ИНФОРМАТИКА И ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ/INFORMATICS AND INFORMATION PROCESSES

DOI: https://doi.org/10.60797/IRJ.2025.157.2

СЕМАНТИЧЕСКИЕ КЛАСТЕРЫ ПАТЕНТНЫХ ДОКУМЕНТОВ И ГЕНЕРАТОР НАБОРОВ ДАННЫХ ДЛЯ МАШИННОГО ОБУЧЕНИЯ

Научная статья

Горбунов А.В.[1, *], Генин Б.Л.[2], Золкин Д.С.[3], Некрасов И.В.[4]
[1] ORCID : 0009-0001-9503-0880;
[2] ORCID : 0000-0003-3514-1340;
[1, 2, 3, 4] Федеральный институт промышленной собственности, Москва, Российская Федерация

* Корреспондирующий автор (gorbunov[at]rupto.ru)

Аннотация

Современные успехи в развитии методов и средств искусственного интеллекта привели и к новым попыткам создать систему автоматического поиска уровня техники в заданной предметной области. Ключом к успеху здесь является создание и обеспечение доступности наборов данных большого объема для машинного обучения. Также важным для эффективного машинного обучения является определение достаточно просто вычисляемого критерия качества автоматического поиска уровня техники в заданной предметной области.

Настоящая работа посвящена комплексному решению этих двух ключевых задач на основе создания инфраструктуры для исследований в данной области. Предложенная инфраструктура включает средства для формирования и использования наборов данных семантических кластеров патентных документов двух типов — наборы данных для машинного обучения систем патентного поиска и наборы данных для тестирования и оценки качества патентного поиска уровня техники, а также программная утилита оценки качества автоматического патентного поиска.

В статье рассматривается предложенная авторами концепция семантических кластеров патентных документов, определяющих уровень техники в заданной предметной области. Приведено определение таких семантических кластеров. Предложено рассматривать поиск уровня техники как задачу определения элементов семантического кластера патентных документов.

Описан генератор конфигурируемых пользователем наборов данных для машинного обучения на основе коллекции патентных документов США. Генератор датасетов сначала создает базу данных ссылок на документы семантических кластеров. Затем по определенным пользователем параметрам формирует набор размеченных данных для машинного обучения.

Заключительная стадия работы с генератором включает формирование тестового набора данных, предоставляемого для проведения автоматического поиска уровня техники, получение результатов поиска тестируемых систем и вычисление оценок качества поиска с использованием утилиты оценки качества поиска документов уровня техники.

Ключевые слова: патентный поиск, уровень техники, семантический кластер, набор данных, датасет, релевантность, оценка качества поиска, машинное обучение, тестовая коллекция, генератор коллекций.

SEMANTIC CLUSTERS OF PATENT DOCUMENTS AND DATASET GENERATOR FOR MACHINE LEARNING

Research article

Gorbunov A.[1, *], Genin B.L.[2], Zolkin D.S.[3], Nekrasov I.V.[4]
[1] ORCID : 0009-0001-9503-0880;
[2] ORCID : 0000-0003-3514-1340;
[1, 2, 3, 4] Federal Institute of Industrial Property, Moscow, Russian Federation

* Corresponding author (gorbunov[at]rupto.ru)

Abstract

Modern advances in the development of methods and tools of artificial intelligence have also led to new attempts to create a system for automatic search of the state of the art in a given subject area. The key to success here is the creation and availability of large data sets for machine learning. Defining a sufficiently easy to compute quality criterion for automatic prior art search in a given subject area is also important for effective machine learning.

The present work is devoted to a holistic solution to these two key problems by creating an infrastructure for research in this area. The proposed infrastructure includes tools for generating and using datasets of semantic clusters of patent documents of two types — datasets for machine learning of patent search systems and datasets for testing and evaluating the quality of prior art patent search, as well as a software utility for evaluating the quality of automatic patent search.

The article examines the concept of semantic clusters of patent documents defining the state of the art in a given subject area, proposed by the authors. The definition of such semantic clusters is presented. It is suggested to regard the search for the prior art as a task of determining the elements of the semantic cluster of patent documents.

A generator of user-configurable datasets for machine learning based on a collection of US patent documents is described. The dataset generator first creates a database of semantic cluster document references. Then, based on user-defined parameters, it generates a set of marked-up datasets for machine learning.

The final stage of the generator operation involves forming a test dataset provided for performing automated prior art searches, obtaining search results for the systems under test, and calculating search quality scores using a prior art document search quality score utility.

Keywords: patent search, prior art, semantic cluster, dataset, relevance, search quality assessment, machine learning, test collection, collection generator.

Введение

Патентный поиск — это один из самых сложных видов поиска научно-технической информации в очень больших базах данных патентной информации. Поиск в таких базах данных требует специальных знаний и значительных интеллектуальных усилий [1]. В связи с этим актуальной является задача автоматизации патентного поиска для целей экспертизы изобретений или поиска документов уровня техники для данной заявки на изобретение. Во многих современных системах патентного поиска существует функциональность, направленная на решение подобных задач. Такая функциональность часто определяется во многих системах, как поиск документов, «похожих» на заявку на изобретение или, в более общем случае, на документ-образец.

В последнее время в этой области значительные успехи достигнуты с использованием методов искусственного интеллекта. Существенные результаты получены при использовании подходов дистрибутивной семантики — при формировании дистрибутивных тезаурусов и использовании квазисинонимов из дистрибутивного тезауруса вместо классических синонимов [2]. Имеются публикации и о попытках непосредственного использования искусственных нейронных сетей для автоматического поиска [3], [4].

Важно, что оба упомянутых подхода существенно опираются на достаточно большие коллекции размеченных документов, используемые для машинного обучения соответствующих моделей. Формирование указанных коллекций производится преимущественно «вручную», с привлечением квалифицированных экспертов. Это ограничивает разнообразие и объем наборов данных (датасетов) и снижает эффективность обучения и тестирования моделей. Созданию коллекций и применению наборов данных для машинного обучения и тестирования систем патентного поиска посвящен целый ряд публикаций [5], [6], однако до настоящего времени исследователям и создателям систем патентного поиска не было предложено подходов и инструментов, позволяющих автоматизировать процесс подготовки коллекций документов и наборов данных.

Для патентных документов характерна определенная, четко заданная Всемирной Организацией по Интеллектуальной Собственности (ВОИС, WIPO) структура, в которой, в том числе, присутствует поле «Список документов уровня техники» — поле ИНИД (56) в соответствии со стандартом WIPO ST.9 [7]. Это поле заполняется экспертами патентного ведомства при экспертизе заявки на патент и зачастую используется в качестве разметки для обучения систем патентной классификации и поиска [8].

Как и в [8], авторы данной статьи также исходят из предположения, что решением задачи поиска документов уровня техники для рассматриваемой заявки на изобретение является максимизация степени близости результатов автоматического патентного поиска к результатам экспертного поиска, приведенным экспертизой в отчете о поиске по заявке на изобретение и в поле ИНИД (56) [9], [10]. Однако оценка «близости» осложняется тем, что для экспертизы не важно, на какую из публикаций об изобретении приводится ссылка в поле ИНИД (56). Другими словами, для каждого изобретения, оцененного как уровень техники для рассматриваемой заявки, эксперты ведомства указывают лишь один документ из потенциально многих, характеризующих это изобретение. Поэтому при формировании наборов данных для обучения и тестирования систем патентного поиска нужно учитывать и публикации на различных этапах жизненного цикла заявки на изобретение и, что не менее важно, публикации по заявкам семейства патентов-аналогов. Полнота обучающей выборки является критическим условием для успешного решения задач машинного обучения [11]. Следовательно, важным элементом для решения задачи обучения автоматическому поиску уровня техники является формирование полного набора данных, учитывающего все опубликованные документы для каждого документа уровня техники.

В настоящей работе предложен новый подход к созданию коллекций документов для машинного обучения, основанный на объединении документов уровня техники в семантические кластеры, опирающиеся на уже выполненную оценку документов экспертами патентного ведомства. Такой подход к созданию коллекций документов для машинного обучения соответствует задаче обучения нейронных моделей выявлению документов уровня техники, при этом оценка результатов автоматического поиска уровня техники по попаданию в семантический кластер соответствует задачам патентной экспертизы.

Целью настоящей работы являлось создание инфраструктуры, обеспечивающей возможность эффективного проведения исследований в области совершенствования патентного поиска, точнее его разновидности — поиска предшествующего уровня техники, с использованием методов и средств искусственного интеллекта.

В статье описываются практические результаты работ по созданию комплексной инфраструктуры для обучения и тестирования систем патентного поиска: коллекция патентных документов на основе семантических кластеров, генератор наборов данных, программная утилита оценки качества автоматического патентного поиска документов уровня техники.

Основные решения подхода

2.1. Семантические кластеры патентных документов

В патентных исследованиях широко используется понятие «семейство патентов-аналогов». Документы, входящие в семейство патентов-аналогов обычно определяются, как группа документов, относящихся к одному и тому же изобретению и включающая различные публикации на различных этапах жизненного цикла заявки на изобретения и заявки, ссылающиеся на тот же самый пул приоритетных данных. В то же время для задачи поиска документов уровня техники представляет интерес более широкое множество патентных документов по изобретениям, определяющим

уровень техники в некой технической области — области рассматриваемой заявки на изобретение. Для такого множества авторы вводят понятие семантического кластера патентных документов.

Кластеризация давно и эффективно применяются для группировки патентных документов. Алгоритмы кластеризации направлены на организацию данных в группы или кластеры на основе присущих им закономерностей в самих данных [12]. Известен, например, метод кластеризации, позволяющий группировать патентные документы, определяющие связанные ключевые инновации. [13]. Исследования [14] показывают, что эффективность кластеризации сильно зависит от выбора параметров, влияющих на оценку степени принадлежности документа кластеру.

Подход авторов к определению семантических кластеров патентных документов связан не столько с выявлением общности характеристик самих данных, сколько с выявлением влияния этих данных на решение задачи.

Формальное определение семантического кластера патентных документов, предлагаемое авторами, опирается на устойчивые особенности патентных документов, вытекающие из особенностей патентного права.

Семантический кластер патентных документов представляет собой множество патентных документов, включающее:

– базовый патентный документ (базовая публикация), то есть публикация заявки на изобретение или публикация описания изобретения к патенту;

– другие публикации, содержащие полнотекстовую информацию на различных этапах жизненного цикла изобретения; патентные документы семейства патентов — аналогов, в которое входит данная заявка на изобретение (в том числе все публикации, содержащие полнотекстовую информацию на различных этапах жизненного цикла изобретений);

– патентные документы, на которые ссылаются эксперты (цитаты) в публикациях по результатам экспертизы отчетов о поиске по заявке и в поле с кодом ИНИД (56) в публикации о выдаче патента на изобретение;

– патентные документы семейств патентов — аналогов, в которые входят документы, на которые ссылается экспертиза.

Таким образом, семантический кластер состоит из документов, которые, по мнению экспертизы, должны быть учтены при экспертизе рассматриваемой заявки на изобретение. С учетом введенного определения решением задачи автоматического поиска уровня техники для целей экспертизы рассматриваемой заявки на изобретение будем считать автоматический поиск документов, которые входят в соответствующий семантический кластер.

Объединение текстов и библиографических данных патентных документов, входящих в семантические кластеры, позволяет специалистам, организующим машинное обучение, абстрагироваться от специфических особенностей и деталей форматирования патентных документов и сконцентрировать свое внимании на задачах собственно машинного обучения.

2.2. Наборы данных на основе семантических кластеров

Структура набора данных на основе семантических кластеров, представлена на рисунке 1. В качестве примера демонстрируется формирование англоязычного набора данных на основе патентных документов, опубликованных ведомством США по патентам и товарным знакам (USPTO).

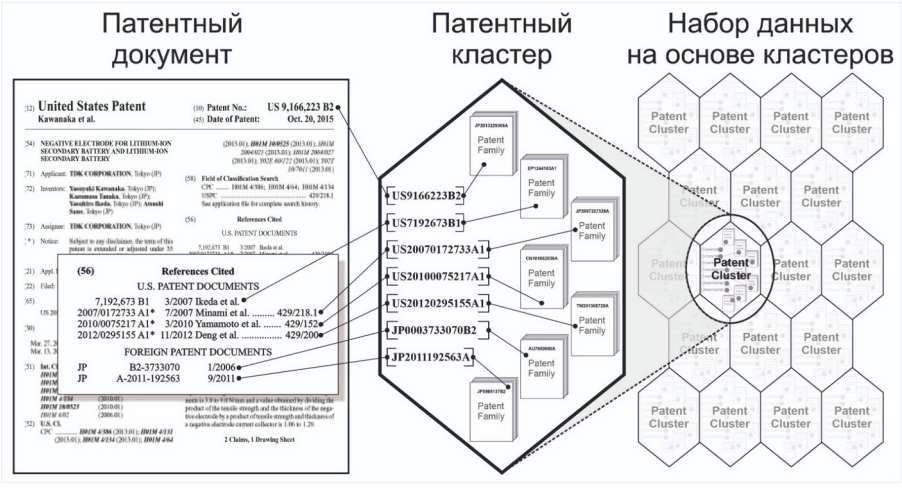

Рисунок 1 - Структура набора данных на основе семантических кластеров патентных документов
DOI: https://doi.org/10.60797/IRJ.2025.157.2.1

Результаты машинного обучения систем автоматического поиска уровня техники для целей экспертизы изобретений могут быть оценены с использованием тестового массива, также представляющего собой набор семантических кластеров, для которых базовый документ представляет собой публикацию патента.

2.3. Критерий качества автоматического патентного поиска

Опираясь на определение семантического кластера, авторы предлагают найденный при автоматическом поиске уровня техники по документу-образцу патентный документ считать релевантным, если он входит в семантический кластер, в состав которого входит этот документ-образец. Учитывая, что в кластер попадают патентные документы,

процитированные экспертизой, и документы из их семейств, оценка релевантности найденных документов заключается в том, чтобы определить, сколько изобретений, указанных в цитатах, найдены в списке результатов поиска (за исключением самого базового документа и его аналогов). Каждое изобретение, включенное в кластер, описывается цитированным документом и его аналогами, назовем эти документы семейством цитаты или группой. В списке результатов поиска может быть один или несколько документов из каждой группы или может не быть ни одного. Если найден хотя бы один документ из группы в списке результатов поиска, то считаем, что найдено одно изобретение.

При вычислении оценки качества поиска используется количество найденных изобретений, то есть количество групп (или семейств) цитат, для которых изобретения были найдены. Это и предлагается считать количеством найденных релевантных документов.

Качество системы патентного поиска уровня техники предлагается оценивать по двум показателям [10], ориентированным на вычисление количества «удачных» поисков системы, т.е. поисков, в которых система находила релевантные документы, принадлежащие семантическому кластеру базового документа (документа-образца). При этом количество документов в выдаче системы ограничивается каким-либо разумным числом. Так, для целей патентного поиска авторы считают достаточным K в пределах от 20 до 50 документов.

Первый показатель — доля выполненных тестовых поисков, для которых в результатах поиска находится не менее одного релевантного документа. Идея показателя схожа с известным показателем hit@K, применяемым для оценки систем, возвращающих несколько ответов. Hit@K — доля запросов в систему, для которых система вернула хотя бы один корректный ответ среди первых K [15], [16].

Таким образом, показатель может быть интерпретирован, как вероятность извлечения системой и представления в top(K) хотя бы одного релевантного документа из коллекции для каждого отдельно взятого выполненного поиска.

Второй показатель — доля выполненных тестовых поисков, для которых в результатах поиска находятся все документы, релевантные запросу. Показатель может быть интерпретирован, как способность системы обеспечивать в первых K документах выдачи 100% полноту для каждого поиска.

Основные результаты

3.1. Генератор наборов данных для машинного обучения

Реализация конфигурируемых по запросам пользователей наборов данных семантических кластеров патентных документов включает базу данных семантических кластеров и идентификаторов входящих в них патентных документов, а также программное обеспечение, предоставляющее API для получения самих документов.

База данных ссылок на семантические кластеры и входящие в них патентные документы создается специализированным блоком программного обеспечения генератора наборов данных семантических кластеров. Этот блок последовательно обрабатывает полнотекстовые публикации патентных документов и формирует соответствующие семантические кластеры. При этом блок при необходимости выполняет стандартизацию данных ссылок экспертизы, а также включает в семантические кластеры сведения о документах, входящих в соответствующие семейства патентов-аналогов.

Семантические кластеры сформированы в виде виртуальных наборов данных, определенных в указанной базе данных, представляющих собой наборы ссылок на входящие в семантические кластеры патентные документы. Для получения наборов данных, нацеленных на исследование многоязычного поиска, в наборы ссылок на документы семантического кластера включаются атрибуты страны публикации документа.

Инфраструктура данных семантических кластеров и системы генерации наборов данных семантических кластеров реализована в виде реляционной базы данных под управлением СУБД PostgreSQL.

Такая структура позволяет использовать для машинного обучения формируемый «на лету» набор данных семантических кластеров.

При этом нужный размер набора данных регулируется выбором диапазона дат публикации. Использование дат публикации для такого регулирования приводит к тому, что тематическое распределение патентных документов в наборе данных в среднем соответствует тематическому распределению изобретательской активности в данной стране.

Для формирования тематических наборов данных можно использовать списки документов, полученные из внешней поисковой системы поиском по классификационным данным или поиском по любым, доступным в этой системе тематическим запросам.

Определение состава семантических кластеров требует использования сведений о семействах патентов-аналогов. В представленной реализации использована база данных DocDB от ЕПВ [17], включающая необходимые данные о семействах патентов-аналогов.

Использование генератора наборов данных большого объема для машинного обучения удобно реализовывать путем on-fly получения семантических кластеров и входящих в них документов. Последовательность шагов по формированию наборов данных представлена на рисунке 2.

На первом шаге исследователь формализует свои пожелания по объему и составу генерируемого набора данных исходя из задач исследования. На этом шаге могут быть определены ограничения по датам публикации базовых документов семантических кластеров. Могут быть определены ограничения и на основе полнотекстового поиска базовых документов семантических кластеров. Причем генератор будет выдавать оценку количества семантических кластеров, соответствующих указанным ограничениям. Генератор может использовать и получаемый извне список патентных документов в качестве базовых документов семантических кластеров.

На втором шаге семантические кластеры для определенных на первом шаге базовых документов и входящие в эти кластеры документы получаются через API генератора путем последовательных обращений вплоть до сообщения о том, что все документы требуемого набора данных возвращены в вызывающую программу. Для удобства использования получаемых документов в системах машинного обучения документы выдаются в формате JSON.

Рисунок 2 - Генерация набора данных на основе семантических кластеров
DOI: https://doi.org/10.60797/IRJ.2025.157.2.2

Таким образом, созданный генератор наборов данных семантических кластеров и сформированная база данных ссылок на патентные документы, входящие в семантические кластеры позволяет через API генератора определять и формировать наборы данных на основе семантических кластеров и выдавать входящие в них патентные документы.

Генератор наборов данных семантических кластеров, база данных ссылок на документы семантических кластеров и API генератора выставлены на поисковой платформе Роспатента вместе с необходимой документацией [18] и предоставляются бесплатно заинтересованным лицам по запросу.

3.2. Утилита оценки качества автоматического поиска уровня техники

Выше отмечалось, что наборы данных на основе семантических кластеров и, в частности, тестовый набор, может использоваться для оценки качества автоматического поиска уровня техники или патентного поиска «похожих» и сравнительного анализа систем патентного поиска.

Подготовлен унифицированный инструмент для вычисления таких оценок — программная утилита оценки качества автоматического патентного поиска похожих (далее — Утилита).

Процедура оценки качества поиска с помощью Утилиты включает три этапа.

На первом этапе Утилита отбирает документы для проведения оценки, либо принимает список идентификаторов документов, подготовленный заранее. Все документы проверяются на наличие активных патентных ссылок, что позволяет вычислять критерий качества поиска, и наличие хотя бы одного из текстовых полей описания изобретения abstract, description, claims (это позволяет проводить полнотекстовый поиск). В результате первого шага формируется список идентификаторов отобранных документов.

На втором этапе по каждому из отобранных документов проводится поиск в одной (или всех) из четырех систем по выбору пользователя: Платформа патентного поиска Роспатента, Google.Patent, Yandex.Patent — через API соответствующих систем; четвертой системой может быть любая произвольная поисковая система, в которую загружается файл, содержащий список идентификаторов отобранных документов. Каждая из тестируемых систем производит поиск в доступных ей массивах патентных документов.

Результаты поиска для указанных трех систем Утилита получает через API, для четвертой системы результаты должны быть загружены из файла с результатами поиска в этой системе.

Результатом работы на этом этапе является массив отсортированных по релевантности списков результатов поиска уровня техники (или поиска похожих) для документов-образцов, отобранных на первом этапе. При этом в список результатов включаются первые K (указывается пользователем, но не более 1000) отсортированных по релевантности результатов поиска (top (K)).

На третьем этапе Утилита вычисляет описанные выше показатели качества поиска уровня техники на основе определения релевантности найденных документов, то есть их принадлежности к соответствующим семантическим кластерам.

Для удобства сравнения получаемых результатов дополнительно реализован расчет еще двух показателей, являющихся усреднением широко используемых оценок качества поиска — точности поиска и полноты поиска:

– точность поиска по одному запросу (precision, P@K) — отношение количества релевантных документов, найденных в результате поиска в первых K документов списка результатов поиска, к общему количеству выданных системой документов (K);

– полнота поиска по одному запросу (recall, R@K) — отношение количества релевантных документов, найденных в результате поиска в первых K документов списка результатов поиска, к общему количеству документов, соответствующих запросу, точнее к количеству учитываемых цитат в поле (56) или в отчете о поиске.

Утилита выдает усредненную точность поиска (mean precision, MP@K) и усредненную полноту поиска (mean recall, MR@K), которые являются арифметическим средним точности и полноты для каждого поиска по множеству поисков.

3.3. Коллекция семантических кластеров патентных документов США

Представленная реализация коллекции семантических кластеров сформирована из общедоступных патентных документов США. В массив включены два типа документов — заявки и патенты. Каждый документ содержит основные библиографические данные и текстовые поля.

В коллекцию включены документы массива патентных документов США с датой публикации с 2000 до конца 2020 года, имеющие код вида документа A1 (заявка — Utility Patent Application published on or after January 2, 2001) или B2 (патент — Utility Patent Grant (with pre-grant publication) issued on or after January 2, 2001). В поисковой платформе Роспатента в массиве документов США по состоянию на дату начала создания коллекции 02.06.2023 всего документов в указанном диапазоне дат — 14 232 161, в том числе документов с кодами вида документа A1 или B2 всего 12 475 835 (из них патентов с кодом вида документа B2 — 4 159 569).

Ниже приведены некоторые характеристики и особенности созданной коллекции семантических кластеров патентных документов США.

Общие характеристики коллекции семантических кластеров приведены в таблице 1. Это количество семантических кластеров, общее количество неуникальных документов во всех кластерах (широко цитируемые документы могут входить в несколько кластеров). Показан характер цитирования патентных документов базового патентного ведомства и других ведомств и среднее количество цитированных патентных документов отдельно для выданных патентов и для заявок, а также использование семейств патентов-аналогов.

Таблица 1 - Общие характеристики коллекции семантических кластеров

DOI: https://doi.org/10.60797/IRJ.2025.157.2.3

№	Показатель	A1	B2	Всего
1	Количество кластеров (записей в БД)	8 316 266	4 159 569	12 475 835
2	Количество неуникальных документов во всех кластерах с учётом базового документа	91 532 079	328 969 976	420 502 055
3	Количество уникальных документов во всех кластерах, включая базовые документы	-	-	26 118 166
4	Количество кластеров, в которые входит только базовый документ	1 450 671	24 019	1 474 690
5	Количество кластеров, в которых есть документы помимо базового	6 865 595	4 135 550	11 001 145
6	Количество цитат всего	12 342 294	85 020 365	97 362 659
7	Количество цитат исходного ведомства	12 305 499	84 506 761	96 812 260
8	Количество	36 795	513 604	550 399

№	Показатель	A1	B2	Всего
	цитат других ведомств			
9	Количество кластеров с цитатами только исходного ведомства (без цитат других ведомств и аналогов базового документа)	1 362 299	3 670 843	5 033 142
10	Количество кластеров с цитатами только других ведомств (без цитат исходного ведомства и аналогов базового документа)	5 315	1 826	7 141
11	Количество кластеров с цитатами как исходного, так и других ведомств (но без аналогов базового документа)	22 714	76 757	99 471
12	Количество кластеров только с аналогами базового документа	6 926 858	399 314	7 326 172
13	Количество кластеров без цитат исходного ведомства	6 931 253	411 969	7 343 222
14	Количество кластеров без цитат других ведомств	8 291 237	4 080 986	12 372 223
15	Среднее количество всех цитат	1,484	20,440	7,804
16	Среднее количество цитат исходного ведомства	1,480	20,316	7,760
17	Среднее количество цитат других ведомств	0,004	0,123	0,044
18	Среднее количество патентов-аналогов базового документа	3,898	4,268	4,022

На рисунке 3 показано распределение количества патентных документов, цитированных один и более раз в других документах. Учитывалось общее количество цитирований для заявок и/или патентов.

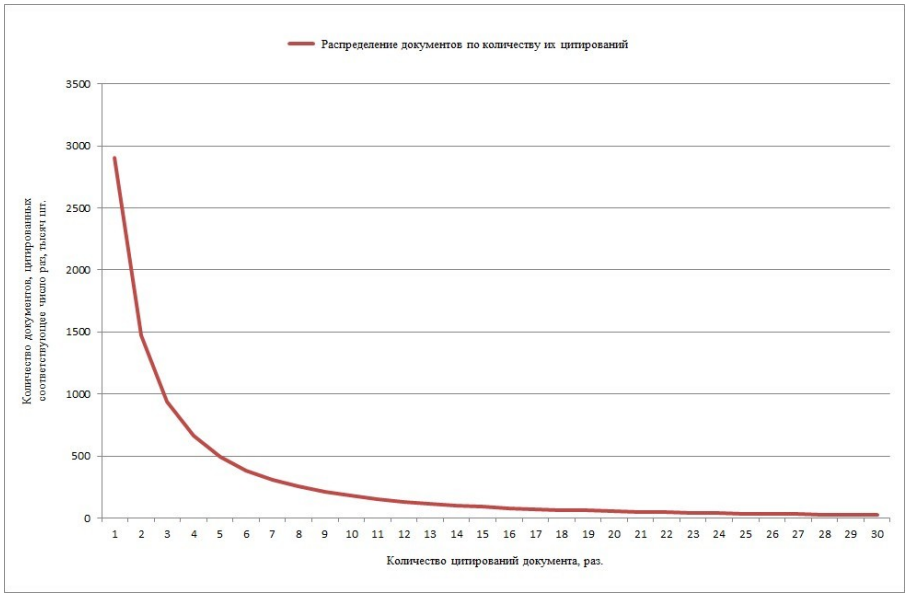

Рисунок 3 - Распределение количества патентных документов, цитированных один и более раз
DOI: https://doi.org/10.60797/IRJ.2025.157.2.4

Отдельно, на рисунке 4 показано распределение количества кластеров с базовым документом с кодом вида B2 по количеству его цитирований.

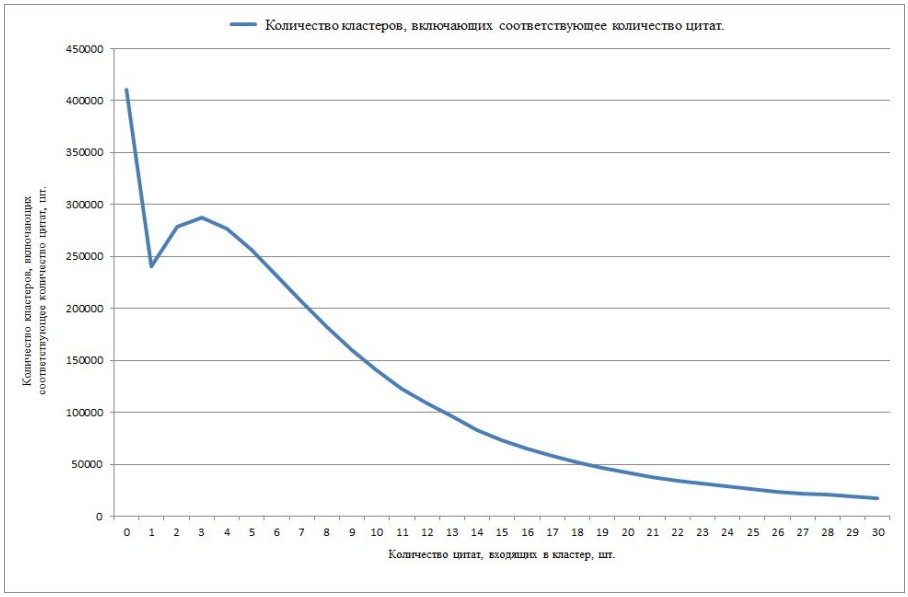

Рисунок 4 - Распределение количества кластеров с базовым документом с кодом вида B2 по количеству его цитирований
DOI: https://doi.org/10.60797/IRJ.2025.157.2.5

В таблице 2 показано распределение цитируемых документов по странам.

Таблица 2 - Распределение цитируемых документов по странам
DOI: https://doi.org/10.60797/IRJ.2025.157.2.6

Ведомство цитируемого документа	Количество цитируемых неуникальных документов
US	96931105
JP	155992

Ведомство цитируемого документа	Количество цитируемых неуникальных документов
WO	87821
EP	45402
CN	43077
KR	28257
CA	24079
DE	20200
RU	6566
CL	6532
TW	6051
AU	4777
GB	378
SE	345
BE	325
FR	213
AT	202
CH	190
ES	13

В таблице 3 показано распределение кластеров по количеству входящих в них документов.

Таблица 3 - Распределение кластеров по количеству входящих в них документов

DOI: https://doi.org/10.60797/IRJ.2025.157.2.7

Количество документов в кластере	Количество кластеров всего	Количество кластеров с базовым A1	Количество кластеров с базовым B2
1	1451135	1448591	2544
2	2618526	2514177	104349
3	1547495	1272907	274588
4	697023	552380	144643
5	531856	366779	165077
6	477485	306253	171232
7	322675	169381	153294
8	309819	159317	150502
9	268019	125349	142670
10	232084	99400	132684
11	212799	88227	124572
12	193804	77536	116628
13	173542	65813	107729
14	160379	59407	100972
15	147284	53446	93838
16	134933	47430	87503
17	126767	44328	82439
18	117083	40296	76787
19	106825	36324	70501
20	100118	33868	66250
21	94470	31688	62782
22	88077	29019	59058
23	82152	27495	54657
24	77162	25225	51937
25	72079	23621	48458

Количество документов в кластере	Количество кластеров всего	Количество кластеров с базовым A1	Количество кластеров с базовым B2
26	68448	22128	46320
27	64676	20934	43742
28	60762	19581	41181
29	57694	18708	38986
30 ... 100	54741 ... 5720	17824 ... 1654	36917 ... 4066

Некоторые из показанных характеристик достаточно интересны. Например из таблицы 3 видно, что наибольшее количество семантических кластеров с базовым патентом содержат всего три патентных документа, но распределение имеет длинный и медленно спадающий «хвост». Из таблицы 1 видно, что публикации заявок в США часто включают цитаты документов уровня техники из отчетов о поиске, обычно отражаемые в поле с кодом ИНИД (56) патента. Таблица 2 демонстрирует, что эксперты USPTO цитируют документы собственного ведомства значительное чаще, чем любые другие. Интересно, что достаточно часто эксперты патентного ведомства США указывают в результатах экспертизы 100 и более патентных документов.

В целом можно отметить, что представленная коллекция семантических кластеров патентных документов США дает возможность проводить машинное обучение систем патентного поиска на достаточно большом объеме размеченных квалифицированными экспертами патентных документов.

Заключение

Рассмотрена задача автоматического поиска уровня техники для целей экспертизы заявок на изобретения.

Предложено считать решением задачи максимизацию степени близости результатов автоматического патентного поиска к результатам экспертного поиска, приведенным экспертизой в отчете о поиске по заявке на изобретение.

Предложено понятие «семантический кластер патентных документов», который должен включать патентные документы с описанием изобретений, определяющих уровень техники в конкретной предметной области.

Описана реализация генератора наборов данных семантических кластеров патентных документов США, предназначенных для машинного обучения и тестирования систем патентного поиска. Реализация содержит базу данных из 12,4 миллиона семантических кластеров, включающих более 120 миллионов неуникальных патентных документов.

Описана утилита оценки качества автоматического патентного поиска документов уровня техники, использующая определяемый исследователем тестовый набор данных семантических кластеров патентных документов.

Созданная инфраструктура для исследований в области автоматического патентного поиска с применением искусственного интеллекта размещена в бесплатном доступе на поисковой платформе Роспатента.

Финансирование

Работа выполнена в рамках НИР 1-ИТ-2022 Федерального института промышленной собственности Роспатента.

Конфликт интересов

Не указан.

Рецензия

Рудой Е.М., ООО «ГК «Иннотех», Москва Российская Федерация
DOI: https://doi.org/10.60797/IRJ.2025.157.2.9
Пикулев А.Н., Казанский национальный исследовательский технический университет им. А.Н. Туполева – КАИ, Казань Российская Федерация
DOI: https://doi.org/10.60797/IRJ.2025.157.2.8
Все статьи проходят рецензирование. Но рецензент или автор статьи предпочли не публиковать рецензию к этой статье в открытом доступе. Рецензия может быть предоставлена компетентным органам по запросу.

Funding

The work was carried out within the framework of R&D 1-IT-2022 of the Federal Institute of Industrial Property of Rospatent.

Conflict of Interest

None declared.

Review

Rudoi E.M., Innotech Group LLC, Moscow Russian Federation
DOI: https://doi.org/10.60797/IRJ.2025.157.2.9
Pikulev A.N., Kazan National Research Technical University named after A.N. Tupolev – KAI, Kazan Russian Federation
DOI: https://doi.org/10.60797/IRJ.2025.157.2.8
All articles are peer-reviewed. But the reviewer or the author of the article chose not to publish a review of this article in the public domain. The review can be provided to the competent authorities upon request.

Список литературы / References

1. Shalaby W. Patent Retrieval: A Literature Review / W. Shalaby, W. Zadrozny // Knowledge and Information Systems. — 2019. — Vol. 61. — № 5–6. — P. 1–24. — DOI: 10.1007/s10115-018-1322-7.

2. Sahlgren M. The Distributional Hypothesis / M. Sahlgren // The Italian Journal of Linguistics. — 2008. — Vol. 20. — P. 33–53. — URL: https://linguistica.sns.it/RdL/20.1/Sahlgren.pdf (accessed: 11.05.2025).

3. Kang I.-S. Cluster-based patent retrieval using international patent classification systeml / I.-S. Kang, S.-H. Na, J. Kim, J.-H. Lee [et al.] // Information Processing and Management. — 2007. — Vol. 43. — № 5. — P. 1173–1182.

4. Vowinckel K. SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search / K. Vowinckel, V.D. Hähnke // World Patent Information. — 2023. — Vol. 73. — № 4. — 102192 p. — DOI: 10.1016/j.wpi.2023.102192.

5. Lupu M. A Horizontal Patent Test Collection / M. Lupu, A. Bampoulidis, L. Papariello // Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. — 2019. — P. 1333–1336. — DOI: 10.1145/3331184.3331346.

6. Горбунов А.В. Задача выявления элементов семантического кластера патентных документов для поиска уровня техники / А.В. Горбунов, Б.Л. Генин, Д.С. Золкин // НТИ. Серия 1: Организация и методика информационной работы. — 2023. — № 8. — С. 27–32. — DOI: 10.36535/0548-0019-2023-08-5. — EDN YSDWZP.

7. WIPO Standard ST.9 Recommendation Concerning Bibliographic Data on and Relating to Patents and SPCS. — 2024. — URL: https://www.wipo.int/documents/d/standards/docs-en-03-09-01.pdf (accessed: 01.07.2024).

8. Acikalin U.U. Patent Search Using Triplet Networks Based Fine-Tuned SciBERT / U.U. Acikalin, M. Kutlu // arXiv. — 2022. — URL: https://arxiv.org/abs/2207.11497 (accessed: 01.07.2024).

9. Genin B.L. Similarity search in patents databases. The evaluations of the search quality / B.L. Genin, D.S. Zolkin // World Patent Information. — 2021. — Vol. 64. — 102022 p. — DOI: 10.1016/j.wpi.2021.102022.

10. Горбунов А.В. Искусственный интеллект в работе патентных ведомств / А.В. Горбунов, Б.Л. Генин, Д.С. Золкин // Информационные ресурсы России. — 2021. — № 3 (181). — С. 18–23. — DOI: 10.46920/0204-3653_2021_03181_18. — EDN SMKYPW.

11. Парасич А.В. Формирование обучающей выборки в задачах машинного обучения. Обзор / А.В. Парасич, В.А. Парасич, И.В. Парасич // Информационно-управляющие системы. — 2021. — № 4 (113). — С. 61–70. — DOI: 10.31799/1684-8853-2021-4-61-70. — EDN SYIIYB.

12. Yin H. A rapid review of clustering algorithms / H. Yin, A. Aryani, S. Petrie [et al.] // arXiv. — 2024. — 2401.07389. — URL: https://arxiv.org/abs/2401.07389 (accessed: 01.07.2024).

13. Yin H. A rapid review of clustering algorithms / H. Yin, A. Aryani, S. Petrie [et al.] // arXiv. — 2024. — URL: https://arxiv.org/abs/2401.07389 (accessed: 01.07.2024).

14. Trappey C.V. Clustering patents using non-exhaustive overlaps / C.V. Trappey, A.J.C. Trappey, C.-Y. Wu // Journal of Systems Science and Systems Engineering. — 2010. — Vol. 19. — P. 162–181. — DOI: 10.1007/s11518-010-5134-x.

15. Kukolj D. Comparison of Algorithms for Patent Documents Clusterization / D. Kukolj, Z. Tekic, Lj. Nikolic [et al.] // Proceedings of the 35th International Convention MIPRO. — 2012. — P. 1176–1178.

16. Frome A. DeViSE: A Deep Visual-Semantic Embedding Model / A. Frome, G.S. Corrado, J. Shlens [et al.] // Advances in Neural Information Processing Systems 26 (NIPS 2013). — 2013. — P. 2121–2129.

17. EPO worldwide bibliographic data (DOCDB). — URL: https://www.epo.org/en/searching-for-patents/data/bulk-data-sets/docdb (accessed: 01.07.2024).

18. Федеральная служба по интеллектуальной собственности (Роспатент). Платформа поиска патентов. — URL: https://searchplatform.rospatent.gov.ru/ (дата обращения: 01.07.2024).

Список литературы на английском языке / References in English

1. Shalaby W. Patent Retrieval: A Literature Review / W. Shalaby, W. Zadrozny // Knowledge and Information Systems. — 2019. — Vol. 61. — № 5–6. — P. 1–24. — DOI: 10.1007/s10115-018-1322-7.

2. Sahlgren M. The Distributional Hypothesis / M. Sahlgren // The Italian Journal of Linguistics. — 2008. — Vol. 20. — P. 33–53. — URL: https://linguistica.sns.it/RdL/20.1/Sahlgren.pdf (accessed: 11.05.2025).

3. Kang I.-S. Cluster-based patent retrieval using international patent classification systeml / I.-S. Kang, S.-H. Na, J. Kim, J.-H. Lee [et al.] // Information Processing and Management. — 2007. — Vol. 43. — № 5. — P. 1173–1182.

4. Vowinckel K. SEARCHFORMER: Semantic patent embeddings by siamese transformers for prior art search / K. Vowinckel, V.D. Hähnke // World Patent Information. — 2023. — Vol. 73. — № 4. — 102192 p. — DOI: 10.1016/j.wpi.2023.102192.

5. Lupu M. A Horizontal Patent Test Collection / M. Lupu, A. Bampoulidis, L. Papariello // Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval. — 2019. — P. 1333–1336. — DOI: 10.1145/3331184.3331346.

6. Gorbunov A.V. Zadacha vyyavleniya elementov semanticheskogo klastera patentnyh dokumentov dlya poiska urovnya tekhniki [Prior art search as a problem of determining the elements of the semantic cluster of patent documents] / A.V. Gorbunov, B.L. Genin, D.S. Zolkin // NTI. Seriya 1: Organizaciya i metodika informacionnoj raboty [STI. Series 1: Organization and Methods of Information Work]. — 2023. — № 8. — P. 27–32. — DOI: 10.36535/0548-0019-2023-08-5. — EDN YSDWZP. [in Russian]

7. WIPO Standard ST.9 Recommendation Concerning Bibliographic Data on and Relating to Patents and SPCS. — 2024. — URL: https://www.wipo.int/documents/d/standards/docs-en-03-09-01.pdf (accessed: 01.07.2024).

8. Acikalin U.U. Patent Search Using Triplet Networks Based Fine-Tuned SciBERT / U.U. Acikalin, M. Kutlu // arXiv. — 2022. — URL: https://arxiv.org/abs/2207.11497 (accessed: 01.07.2024).

9. Genin B.L. Similarity search in patents databases. The evaluations of the search quality / B.L. Genin, D.S. Zolkin // World Patent Information. — 2021. — Vol. 64. — 102022 p. — DOI: 10.1016/j.wpi.2021.102022.

10. Gorbunov A.V. Iskusstvennyj intellekt v rabote patentnyh vedomstv [Artificial intelligence in patent office's procedures] / A.V. Gorbunov, B.L. Genin, D.S. Zolkin // Informacionnye resursy Rossii [Information Resources of Russia]. — 2021. — № 3 (181). — P. 18–23. — DOI: 10.46920/0204-3653_2021_03181_18. — EDN SMKYPW. [in Russian]

11. Parasich A.V. Formirovanie obuchayushchej vyborki v zadachah mashinnogo obucheniya. Obzor [Training set formation in machine learning tasks. Survey] / A.V. Parasich, V.A. Parasich, I.V. Parasich // Informacionno-upravlyayushchie sistemy [Information and Control Systems]. — 2021. — № 4 (113). — P. 61–70. — EDN SYIIYB. [in Russian]

12. Yin H. A rapid review of clustering algorithms / H. Yin, A. Aryani, S. Petrie [et al.] // arXiv. — 2024. — 2401.07389. — URL: https://arxiv.org/abs/2401.07389 (accessed: 01.07.2024).

13. Yin H. A rapid review of clustering algorithms / H. Yin, A. Aryani, S. Petrie [et al.] // arXiv. — 2024. — URL: https://arxiv.org/abs/2401.07389 (accessed: 01.07.2024).

14. Trappey C.V. Clustering patents using non-exhaustive overlaps / C.V. Trappey, A.J.C. Trappey, C.-Y. Wu // Journal of Systems Science and Systems Engineering. — 2010. — Vol. 19. — P. 162–181. — DOI: 10.1007/s11518-010-5134-x.

15. Kukolj D. Comparison of Algorithms for Patent Documents Clusterization / D. Kukolj, Z. Tekic, Lj. Nikolic [et al.] // Proceedings of the 35th International Convention MIPRO. — 2012. — P. 1176–1178.

16. Frome A. DeViSE: A Deep Visual-Semantic Embedding Model / A. Frome, G.S. Corrado, J. Shlens [et al.] // Advances in Neural Information Processing Systems 26 (NIPS 2013). — 2013. — P. 2121–2129.

17. EPO worldwide bibliographic data (DOCDB). — URL: https://www.epo.org/en/searching-for-patents/data/bulk-data-sets/docdb (accessed: 01.07.2024).

18. Federalnaya sluzhba po intellektualnoy sobstvennosti (Rospatent). Platforma poiska patentov [Federal Service for Intellectual Property (Rospatent). Patent Search Platform]. — URL: https://searchplatform.rospatent.gov.ru/ (accessed: 01.07.2024). [in Russian]