

ИНТЕЛЛЕКТУАЛЬНЫЙ АНАЛИЗ ПАТЕНТНЫХ МАТЕРИАЛОВ НА ПРИМЕРЕ КВАНТОВОЙ И ИНТЕЛЛЕКТУАЛЬНОЙ РОБОТОТЕХНИКИ

Д. П. Зрелова^{1,*}, *П. В. Зрелов*¹, *В. В. Кореньков*¹,
*И. Д. Соколов*²

¹ Объединенный институт ядерных исследований, Дубна, Россия

² ООО «Системы информационной аналитики», Москва, Россия

Осуществляется выбор методики интеллектуального анализа патентной информации, включая этапы сбора данных, их предобработки, статистического и семантического анализа. Особое внимание уделяется применению нейросетевых моделей и LDA-моделирования для выявления скрытых закономерностей в массивах патентных документов.

The paper presents a methodology for the intelligent analysis of patent information, covering data collection, preprocessing, and both statistical and semantic analysis. Special emphasis is placed on the use of neural network models and Latent Dirichlet Allocation (LDA) to reveal latent structures in large-scale patent corpora.

PACS: 89.20.Ff

ВВЕДЕНИЕ

В условиях стремительного развития технологий и усиления глобальной конкуренции анализ научно-технической информации приобретает ключевое значение. Одним из наиболее информативных источников таких данных являются патентные документы, которые не только фиксируют инновационные решения, но и отражают стратегии компаний, государств и научных организаций. Интеллектуальный анализ патентов позволяет выявлять технологические тренды, оценивать конкурентную среду, оптимизировать научные исследования и определять недостаточно изученные направления для патентования в исследуемых областях знаний.

МЕТОДИКА ПРОВЕДЕНИЯ ПАТЕНТНОГО АНАЛИЗА

Методика исследования включает комплексный подход к обработке и интерпретации патентных данных. На первом этапе была определена аналитическая задача — выявление технологических трендов в области

* E-mail: zrelova@jinr.ru

робототехники с акцентом на интеллектуальные и квантовые технологии. Для ее решения использовались данные патентной базы Lens.org*, отобранные с помощью специализированных поисковых запросов и фильтров по временному интервалу. Выбор платформы Lens.org был обусловлен ее интеграцией данных из ведущих патентных ведомств и научных репозиторий более чем из 100 стран.

Аналитическая работа состояла из двух взаимодополняющих этапов. Первый этап включал статистическую обработку данных, в ходе которой осуществлялась группировка патентов по технологическим направлениям, компаниям-заявителям и странам происхождения. Второй этап исследования представлял собой углубленный семантический анализ текстов патентных документов с целью обнаружения скрытых взаимосвязей между технологическими направлениями.

Заключительной фазой исследования стала визуализация полученных результатов. С помощью специализированных инструментов были созданы интерактивные графики, диаграммы и тематические карты, наглядно демонстрирующие выявленные закономерности.

СБОР И ОБРАБОТКА ПАТЕНТНЫХ ДАННЫХ. СТАТИСТИЧЕСКИЙ АНАЛИЗ

Формирование выборки для исследования осуществлялось по запросу «robotics AND (intelligent OR quantum)» за 2013–2024 гг. — период, наиболее репрезентативный для анализа современных тенденций в интеллектуальных и квантовых технологиях робототехники. На основе результатов патентного поиска была сформирована выборка из 48 052 патентных документов, соответствующих заданным критериям.

В связи с ограничением используемого API, допускающего загрузку не более 1000 документов за один запрос, был разработан специализированный алгоритм последовательной выгрузки и структурирования данных. Этот алгоритм обеспечивает извлечение ключевых метаданных — заголовков, аннотаций, дат подачи и публикации, МПК-кодов, сведений о заявителях и правообладателях — с сохранением в формате JSON, оптимальном для последующего анализа.

Для обработки и анализа больших объемов данных использовалась распределенная поисковая платформа Elasticsearch, обеспечивающая эффективное хранение, индексацию и полнотекстовый поиск. Визуализация полученных результатов осуществлялась с помощью аналитического инструмента Kibana.

Статистический анализ выявил значительный рост патентной активности в исследуемой области — пик пришелся на 2022 г. (11 870 за-

* Global patent and scholarly knowledge public resource Lens.org. <https://lens.org/>.

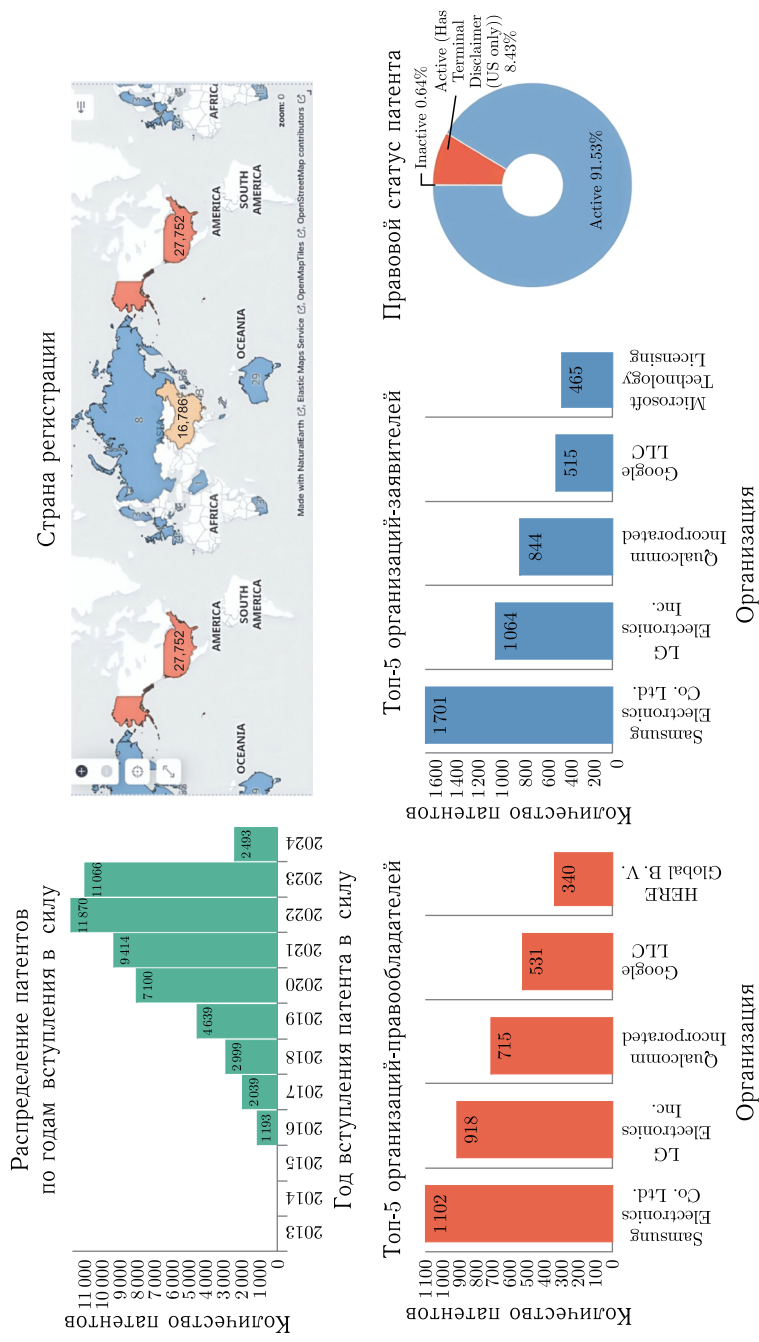


Рис. 1. Фрагмент аналитической панели, созданной в среде Kibana: динамика патентования, ведущие организации-заявители и правообладатели, правовой статус патентов

явок). 91,53 % патентов сохраняют активный статус, что подчеркивает их устойчивую технологическую и коммерческую значимость. Лидерами по числу заявок выступают Samsung, LG и Qualcomm, что свидетельствует об их стратегической ориентации на развитие передовых технологий в робототехнике. Анализ распределения патентов по странам продемонстрировал выраженное доминирование США (27 752 патента) и Китая (16 786 патентов) в общем массиве патентной документации (рис. 1).

Анализ структуры патентного массива по Международной патентной классификации (МПК, IPC) — стандартизированной системе категоризации изобретений по техническим областям — позволил выделить три ключевые технологические области: 1) G05D 1/02 — управление местоположением или курсом в системах двух координат (1986 патентов); 2) B25J 11/00 — манипуляторы, не отнесенные к другим рубрикам (1865 патентов); 3) G06K 9/00 — способы и устройства для считывания и распознавания напечатанных или написанных знаков или распознавания образов, например отпечатков пальцев (1839 патентов). Это свидетельствует о высокой патентной активности в сферах автономного управления, специализированных манипуляторов и технологий распознавания образов.

АНАЛИЗ ПАТЕНТНОГО КОРПУСА: КЛАСТЕРИЗАЦИЯ И ТЕМАТИЧЕСКОЕ МОДЕЛИРОВАНИЕ

Семантический анализ собранных данных выполнялся с применением методов NLP* [1, 2] и библиотек языка R**, включая предварительный анализ по рубрикам, хронологическим периодам, странам и иным параметрам. Предобработка включала удаление знаков разметки и пунктуации, удаление стоп-слов и др. Отсутствие аннотаций было выявлено для 5,4 % из 48 052 патентов, незначительная доля документов (0,6 %) была представлена на национальных языках (китайском, японском, корейском, французском, немецком и др.), и для дальнейшего анализа отбирались только патенты на английском языке. Окончательный список патентов включал 42 618 аннотаций. На основе текстов аннотаций был сформирован корпус документов.

Для выявления семантических связей документы подвергались кластеризации в векторном пространстве, построенном с помощью нейросетевых моделей Word2Vec [3] и Doc2Vec [4]. Для униграмм и биграмм кластеризация проводилась отдельно. Размерность семантического пространства полагалась равной $n = 200$, кластеризация проводилась методом k -средних после редукции признакового пространства до 10

* NLP — Natural Language Processing (обработка естественного языка).

** The R Project for Statistical Computing. <https://www.r-project.org/>.

с помощью автокодирующей нейронной сети (АЕ) с числом кластеров, равным 8. На основе анализа с помощью TF-IDF были выявлены характерные слова (униграммы) для каждого кластера. Например, для первого кластера — *grille*, *organotypic*, *deterrent* (биологические и медицинские термины), для второго — *fland*, *porridge*, *defrosting* (пищевые технологии) и т. д.

С целью выявления скрытых тем в корпусе патентных текстов использовалась вероятностная модель тематического моделирования LDA (Latent Dirichlet Allocation) [5] отдельно для униграмм и биграмм. Количество тем было задано равным $k = 5$. Для каждого документа определялась ведущая LDA-тема. Информация о корреляциях между кластерами и темами определялась с помощью матрицы сопряженности «кластеры $\times$ темы».

Результаты LDA-моделирования на униграммах не позволили четко определить названия тем, поэтому в дальнейшем анализе использовались биграммы. На рис. 2 представлены пять тем LDA, каждая в виде 10 ранжированных списков биграмм. Первые 3 биграммы каждой темы фор-

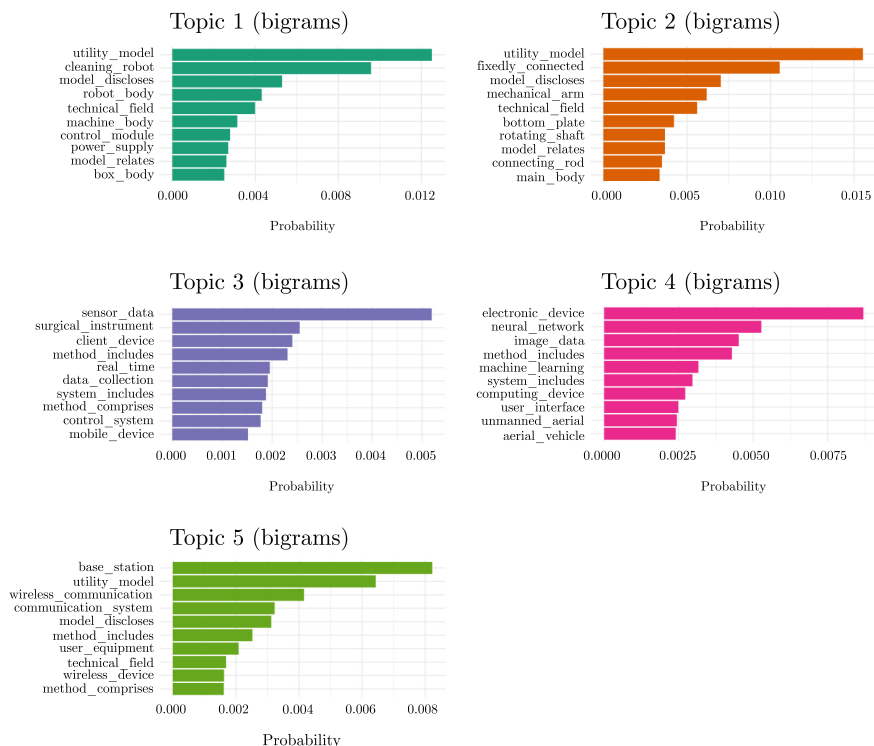


Рис. 2. Темы LDA, представленные в виде 10 ранжированных списков биграмм

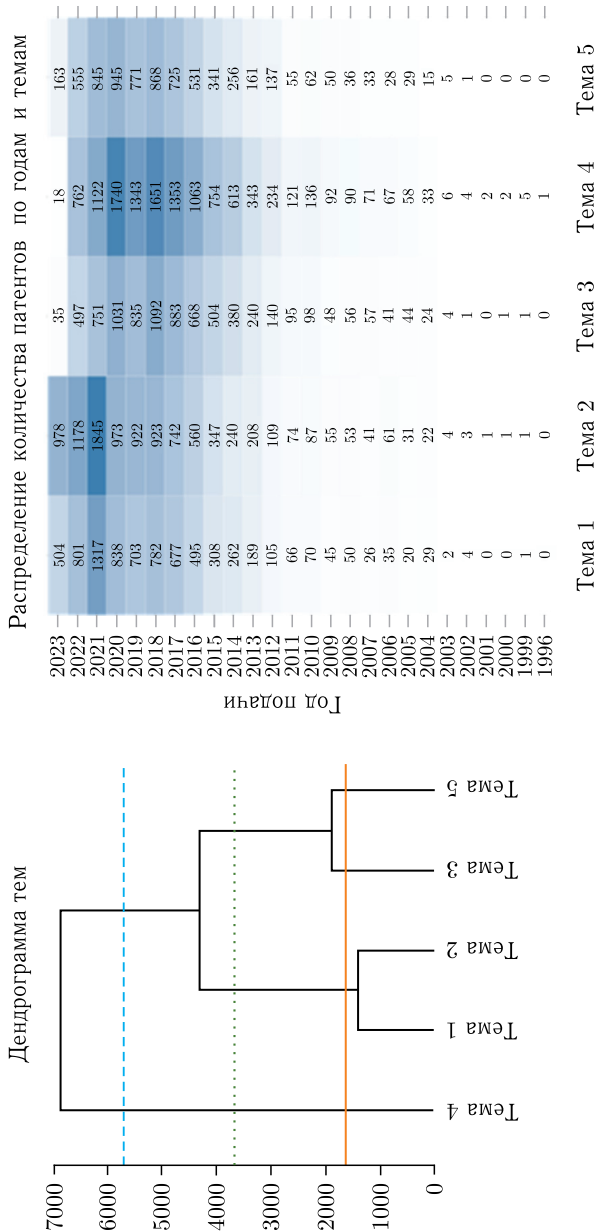


Рис. 3. Слева — дендрограмма тем LDA: на нижнем уровне представлены все 5 тем, «разрез» по сплошной линии определяет 4 кластера, «разрез» по пунктирной линии — 3 кластера, «разрез» по штриховой линии — 2 кластера. Справа — распределение заявок на патенты по годам подачи для 5 выявленных с помощью LDA тем (на биграммах)

мировали ее название*: тема 1 «Cleaning robot, robot body and machine body»; тема 2 «Fixedly connected, mechanical arm and rotating shaft»; тема 3 «Sensor data, surgical instrument and real time»; тема 4 «Electronic device, neural network and image data»; тема 5 «Base station, wireless communication and user equipment».

На рис. 3, слева представлена дендрограмма для пяти LDA-тем, построенная на основании распределений биграмм. Хорошо видно, что в зависимости от поставленной цели результаты могут быть представлены в виде 1, 2, 3, 4 кластеров (новых тем) или 5 начальных тем. При объединении тем строится общий ранжированный список с учетом параметра «заметности» (saliency): тема (1 + 2) «Cleaning robot, fixedly connected and mechanical arm»; тема (3 + 5) «Sensor data, base station and surgical instrument»; тема (1 + 2 + 3 + 5) «Cleaning robot, sensor data and fixedly connected»; тема (1 + 2 + 3 + 4 + 5) «Cleaning robot, sensor data and electronic device».

Для временного анализа год подачи заявки или год выдачи патента использовались как временные метки. На рис. 3, справа представлено распределение количества заявок по годам их подачи отдельно для каждой темы. Хорошо видно, что темы имеют разные временные характеристики.

ЗАКЛЮЧЕНИЕ

В рамках настоящего исследования разработана методика комплексного интеллектуального анализа патентных документов, интегрирующая этапы автоматизированного сбора, хранения и предобработки текстовых данных с последующим статистическим анализом метаинформации. Семантическое моделирование реализовано на основе нейросетевых архитектур Word2Vec и Doc2Vec, а также LDA-моделирования. Методика включает процедуры выявления семантических паттернов и тематических кластеров в массиве патентной документации.

В ходе исследования выявлены значимые закономерности в развитии исследуемых технологических направлений, идентифицированы ключевые технологические тренды и определены перспективные направления исследований.

Практическая значимость работы заключается в возможности применения разработанной методики для прогнозирования технологических трендов, анализа конкурентной среды и поддержки принятия стратегических решений в области исследований и разработок. Перспективы дальнейших исследований связаны с применением трансформерных моделей машинного обучения, искусственного интеллекта и систем предиктивной аналитики.

* Для формирования названий биграммы фильтровались с помощью дополнительного «стоп-листа», состоящего из «патентных шаблонов». Ранг биграмм корректировался в соответствии с величиной их «заметности» (saliency).

СПИСОК ЛИТЕРАТУРЫ

1. *Eisenstein J.* Introduction to Natural Language Processing. MIT Press, 2019. P. 1.
2. Белов С. Д., Зрелова Д. П., Зрелов П. В., Кореньков В. В. Обзор методов автоматической обработки текстов на естественном языке // Систем. анализ в науке и образовании: Сетевое науч. изд. 2020. № 3. С. 8–22; <http://sanse.ru/download/401>.
3. *Mikolov T. et al.* Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781 [cs.CL]. 2013.
4. *Quoc V. Le, Mikolov T.* Distributed Representations of Sentences and Documents // Proc. 31st Intern. Conf. on Machine Learning (ICML 2014). arXiv:1405.4053 [cs.CL]. 2014.
5. *Blei D. M., Ng A. Y., Jordan M. I.* Latent Dirichlet Allocation // J. Mach. Learn. Res. 2003. V. 3, No. 4–5. P. 993–1022; doi: 10.1162/jmlr.2003.3.4-5.993.