	Description
6	мандрыкин	Description
7	россельхозиздат	Description
8	защитный элемент	Description
9	колышки	Description
10	песок	Description
11	повышение уровня воды	Description
12	лед в болотистой местности	Description
13	земля	Description
14	рассада	Description
15	болотистая почва	Description
16	защита от льда	Description

Рисунок 8 – Извлеченные ключевые фразы и слова к патенту
Figure 8 – Extracted keywords and words for the patent

В результате проведенного тестирования системы подтверждена ее работоспособность и соответствие требованиям. Тестирование включало проверку загрузки и хранения патентных файлов, поиска по атрибутам патента, отображения результатов с ключевыми фразами и словами, доступа к полному описанию патента и использования системы через веб-интерфейс.

Положительные результаты тестирования подтвердили правильность работы системы, ее эффективность в обработке и хранении патентов, а также удобство использования через веб-интерфейс для пользователей. В целом система успешно прошла функциональное тестирование и соответствует требованиям по загрузке, хранению, поиску и предоставлению информации о патентах.

Результаты

В результате обучения модели keyT5-large, а также интеграции этой модели в разработанную систему, эффективность извлечения ключевых фраз и слов была повышена, и время, затраченное на этот процесс, сократилось.

Таблица 3 – Метрики результатов модели
Table 3 – Metrics of model results

Метрика	Точность	Полнота	F1-мера
Точность	0,764	0,513	0,614

Значение метрики точности составляет 0,764, что означает, что примерно 76,4 % предсказанных меток являются верными. То есть из всех меток, которые модель предсказала как положительные, около 76,4 % действительно являются положительными.

Значение метрики полноты составляет 0,513, что означает, что модель смогла извлечь около 51,3 % от всех истинных положительных меток. То есть из всех реальных положительных меток, модель обнаружила примерно 51,3 % и корректно предсказала их.

Значение F1-меры составляет 0,614, которая является гармоническим средним между точностью и полнотой. F1-мера используется для оценки сбалансированности модели, учитывая и точность, и полноту. Чем ближе значение F1-меры к 1, тем лучше сбалансирована модель в предсказании положительных и отрицательных меток [9].

На основе полученных значений метрик, как показано в Таблице 3, можно сделать следующие выводы:

1) модель достаточно точно предсказывает истинные метки, так как значение точности составляет 0,764;

2) однако модель имеет невысокую полноту в извлечении релевантной информации, так как значение полноты составляет 0,513;

3) F1-мера равна 0,614, что указывает на сбалансированность между точностью и полнотой модели.

Таблица 4 – Скорость извлечения ключевых фраз и слов
Table 4 – Rate of extracting key phrases and words

Критерии оценки	Существующие системы («Яндекс Патенты», ФИПС «Роспатент»)	Автоматизированная система
Время извлечения ключевых фраз и слов	приблизительно 9 мин.	1,5–2 мин.

В ручном анализе патентных заявок для определения ключевых фраз и слов тратится значительное время до 9 минут, особенно для русскоязычных патентов [10]. Однако автоматизированная система может выполнить эту задачу за 1,5–2 минуты, что значительно сокращает время и повышает эффективность анализа, как показано в Таблице 4.

Таким образом, модель имеет неплохую точность в предсказании, но есть потенциал для улучшения полноты в извлечении релевантной информации. Скорость извлечения ключевых фраз и слов из патента уменьшилась.

Заключение

В результате проделанной работы была создана система информационного поиска для определения уровня техники. Были проанализированы существующие системы и процессы поиска патентов. На основе сравнительного анализа были сформулированы этапы разработки системы. Для способа извлечения ключевой информации из текста была использована модель обработки естественного языка keyT5-large, обученная на русских данных, позволяющая извлекать ключевые слова и фразы из текста. На этапах проектирования системы были разработан алгоритм парсинга патентов, формирования обучающей выборки, обучения модели и алгоритм извлечения ключевых фраз и слов на русском языке для патентного массива. В результате алгоритмов были получены данные, которые в процессе разработки системы загружены в распределенную файловую систему HDFS. Для полнотекстового поиска использовалась система Apache Solr, которая просто настраивается с системой. Это позволило проиндексировать данные из HDFS в Solr, после чего осуществлялся поиск по данным из HDFS. В результате перечисленных действий весь процесс был интегрирован во фреймворк Django для более простого использования системы.

Разработанная система позволяет осуществлять поиск по всем атрибутам патента, а также по полнотекстовому описанию библиографических данных патента и по ключевым фразам и словам, извлеченным из полнотекстового описания патента с использованием модели keyT5-large, обученной на патентных данных.

СПИСОК ИСТОЧНИКОВ

1. Жарова А.К. Интеллектуальное право. *Защита интеллектуальной собственности*. М.: Издательство «Юрайт»; 2023. 379 с.
2. Ишков А. Д., Степанов А.В. *Промышленная собственность. Оформление заявки на выдачу патента на полезную модель*. М.: Флинта; 2013. 98 с.
3. Zhiwei F. *Formal Analysis for Natural Language Processing: A Handbook*. Singapore, Springer; 2023. 796 p.
4. Романадзе Е.Л., Судаков В.А., Кислинский В.Г. Разработка метода извлечения ключевых слов на основе вероятностной тематической модели. *Моделирование и анализ данных*. 2022;12(2). URL: https://psyjournals.ru/journals/mda/archive/2022_n2/Romanadze_et_al. DOI: 10.17759/mda.2022120202 (дата обращения: 22.07.22).
5. Феногенова А., Тихонова М., Михайлов В. Русский SuperGLUE 1.1: пересмотр уроков, не усвоенных российскими моделями НЛП. ArXiv. 2022; 2202.07791. URL: https://arxiv.org/abs/2202.07791. DOI: 10.28995/2075-7182-2021-20-235-245 (дата обращения: 12.11.2022).
6. Koitzsch K. Advanced Search Techniques with Hadoop, Lucene, and Solr. In: *Pro Hadoop Data Analytics*. Berkeley, Apress; 2017. 298 p. DOI: 10.1007/978-1-4842-1910-2.
7. Wadkar S., Siddalingaiah M. *Pro Apache Hadoop.* CA: Apress Berkeley; 2014. 413 p. DOI: 10.1007/978-1-4302-4864-4.
8. Abu-Salih B., Wongthongtham P., Zhu D., Chan K.Y., Rudra A. *Introduction to Big Data Technology*. Singapore, Springer; 2021. 218 p.
9. Дудченко П.В. Метрики оценки классификаторов в задачах медицинской диагностики. *Молодежь и современные информационные технологии: сборник трудов XVI Международной научно-практической конференции студентов,*

аспирантов и молодых учёных, 3–7 декабря 2018, Томск. Томск: Изд-во ТПУ; 2019. Режим доступа: http://earchive.tpu.ru/handle/11683/52692 (дата обращения: 12.02.2023).
10. Николаев А.С. *Патентная аналитика: учебно-методическое пособие.* СПб: Университет ИТМО; 2022. 98 с.

REFERENCES

1. Zharova A.K. *Intellectual law. Intellectual property protection*. Moscow, Yurite Publishing House; 2023. 379 p. (In Russ.).
2. Ishkov A.D., Stepanov A.V. *Industrial property. Registration of a patent application for a utility model*. Moscow, Flinta; 2013. 98 p. (In Russ.).
3. Zhiwei F. *Formal Analysis for Natural Language Processing: A Handbook*. Singapore, Springer; 2023. 796 p.
4. Romanadze E.L., Sudakov V.A., Kislinsky V.G. Development of a method for extracting keywords based on a probabilistic thematic model. *Data modeling and analysis*. 2022;12(2). URL: https://psyjournals.ru/journals/mda/archive/2022_n2/Romanadze_et_al. DOI: 10.17759/mda.2022120202 (accessed on 22.07.22). (In Russ.).
5. Phenogenova A., Tikhonova M., Mikhailov V. Russian SuperGLUE 1.1: revision of lessons not learned by Russian NLP models. ArXiv. 2022; 2202.07791. URL: https://arxiv.org/abs/2202.07791. DOI: 10.28995/2075-7182-2021-20-235-245 (accessed on 12.11.2022). (In Russ.).
6. Koitzsch K. Advanced Search Techniques with Hadoop, Lucene, and Solr. In: *Pro Hadoop Data Analytics*. Berkeley, Apress; 2017. 298 p. DOI: 10.1007/978-1-4842-1910-2.
7. Wadkar S., Siddalingaiah M. *Pro Apache Hadoop.* CA: Apress Berkeley; 2014. 413 p. DOI: 10.1007/978-1-4302-4864-4.
8. Abu-Salih B., Wongthongtham P., Zhu D., Chan K.Y., Rudra A. *Introduction to Big Data Technology*. Singapore, Springer; 2021. 218 p.
9. Dudchenko P.V. Metrics for assessing classifiers in medical diagnostic tasks. In: The youth and modern information technologies: Proceedings of XVI International scientific and practical conference for undergraduate, postgraduate students and young researchers, *3–7 December 2018, Tomsk.* Tomsk, Publishing House of TPU; 2019. URL: http://earchive.tpu.ru/handle/11683/52692 (accessed on 12.02.2023). (In Russ.).
10. Nikolaev A.S. *Patent analytics: educational and methodological manual*. Saint Petersburg, ITMO University; 2022. 98 p. (In Russ.).

ИНФОРМАЦИЯ ОБ АВТОРАХ / INFORMATION ABOUT THE AUTHORS

Бобунов Артем Владимирович, магистр, Волгоградский государственный технический университет, Волгоград, Российская Федерация.
e-mail: btema1999@yandex.ru

Artem Vladimirovich Bobunov, Master's Student, 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 Mikhailovich 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

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

Статья поступила в редакцию 20.06.2023; одобрена после рецензирования 29.07.2023; принята к публикации 20.09.2023.

The article was submitted 20.06.2023; approved after reviewing 29.07.2023; accepted for publication 20.09.2023.