ИНФОРМАЦИОННО-ПОИСКОВЫЕ ЯЗЫКИ

УДК 025.4.036
https://doi.org/10.33186/1027-3689-2026-2-122-137

Повышение качества автоматического патентного поиска уровня техники на основе дистрибутивной семантики и библиографических данных

А. В. Горбунов[1], Б. Л. Генин[2], Д, С. Золкин[3], И. В. Некрасов[4]

[1, 2, 3, 4]Федеральный институт промышленной собственности, Москва, Российская Федерация

Автор ответственный за переписку: Александр Владимирович Горбунов, gorbunov@rupto.ru, https://orcid.org/0009-0001-9503-0880

Аннотация. В статье описан подход к повышению качества автоматического патентного поиска уровня техники, решающий проблему недостаточной эффективности существующих систем. Подход основан на автоматическом формировании терминологического вектора запроса из текста заявки с последующим его расширением квазисинонимами из дистрибутивного тезауруса, построенного на корпусе патентных документов, и обогащением библиографическими данными – кодами Международной патентной классификации (МПК). Дана математическая формализация формирования и расширения вектора запроса, описано построение дистрибутивного тезауруса патентной лексики. Предложены оригинальные показатели оценки качества поиска, учитывающие специфику патентных документов – наличие так называемых «патентных семейств», что позволяет оценивать способность системы находить релевантные изобретения, а не только совпадающие номера документов. Эксперименты на русскоязычной и англоязычной коллекциях показали повышение показателя S@20 на 10% по сравнению с базовым поиском по ключевым словам, продемонстрировано влияние учета патентных семейств на оценку успешности результатов поиска. Независимая экспертная оценка поисков в русскоязычной коллекции патентных документов подтвердила, что система находит хотя бы один релевантный документ в 96,25% случаев. Разработанные алгоритмы внедрены в поисковую платформу Роспатента.

Ключевые слова: автоматический патентный поиск, дистрибутивная семантика, квазисинонимы, патентные семейства, информационный поиск, машинное обучение

Благодарности. Авторы благодарят доктора экономических наук О. П. Неретина и доктора педагогических наук Н. В. Лопатину за методологическую помощь, С. В. Сапожникова – за участие в разработке инфраструктуры.

Для цитирования: Горбунов А. В., Генин Б. Л., Золкин Д. С., Некрасов И. В. Повышение качества автоматического патентного поиска уровня техники на основе дистрибутивной семантики и библиографических данных // Научные и технические библиотеки. 2026. № 2. С. 122–137. https://doi.org/10.33186/1027-3689-2026-2-122-137

INFORMATION RETRIEVAL LANGUAGES

UDC 025.4.036
https://doi.org/10.33186/1027-3689-2026-2-122-137

Increasing quality of automated patent search based on distributional semantics and bibliographic data

Alexander V. Gorbunov[1], Boris L. Genin[2], Dmitry S. Zolkin[3] and Igor V. Nekrasov[4]

[1, 2, 3, 4]*Federal Institute of Industrial Property, Moscow, Russian Federation*

Corresponding author: Alexander V. Gorbunov, gorbunov@rupto.ru, https://orcid.org/0009-0001-9503-0880

Abstract. The authors describe the approach to increasing quality of automated patent search and emphasize the inefficiency of existing systems. The approach is based on automated formulation of query terminological vector out from the application text with its further extension with quasisynonyms from the distributional thesaurus built on the body of patent documents, and with further enrichment with bibliographic data – IPC (international Patent Classification) codes. Mathematical formalization of acquiring and extending query vector is provided; acquisition of patent distributional thesaurus is discussed. The authors propose original indicators of retrieval quality assessment with account to patent document specific character, i. e. the so-called “patent families”, which enables to evaluate the system’s capability to find the relevant inventions beyond the corresponding numbers. The experi-

ments with Russian and English-language collections demonstrate the S@20 indicator increase by 10% as compared to standard search by keywords. The authors conclude on the effect of inclusion of patent families on the evaluation of search results efficiency. The independent expertise of search performance in the Russian-language patent collection confirmed that the system delivered at least one relevant document in 96.25% cases. The developed algorithms have been implemented into the Rospatent search platform.

Keywords: computerized patent search, distributional semantics, quasisynonym, patent family, information search, machine learning

Acknowledgements. The authors offer their thanks to O. P. Neretin, Doctor of Economics, and N. V. Lopatina, Doctor of Pedagogy, for methodological support, and to S. V. Sapozhnikov, for his assistance in infrastructure development.

Cite: Gorbunov A. V., Genin B. P., Zolkin D. S., Nekrasov I. V. Increasing quality of automated patent search based on distributional semantics and bibliographic data // Scientific and technical libraries. 2026. No. 2, pp. 122–137. https://doi.org/10.33186/1027-3689-2026-2-122-137

Введение

Согласно Стратегии научно-технологического развития Российской Федерации (указ Президента РФ от 28.02.2024 № 145), обеспечение технологического суверенитета требует оперативного выявления и анализа мирового уровня техники в приоритетных отраслях. Одним из ключевых инструментов такого анализа является поиск уровня техники – ключевой этап патентной экспертизы, на который приходится до 70% времени работы эксперта [**1**]. Уровень техники, согласно ст. 1350 Гражданского кодекса РФ, включает любые сведения, ставшие общедоступными в мире до даты приоритета изобретения [**2**]. Качественный поиск позволяет оценить новизну и изобретательский уровень технического решения и имеет важное значение как для оценки патентоспособности, так и для задач стандартизации и технического регулирования, в которых требуется объективная оценка состояния технологий в конкретной области [**3**].

Как показали M. Lupu и A. Hunbury [**4**], системы поиска, основанные на ключевых словах и ручном составлении запросов, не обеспечивают достаточной полноты и точности – в основном из-за вариативности патентной лексики и роста объемов патентного фонда. Современные подходы к автоматизации поиска можно условно разделить на три направления:

1. Автоматическое расширение запросов, составленных вручную, с использованием тезаурусов и синонимов [**4**].

2. Полностью автоматическое построение запросов на основе текста заявки [**5**].

3. Методы векторизации текстов и сравнения векторных представлений документов [**6**].

Недостатком первых двух подходов является зависимость от лингвистического обеспечения и игнорирование контекстной семантики. Третий подход, хотя и учитывает семантику, но имеющиеся публикации не показали высоких результатов [**7**, **8**].

Кроме того, традиционные подходы к оценке качества поиска, используемые в информационном поиске и оценивающие успешность поиска релевантного документа, не соответствуют специфике патентного поиска, где релевантность определяется изобретениями, зачастую представленными в виде патентных семейств [**9**], то есть целой группой документов, любой из которых описывает искомый уровень техники.

В настоящей работе предлагается комплексный подход, объединяющий методы дистрибутивной семантики, статистического анализа текста и патентной аналитики для повышения качества автоматического поиска уровня техники.

Использованы следующие ключевые понятия:

Терминологический вектор документа – множество терминов документа с весами, вычисленными по формуле *tf-idf* (term frequency – inverse document frequency). Вес термина вычисляется как произведение частоты его встречаемости в документе *(tf)* на логарифм обратной частоты документов, содержащих этот термин *(idf),* и отражает важность термина для данного документа относительно всего корпуса документов [**10**].

Дистрибутивный тезаурус – лингвистический ресурс, в котором каждому термину сопоставлены квазисинонимы – термины, имеющие схожее распределение в корпусе патентных документов. В отличие от

лингвистических синонимов, квазисинонимы не обязательно совпадают по значению, но часто встречаются в одном контексте [**11**].

Сходство терминов понимается как мера того, насколько часто термины встречаются в одном контексте – чем выше сходство, тем вероятнее, что термины встречаются совместно в одних и тех же текстах, то есть являются квазисинонимами. Сходство терминов измеряется с помощью косинусной меры между их векторными представлениями, в данной работе векторные представления терминов получены с помощью модели FastText [**12**].

В работе используется особенность структуры патентных документов – наличие библиографического поля, так называемого поля ИНИД (56), в котором эксперт патентного ведомства указывает (цитирует) документы, отражающие уровень техники для данного изобретения [**13**]. В процессе обучения система поощряется, если находит в патентной базе документы или представителей патентных семейств, которые эксперт уже признал релевантными, то есть процитировал ссылки на них. Таким образом, документы, процитированные экспертизой, рассматриваются в качестве «правильных ответов» для обучения системы.

Методология

Общая схема подхода

Описываемый подход базируется на методе автоматического формирования запроса путем извлечения ключевых терминов из документа, для которого выполняется поиск уровня техники и расширения набора ключевых терминов (терминологического вектора) квазисинонимами, хранящимися в предварительно построенном дистрибутивном тезаурусе. Расширенный терминологический вектор преобразуется в поисковый запрос в любую промышленную поисковую машину (в данном случае – ElasticSearch), далее полученный отзыв ограничивается каким-то разумным числом документов, которые сможет проанализировать эксперт в приемлемое время. В данной работе использовались первые 20 документов выдачи (топ-20), которые и предъявляются эксперту в качестве документов уровня техники для сопоставления с рассматриваемым документом.

Из сказанного следует, что еще до того, как система сформирует свой первый запрос на поиск, должен быть подготовлен дистрибутивный тезаурус. В качестве исходных данных для обучения/дообучения дис-

трибутивного тезауруса предлагается использовать базу данных патентных документов. Этим решается сразу несколько важных вопросов:

1. Доступность. Патентные документы (как минимум, выданные патенты, но во многих случаях – и заявки, и ряд других документов, сопутствующих рассмотрению заявок) открыто публикуются всеми патентными ведомствами, их использование для машинного обучения не нарушает ничьих авторских прав.

2. Объем. Опубликованные патентные базы обширны и постоянно пополняются. На начало 2025 г. мировой патентный фонд содержал около 150 млн документов. Даже если принять оценку (заниженную), что только четверть этих документов является патентами – то есть размечена экспертами ведомств – их число, 35–40 млн, будет вполне достаточным для обучения и тестирования систем.

3. Соответствие задаче. Обученный дистрибутивный тезаурус учитывает лексику патентных документов и запросы на поиск также строятся на этой лексике.

Задача построения дистрибутивного тезауруса может быть описана следующим образом: из существующей коллекции документов *DC*, представленной как множество документов $d_1 \ldots d_n$, то есть $DC = \{d_1, d_2, \ldots, d_n\}$, извлечь все уникальные термины, сформировать множество уникальных терминов $T = \{t_1, t_2, \ldots, t_m\}$. Для каждого уникального термина $t \in T$ построить множество квазисинонимов $S^K(t)$, дистрибутивный тезаурус DT сформировать как объединенное множество терминов и их квазисинонимов.

Представим каждый документ коллекции DC как множество входящих в него терминов. Тогда множество всех уникальных терминов коллекции определяется как $T = \bigcup_{d \in D} d$, множество квазисинонимов s для $t \in T$ будет определено как первые K терминов из множества T (исключая сам термин t), превысивших заданный порог сходства:

$$S^K(t) = \{top - K_{s \in T \setminus \{t\}} sim(t, s) \geq \theta\}, \quad (1)$$

где $sim(t, s)$ – функция близости векторного представления термина и его квазисинонима (в данной работе используется косинусная мера сходства); $\theta \epsilon [0,1]$ – порог сходства.

Тогда дистрибутивный тезаурус DT определяется как

$$DT = \{(t, S^K(t)) | t \in T\}. \quad (2)$$

Как уже было сказано выше, для построения дистрибутивного тезауруса используется модель глубокого обучения по алгоритму FastText. Обученный дистрибутивный тезаурус используется для расширения терминологического вектора искомого документа.

После того как дистрибутивный тезаурус подготовлен, становится возможным выполнять поиск документов уровня техники с использованием автоматического извлечения ключевых слов (с учетом весов) и расширения полученного терминологического вектора квазисинонимами. Алгоритм поиска может быть представлен как последовательность шагов, направленных на решение следующей задачи: для произвольного документа d, представленного как множество терминов $d = \{t_1, t_2, \ldots, t_n\}$ сформировать (базовый) терминологический вектор V_d, расширить его квазисинонимами из дистрибутивного тезауруса, учесть положение терминов в разделах патентного документа (заголовок, реферат, формула, описание), учесть в векторе коды МПК, сформировать и направить запрос на поиск в ElasticSearch.

Рассмотрим шаги алгоритма.

1. Извлечение ключевых терминов из текстовых полей документа (название, реферат, формула, описание). В данной работе принята размерность базового терминологического вектора в 50 терминов, в качестве ключевых выбираются термины документа с максимальными весами *tf-idf*. Перед применением весов выполняется шкалирование *tf-idf* в диапазон от 1 до 2 включительно, это позволит избежать чрезмерного влияния терминов с высокими весами и полного игнорирования терминов, которые одновременно имеют небольшой вес и находятся в наименее влиятельной части документа:

$$w_i' = 1 + \frac{w_i - \min(W)}{\max(W) - \min(W)}, \quad (3)$$

где W – массив исходных значений весов; w_i – шкалируемое значение веса; w_i' — шкалированное значение веса, $w_i' \in [1, 2]$.

2. Коррекция весов терминов в зависимости от положения термина в документе. Интуитивно понятно, что термины в названии, реферате, формуле заявки или описании имеют разное значение для понимания

смысла документа. Следовательно, необходимо подобрать веса терминов таким образом, чтобы метрики отзыва системы были максимальными. Это было сделано на этапе эксперимента. После подбора весов все последующие поиски выполняются с учетом полученных значений.

Итоговый вес термина: $w_t = w'_t \cdot w_{\text{поле}}$.

3. Формирование базового терминологического вектора документа V_D происходит с учетом скорректированных весов терминов:

$$V_d = \{(t, w_t) | t \in d, w_t = tf - idf(t, d)\}_{top-50 \text{ по } w_t}. \quad (4)$$

4. Расширение квазисинонимами из тезауруса. Выбирается k наиболее близких квазисинонимов, их веса соответствуют наследуемому весу соответствующих терминов базового вектора, умноженному на косинусное расстояние между базовым термином и квазисинонимом:

$$V'_D = V_D \cup \{(s, w_t) | (t, w_t) \in V_D, (t, S^K(t)) \in DT, s \in top - k(S^K(t))\}. \quad (5)$$

5. Формирование запроса в ElasticSearch. Запрос формируется путем соединения терминов с весами по «ИЛИ». Библиографические данные (в данной работе использованы коды МПК [**13**], а именно уровни «раздел» и «класс») так же соединяются по «ИЛИ», затем добавляются к строке терминов по «И».

Оценка качества поиска

Традиционные метрики, включая precision и recall и др., ориентированы на оценку результатов одного поиска, но не позволяют оценить работу поисковой системы в целом, на множестве поисков, так как их усреднение маскирует крайние случаи [**14**]. В работе предложены и используются метрики, ориентированные на оценку поисковой системы.

Пусть TP_i – множество документов, указанных в поле ИНИД (56) для документа, размещенного в поисковую систему в *i-м* тестовом запросе;

TF_i – документы из патентных семейств TP_i;

$TPF_i = TP_i \cup TF_i$ – расширенное множество релевантных документов для *i-го* поиска.

Полезность выдачи системы оценивается тремя показателями: $S@20$ – доля поисков, в которых в топ-20 документов, найденных системой, попал хотя бы один документ, цитированный экспертом; $S_F@20$ – доля поисков, в которых в топ-20 документов, найденных системой, по-

пали документы, цитированные экспертом, либо документы их патентных семейств. Этот показатель лучше отвечает задачам патентного поиска – найти любые документы, описывающие изобретения, формирующие уровень техники для оцениваемого документа. $H@20$ показывает долю поисков, в которых система в топ-20 разместила максимально возможное количество документов, цитированных экспертом, то есть результат работы системы был максимально полезен:

$$S@20 = \frac{1}{N}\sum_{i=1}^{N}(s_i), \quad (6)$$

где s_i- индикатор, принимающий значение 0 или 1, в зависимости от того, присутствуют ли ссылки на цитаты экспертизы ведомства в первых 20 документах выдачи, то есть для ранжированного списка из 20 первых документов в выдаче системы $F_i^{(20)} = (f_1, ..., f_{20})$ по запросу q_i:

$$s_i = \mathbf{1}_{\{f_1,...,f_{20}\} \cap TP_i \neq \emptyset}. \quad (7)$$

$S_F@20$ рассчитывается следующим образом:

$$S_F@20 = \frac{1}{N}\sum_{i=1}^{N}(s_{Fi}). \quad (8)$$

Здесь документами, релевантными запросу q, считаются и документы, указанные экспертизой ведомства в поле (56), и члены их простых патентных семейств [**13**]. Объединение этих документов и простого патентного семейства, в которое входит рассматриваемый документ, формируют так называемый «семантический кластер патента» [9] – множество документов, объединенных общей предметной областью. Индикатор s_F учитывает попадание в выдачу не только цитирований экспертизы ведомства, но и любых других документов, принадлежащих семантическому кластеру рассматриваемого документа, точнее – документов из патентных семейств цитат экспертизы ведомства, то есть s_F реагирует на удачно выполненный поиск документов из одной предметной области (документов уровня техники) рассматриваемого документа. Для *i-го* поиска:

$$s_{Fi} = \mathbf{1}_{\{f_1,...,f_{20}\} \cap TPF_i \neq \emptyset}. \quad (9)$$

Наконец, $H@20$ рассчитывается по формуле

$$H@20 = \frac{1}{N}\sum_{i=1}^{N}(h_i), \quad (10)$$

где для *i-го* запроса на поиск q_i при $|TP_i| \geq 20$ индикатор h_i принимает значение 1, если все документы из топ-20 списка найденных системой документов являются ссылками на документы из поля (56) целевого документа. Для $|TP_i| < 20$ показатель h_i принимает значение 1, если все документы из поля (56) целевого документа попали в топ-20 списка найденных системой:

$$h_i = \mathbf{1}_{\left\{\left(|TP_i| \geq 20 \wedge F_i^{(20)} \subseteq TP_i\right) \vee \left(|TP_i| < 20 \wedge TP_i \subseteq F_i^{(20)}\right)\right\}}. \quad (11)$$

Эксперимент

Для эксперимента были подготовлены следующие коллекции и сформированы тестовые выборки:

Русскоязычная коллекция: 1 112 502 кластеров (Роспатент, 2000–2022). Тестовая выборка: 1150 патентов (документы кодов вида C1/C2 из публикаций 2020 г.).

Англоязычная коллекция: 14 415 916 кластеров (Патентное ведомство США, 2001–2022). Тестовая выборка: 1575 патентов (документы кодов вида B2, из публикаций 2020 г.).

Эксперимент состоял из последовательности шагов поиска уровня техники для каждого документа тестовой выборки, причем для русскоязычных документов первым шагом выполнялся поиск с использованием только базового терминологического вектора, а на каждом следующем шаге базовый терминологический вектор дополнялся, как представлено в табл. 1. Показатель $S@20$ вычислялся для каждого шага, остальные показатели были вычислены только для финальной версии поискового запроса.

Эксперимент по поиску для англоязычных документов имел своей целью подтвердить значимость учета патентных семейств при поиске в тех коллекциях, для которых характерно наличие обширных патентных семейств, – именно так обстоит дело для патентов США. Эксперимент был проведен только для поисковых запросов по расширенному терминологическому вектору. Достигнутые значения всех измеренных показателей для документов русскоязычного (RU) и англоязычного (EN) тестовых наборов представлены в табл. 2.

Таблица 1

Эксперимент для русскоязычной тестовой выборки

Шаг эксперимента	Терминологический вектор	Значение $S@20$
1	Базовый поиск (50 ключевых терминов, *tf-idf*)	0,593
2	К каждому термину добавлены $k = 2$ квазисинонимов.	0,614
3	Веса терминов их квазисинонимов скорректированы в зависимости от положения терминов в документе (поля документа «Название», «Реферат», «Формула изобретения», «Описание»). Веса для корректировки были определены путем опроса экспертов.	0,636
4	Произведен подбор оптимальных весов, полученные веса: «Название» = 1, «Реферат» = 1,1, «Формула изобретения» = 1,3, «Описание» = 1.	0,644
5	К терминологическому вектору добавлены коды МПК документа (section и class).	0,652

Таблица 2

Показатели для поисков с полным терминологическим вектором

Тестовый набор	$S@20$	$S_F@20$	$H@20$
RU	0,652	0,653	0,272
EN	0,617	0,672	0,010

Разница между $S_F@20$ и $S@20$ для англоязычных документов (+8,9%) подтверждает важность учета патентных семейств.

Экспертная оценка

Как было сказано выше, результаты системы оцениваются сравнением с уже известными результатами рассмотрения документов тестовой выборки экспертами патентного ведомства: система получает баллы, если находит документы, которые были процитированы экспертами ведомства как документы уровня техники (для показателя $S_F@20$ учитываются и документы из патентных семейств). «Ручная» оценка экспертами полученного отзыва системы может дать ответ на вопрос: «Приводит ли использование предложенного подхода к формированию запроса на поиск к выявлению документов, так же хорошо описывающих уровень техники, как и те, которые уже нашел и процитировал эксперт ведом-

ства?» Положительный ответ на этот вопрос показывает, что поиск при таком подходе – это шаг от поиска релевантных к пертинентным документам в отзыве системы.

Для проверки были отобраны 80 русскоязычных заявок на патенты, по которым уже были приняты положительные решения экспертизы (то есть поиск на уровень техники проведен и документы уровня техники указаны в патенте). Все документы имеют код МПК G01N, по каждому из документов был проведен автоматический поиск. Для результатов поиска по каждому документу (всего 80 поисков, топ-20 документов отзыва системы в каждом поиске) был проведен расчет показателя $S@20$: а) на основании учета только цитат экспертизы ведомства, приведенных в поле ИНИД (56) и б) на основании «ручной» оценки экспертами – патентными поверенными ЦИС «Сколково». Результаты представлены в табл. 3.

Таблица 3

Результаты «ручной» проверки работы системы поиска

Характеристика	Значение
$S@20$: а) рассчитанный на основании поля ИНИД (56), б) рассчитанный с учетом суждений внешних экспертов о соответствии отзыва уровню техники.	а) 68,29% б) 96,25%
Среднее число документов уровня техники в отзыве системы, подтвержденное «ручной» оценкой.	7,7
Среднее число ссылок в поле ИНИД (56).	3,4
Количество заявок из выборки, для которых система не нашла ни одного документа уровня техники.	3

Результаты экспертной оценки подтверждают, что автоматический поиск может быть полезен экспертам при рассмотрении заявок на изобретения.

Обсуждение

1. Патентные семейства как основа оценки

Переход от оценки «документ ↔ документ» к «изобретение ↔ изобретение» соответствует сути патентного поиска. Система ищет не конкретный номер патента, а техническое решение, которое может быть описано в нескольких документах, одинаково релевантных с точки

зрения эксперта ведомства. Использование оценки, учитывающей патентные семейства, приближает результаты поиска не просто к выявлению релевантных, но и к получению пертинентных целям экспертизы документов.

2. Квазисинонимы или лингвистические синонимы?

Дистрибутивный тезаурус учитывает контекстную близость в патентном корпусе, что критически важно для технической лексики. В отличие от лингвистических тезаурусов, он лучше адаптируется к вариативности в терминологии и способен выявлять семантические связи. Например, «катод» и «анод» не являются синонимами, но часто встречаются вместе в описаниях электронных устройств, что делает их квазисинонимами в дистрибутивном тезаурусе.

3. Роль МПК

Добавление кодов МПК (уровни «раздел» и «класс») сужает поиск до релевантной технической области или областей, снижая уровень шума в отзыве системы. Это особенно важно в крупных коллекциях, где один и тот же термин может использоваться в разных отраслях с разным смыслом. Учет МПК позволяет избежать ложных срабатываний и повысить точность выдачи.

4. Практическая применимость

Алгоритм внедрен в поисковую платформу Роспатента, эксплуатируемую в настоящее время. Кроме того, подход может быть адаптирован для любых задач, где требуется объективная оценка уровня технологий в конкретной области.

Заключение

Предложенный подход обеспечивает значительное (на 10%) повышение качества автоматического патентного поиска за счет:

расширения запроса квазисинонимами из дистрибутивного тезауруса,

учета структуры документа и библиографических данных,

введения адекватной метрики оценки с учетом патентных семейств.

Подход инвариантен к языку, что позволяет применять его на многоязычных коллекциях, и может быть применен для совместного поиска патентной и непатентной литературы.

Список источников

1. **Lahorte P.** Inside the mind of an EPO examiner // World Patent Information. 2018. Vol. 54, Suppl. P. S18–S22. DOI 10.1016/j.wpi.2017.03.005.

2. **Гражданский** кодекс Российской Федерации (часть четвертая) от 18.12.2006 № 230-ФЗ. Ст. 1350.

3. **Указ** Президента Российской Федерации от 28.02.2024 № 145 «О Стратегии научно-технологического развития Российской Федерации» // Собрание законодательства РФ. 2024. № 9. Ст. 1123.

4. **Lupu M., Hanbury A.** Patent retrieval // Foundations and Trends in Information Retrieval. 2013. Vol. 7, № 1. P. 1–97.

5. **Mahdabi P., Crestani F.** Learning-based pseudo-relevance feedback for patent retrieval // Information Retrieval Facility Conference. Springer, 2012. P. 1–11.

6. **Hain, Daniel & Jurowetzki, Roman & Buchmann, Tobias & Wolf, Patrick.** (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. Technological Forecasting and Social Change. 177. 121559. 10.1016/j.techfore.2022.121559.

7. **Haghighian Roudsari, Arousha & Afshar, Jafar & Lee, Suan & Lee, Wookey.** (2021). Comparison and Analysis of Embedding Methods for Patent Documents. 152–155. 10.1109/BigComp51126.2021.00037.

8. **Ali, Amna & Tufail, Ali & De Silva, Liyanage & Abas, Pg Emeroylariffion.** (2024). Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches. Applied System Innovation. 7. 91. 10.3390/asi7050091.

9. **Горбунов А. В.** Кластерный подход к формированию наборов патентных данных и оценивание качества поиска «уровня техники» // Научные и технические библиотеки. 2025. № 5. С. 58–80. DOI 10.33186/1027-3689-2025-5-58-80.

10. **Manning C. D., Raghavan P., Schütze H.** Introduction to Information Retrieval. Cambridge University Press, 2008. 496 p.

11. **Harris C. G., Arens R., Srinivasan P.** Using classification code hierarchies for patent prior art searches // Current Challenges in Patent Information Retrieval. Springer, 2017. P. 287–304.

12. **Mikolov T. et al.** Efficient Estimation of Word Representations in Vector Space // arXiv preprint arXiv:1301.3781. 2013.

13. **WIPO.** Handbook on Industrial Property Information and Documentation. Glossary of Terms. URL: http://www.wipo.int/standards/en/pdf/08-01-01.pdf.

14. **Sakai T.** Metrics, Statistics, Tests: An IR Researcher's Guide to Analysing and Comparing Systems. Springer Nature Singapore, 2021. 335 p.

References

1. **Lahorte P.** Inside the mind of an EPO examiner // World Patent Information. 2018. Vol. 54, Suppl. P. S18–S22. DOI 10.1016/j.wpi.2017.03.005.

2. **Grazhdanskii`** kodeks Rossii`skoi` Federatcii (chast` chetvertaia) ot 18.12.2006 № 230-FZ. St. 1350.

3. **Ukaz** Prezidenta Rossii`skoi` Federatcii ot 28.02.2024 № 145 «O Strategii nauchno-tekhnologicheskogo razvitiia Rossii`skoi` Federatcii» // Sobranie zakonodatel`stva RF. 2024. № 9. St**.** 1123.

4. **Lupu M., Hanbury A.** Patent retrieval // Foundations and Trends in Information Retrieval. 2013. Vol. 7, № 1. P. 1–97.

5. **Mahdabi P., Crestani F.** Learning-based pseudo-relevance feedback for patent retrieval // Information Retrieval Facility Conference. Springer, 2012. P. 1–11.

6. **Hain, Daniel & Jurowetzki, Roman & Buchmann, Tobias & Wolf, Patrick.** (2022). A text-embedding-based approach to measuring patent-to-patent technological similarity. Technological Forecasting and Social Change. 177. 121559. 10.1016/j.techfore.2022.121559.

7. **Haghighian Roudsari, Arousha & Afshar, Jafar & Lee, Suan & Lee, Wookey.** (2021). Comparison and Analysis of Embedding Methods for Patent Documents. 152–155. 10.1109/BigComp51126.2021.00037.

8. **Ali, Amna & Tufail, Ali & De Silva, Liyanage & Abas, Pg Emeroylariffion.** (2024). Innovating Patent Retrieval: A Comprehensive Review of Techniques, Trends, and Challenges in Prior Art Searches. Applied System Innovation. 7. 91. 10.3390/asi7050091.

9. **Gorbunov A. V.** Clasterny`i` podhod k formirovaniiu naborov patentny`kh danny`kh i ocenivanie kachestva poiska «urovnia tekhniki» // Nauchny`e i tekhnicheskie biblioteki. 2025. № 5. S. 58–80. DOI 10.33186/1027-3689-2025-5-58-80.

10. **Manning C. D., Raghavan P., Schütze H.** Introduction to Information Retrieval. Cambridge University Press, 2008. 496 p.

11. **Harris C. G., Arens R., Srinivasan P.** Using classification code hierarchies for patent prior art searches // Current Challenges in Patent Information Retrieval. Springer, 2017. P. 287–304.

12. **Mikolov T. et al.** Efficient Estimation of Word Representations in Vector Space // arXiv preprint arXiv:1301.3781. 2013.

13. **WIPO.** Handbook on Industrial Property Information and Documentation. Glossary of Terms. URL: http://www.wipo.int/standards/en/pdf/08-01-01.pdf.

14. **Sakai T.** Metrics, Statistics, Tests: An IR Researcher's Guide to Analysing and Comparing Systems. Springer Nature Singapore, 2021. 335 p.

Информация об авторах / Authors

Горбунов Александр Владимирович – начальник Центра развития научного направления «Искусственный интеллект» Федерального института промышленной собственности, Москва, Российская Федерация

gorbunov@rupto.ru, https://orcid.org/0009-0001-9503-0880

Генин Борис Лемелевич – канд. техн. наук, ведущий научный сотрудник отдела проектирования информационно-поисковых систем Федерального института промышленной собственности, Москва, Российская Федерация

BGuenine@rupto.ru, https://orcid.org/0000-0003-3514-1340

Золкин Дмитрий Сергеевич – начальник отдела проектирования информационно-поисковых систем Федерального института промышленной собственности, Москва, Российская Федерация

db_dept@rupto.ru, https://orcid.org/0009-0002-1641-4518

Некрасов Игорь Валерьевич – научный сотрудник отдела проектирования информационно-поисковых систем Федерального института промышленной собственности, Москва, Российская Федерация

igor.nekrasov@rupto.ru, https://orcid.org/0009-0005-8322-7247

Alexander V. Gorbunov – Head, Center for Research on "Artificial Intelligence", Federal Institute of Industrial Property, Moscow, Russian Federation

gorbunov@rupto.ru, https://orcid.org/0009-0001-9503-0880

Boris L. Genin – Cand. Sc. (Engineering), Leading Researcher, Department for Information Retrieval Systems Design, Federal Institute of Industrial Property, Moscow, Russian Federation

BGuenine@rupto.ru, https://orcid.org/0000-0003-3514-1340

Dmitry S. Zolkin – Head, Department for Information Retrieval Systems Design, Federal Institute of Industrial Property, Moscow, Russian Federation

db_dept@rupto.ru, https://orcid.org/0009-0002-1641-4518

Igor V. Nekrasov – Researcher, Department for Information Retrieval Systems Design, Federal Institute of Industrial Property, Moscow, Russian Federation

igor.nekrasov@rupto.ru, https://orcid.org/0009-0005-8322-7247