Научная статья
УДК 347.77
doi: 10.17586/2713-1874-2022-2-36-47

МЕТОДЫ И МОДЕЛИ МОНИТОРИНГА РАЗВИТИЯ ТЕХНОЛОГИЙ ПОЛУЧЕНИЯ ВОДОРОДА И СОПУТСТВУЮЩЕЙ ДЕКАРБОНИЗАЦИИ НА ОСНОВЕ АНАЛИЗА СВЕРХБОЛЬШИХ ПАТЕНТНЫХ КОЛЛЕКЦИЙ

Валерий Олегович Ена[1]✉, Федор Александрович Батанов[2]

[1,2]Проектный офис Федерального института промышленной собственности, г. Москва, Россия
[1]valeriy.ena@rupto.ru ✉, https://orcid.org/0000-0002-1271-5316
[2]batanov@rupto.ru, https://orcid.org/0000-0002-7547-8303
Язык статьи – русский

Аннотация: В условиях экстремально высоких темпов смены технологий, появления новых продуктовых групп и цепочек добавленной стоимости всё больше растет спрос на услуги качественного анализа патентной и другой научно-технической информации. Патентная аналитика должна быть всеобъемлющей, должна применять универсальные подходы вне зависимости от сектора экономики и предоставлять практические рекомендации в максимально сжатые сроки. Интерес к патентной аналитике в контексте задач управления наукой, технологиями и инновациями на уровне государства выражается в существенном расширении спектра исследований и значительном повышении размеров патентных коллекций, подвергаемых анализу, что может выразиться в росте трудозатрат на сбор, анализ и интерпретацию результатов патентной аналитики. Настоящая статья посвящена особенностям оптимизации работы с так называемыми сверхбольшими коллекциями (свыше шести тысяч документов) и предлагает адаптацию традиционного подхода к построению патентного ландшафта.

Ключевые слова: интеллектуальный анализ данных, коммерциализация технологий, патентная аналитика, патентная технологическая разведка, патентный ландшафт, скаутинг технологий, стратегия развития, управление наукой, трансфер технологий, управление технологиями

НИР проведена во исполнение тематического плана научно-исследовательских работ Федерального института промышленной собственности (ФИПС), выполняемых за счет средств от приносящей доход деятельности, на 2021–2023 годы и организации выполнения научно-исследовательских работ, утвержденными приказом ФИПС № 494 от 23.11.2021 г.

Ссылка для цитирования: Ена В.О., Батанов Ф.А. Методы и модели мониторинга развития технологий получения водорода и сопутствующей декарбонизации на основе анализа сверхбольших патентных коллекций // Экономика. Право. Инновации. 2022. № 2. С. 36–47. http://dx.doi.org/10.17586/2713-1874-2022-2-36-47.

METHODS AND MODELS FOR MONITORING THE DEVELOPMENT OF HYDROGEN PRODUCTION TECHNOLOGIES AND ASSOCIATED DECARBONIZATION BASED ON THE ANALYSIS OF ULTRA-LARGE PATENT COLLECTIONS

Valeriy O. Ena[1]✉, Fedor A. Batanov[2]

[1,2]Project Office of the Federal Institute of Industrial Property, Moscow, Russia
[1]valeriy.ena@rupto.ru ✉, https://orcid.org/0000-0002-1271-5316
[2]batanov@rupto.ru, https://orcid.org/0000-0002-7547-8303
Article in Russian

Abstract: In conditions of rapid technology changes rates, constant emergence of new product groups and value chains, the demand for qualitative analysis of patent and other scientific and technical information is growing more and more. Patent analytics in that regard should be comprehensive, apply universal approaches regardless of the subject and provide practical recommendations in the shortest possible time. Interest in patent analytics in the context of the tasks of managing science, technology and innovation at the state level is expressed in a wider range of research. Following significant increase in the size of patent collections subjected to analysis, can result in the overall quality drop of work and an increase in labor costs for collection, analysis and interpretation of patent data. This article is devoted to the features of optimizing work with the so-called. super-large collections (over 6,000 patent documents), and offers an adaptation of the traditional approach to building a patent landscape.

Keywords: data mining, development strategy, patent analytics, patent landscape, science management, technology commercialization, technology management, technology scouting, technology transfer

The research was carried out in accordance to the thematic plan of research works of the Federal Institute of Industrial Property (FIPS at the expense of funds from income-generating activities, for 2021-2023 and the organization of research work, approved by the order of FIPS No. 494 of 23.11.2021.

For citation: Ena V.O., Batanov F.A. Methods and Models for Monitoring the Development of Hydrogen Production Technologies and Associated Decarbonization Based on the Analysis of Ultra-Large Patent Collections. *Ekonomika. Pravo. Innovacii.* 2022. No. 2. pp. 36–47. (In Russ.). http://dx.doi.org/10.17586/2713-1874-2022-2-36-47.

Введение. В современном мире технологии меняются и совершенствуются всё быстрее, объем информации при этом растет такими темпами, что традиционные подходы к обработке информации зачастую перестают быть эффективными. С одной стороны информация, на основании которой делается анализ, должна быть объективной и обладать практической ценностью, а с другой – ответы на все новые вызовы, которые встают перед лицами, принимающими решения, нужно получать в относительно разумные сроки.

Один из возможных путей решения вышеописанных проблем – использование патентной информации. Обоснованность использования патентных баз как источника информации определяется следующими свойствами:

1. Патентная информация содержит ценную научно-техническую информацию практического характера, причем значительная ее часть не встречается в других источниках [1].

2. Техническое раскрытие информации на уровне, достаточном для воспроизведения специалистом описанного в патенте изобретения – необходимое условие получения патента (специалист в данной области мог его осуществить), поэтому, как правило, сведения в патентах более полные и подробные, чем в других видах источников.

3. Патенты выдаются после детальной технической экспертизы, таким образом, почти каждый патентный документ верифицируется на уровне государства.

4. Патентная информация охватывает все области техники и структурирована с помощью классификаторов (МПК, СПК и т.д.).

5. Получение патентов связано с серьезными затратами, поэтому выданный патент – это не только технически ценная информация, но и индикатор бизнес-намерений компаний.

Патентная информация широко используется как для анализа технологических трендов, так и для патентных исследований при патентовании изобретения. В первом случае создаются патентные ландшафты, которые показывают тенденции развития области техники [2]. В подобных исследованиях в качестве основы для анализа могут использоваться коллекции патентных документов от нескольких десятков тысяч до нескольких сотен тысяч патентных документов [3].

Однако в данном подходе есть существенный недостаток: поскольку подбор документов ведется автоматически, то далеко не все документы релевантны тематике исследования, что снижает практическую ценность подобных исследований [4]. Во втором случае проводится экспертный отбор патентных документов, что исключает учет нерелевантных документов, однако область поиска при этом очень узкая, и коллекции составляют от нескольких десятков до нескольких сотен патентных документов, при этом область техники «в целом» не рассматривается, что существенно снижает бизнес-ценность исследований.

Отраслевой патентный ландшафт объединяет преимущества и нивелирует недостатки двух вышеописанных подходов, позволяя показать объективную картину области техники и включая релевантную коллекцию патентных документов. Это достигается комбинацией сложной поисковой стратегии и экспертным просмотром всех патентных документов [5]. У данного подхода есть ограничение в 4 000 – 6 000 патентных семейств (совокупности всех патентных документов, имеющих отношение к одному техническому решению), связанное с экспертной частью работы.

Однако существует много областей техники, которые содержат более 6 000 патентных семейств, что затрудняет их детальное изучение с точки зрения патентной информации.

Цель и гипотеза исследования – предоставить рекомендации на основе практики и экспертизы ПО ФИПС по работе с коллекциями больше 6 000 патентных семейств и сохранения оптимальных показателей релевантности и полноты.

Проблематика и современное состояние вопроса. Традиционный отраслевой патентный ландшафт состоит из нескольких этапов, которые можно свести к основным четырем:

1. Исследование предметной области, составление модели предметной области и поисковой стратегии.

2. Поиск и сбор коллекции патентных документов.

3. Ручной просмотр коллекции и соотнесение патентных документов коллекции с одним или несколькими элементами модели предметной области (далее «тегирование»).

4. Анализ полученных итоговых коллекций, выводы и рекомендации на основе выявленных тенденций.

С точки зрения ограничений по числу документов, «узким горлышком» исследования, которое ограничивает максимальное число документов, является третий этап. При проведении исследования все патентные документы, относящиеся к коллекции, сформированной на первом и втором этапах, просматриваются и соотносятся с более узкими тематиками, входящими в предметную область. Для этого в специализированной среде, либо с помощью программы Excel создается матрица, где в столбце представлены все патентные документы, а в строчке элементы модели предметной области (Рисунок 1).

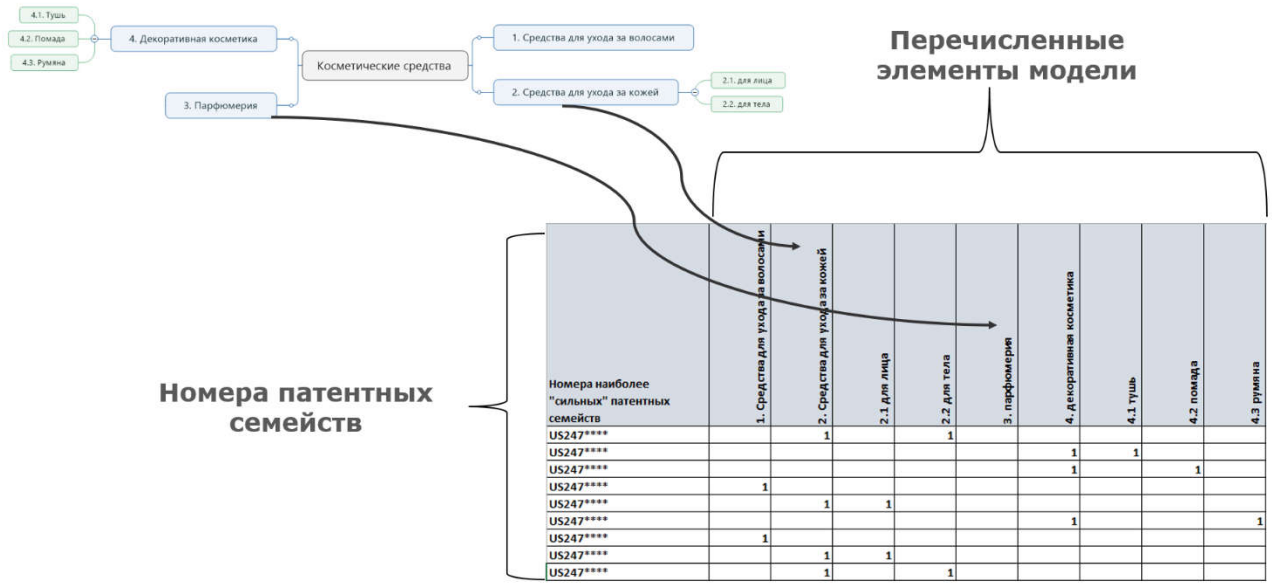

Рисунок 1 – Пример построения матрицы тегирования

Эксперты вручную заполняют данную таблицу после прочтения патентных документов, тем самым формируя базис для последующего многомерного анализа [3]. При решении подобных задач важно обеспечить единый подход экспертов к оценке патентных документов, общую рабочую среду, контроль качества и приемлемые сроки, которые позволяют итоговому исследованию оставаться актуальным в течение продолжительного времени.

Такие подходы подразумевают необходимость ручной чистки документов и их тегирования. В связи с тем, что это достаточно трудоемкий процесс, которому еще не было представлено автоматизированных альтернатив с достаточно высокой степенью качества проработки, полностью заменить экспертную оценку документов патентными аналитиками не представляется возможным. Особенно это касается тематических областей, где наполнение самих документов представлено достаточно неоднородно, чтобы однозначно применять одни и те же критерии отбора документов сразу ко всей коллекции.

Все вышеперечисленные факторы сказываются на общей продолжительности процесса, в связи с чем существует определенный количественный порог по числу документов в общей коллекции, подлежащей обработке и дальнейшему анализу. Как показывает практика Проектного офиса ФИПС, такой пороговой точкой является лимит в 4–6 тысяч патентных документов, что по большей части не вызывает затруднений при проведении исследований на хронологическом горизонте 10–20 лет.

Тем не менее, проблемы начинают возникать в случаях, когда количество найденных патентных документов (до стадии очистки, а в некоторых случаях и уже вся генерализованная коллекция) превышает 6 тысяч документов (так называемые сверхбольшие коллекции). В таких ситуациях описанные выше подходы не позволяют провести все этапы исследования в полной мере, а где-то это не представляется возможным в принципе.

При этом текущее состояние развития области патентной аналитики, все большее внимание и внедрение данного направления в сторону управления наукой, технологиями и инновациями на уровне государства указывает на смещение фокуса исследований на более глобальные и широкие тематики, чем при разработке традиционных патентных ландшафтов. В сложившихся условиях и проведении исследований на макроуровне, можно прогнозировать значительное увеличение сверхбольших коллекций, подлежащих анализу.

Тем не менее, проведение работы в традиционной практике создания патентных ландшафтов в такой ситуации не представляется возможным, так как ручной анализ такого масштаба не целесообразен с точки зрения трудовых ресурсов и затрачиваемого времени.

Вызвано это фактом наличия ощутимого временного лага, негативно влияющего на ценность и практическую актуальность работы по мере роста потраченного на него времени, так как выделенные в нем выводы применимы только на момент формирования коллекции.

В связи с этим, задача качественной и своевременной обработки сверхбольших коллекций патентных документов с минимизацией доли нерелевантных документов является в высшей степени актуальной. Подход, описанный в данной статье, применялся на реальном проекте и при определенных условиях может быть использован универсально.

Методология.

Подход к оптимизации работ со сверхбольшими коллекциями.

В рамках данного исследования рассматриваются изменения, которые можно предпринять при анализе сверхбольших коллекций – коллекций патентных документов большого размера (более 6 тысяч патентов). Почти все эти изменения носят методологический характер и относятся к стадиям предварительной очистки коллекции и последующего тегирования перед стадией проведения непосредственного патентного анализа.

После традиционных подходов к проведению поисков по общему запросу/-ам образуется первичная (до чистки) генерализованная коллекция, которая уже может подлежать обработке [4]. Тем не менее, перед тем как переходить к тегированию документов по элементам модели, необходимо проведение хотя бы первичной чистки. Вызвано это потенциальной возможностью сохранения нерелевантных документов в коллекции даже в случае применения специальных техник тегирования по отношению к сверхбольшим коллекциям.

При традиционном подходе разработки патентных ландшафтов стадии чистки и тегирования выполняются параллельно в рамках одного масштабного процесса. В отношении сверхбольших коллекций целесообразно разделить этапы чистки и тегирования в две отдельные независимые друг от друга группы задач.

В целях корректной работы предложенных ниже алгоритмов тегирования необходима хотя бы первичная экспертная чистка нерелевантных документов. При этом в отличие от тегирования, в случае с чисткой не получится полностью избежать ручного анализа документов экспертной группой. Представленным компромиссом, позволяющим соблюсти необходимый баланс между качеством работ и трудозатратами, было принято

решение ограничиться первичным поиском через определенные поля документов (в подобных случаях целесообразно остановиться только на полях названия и реферата документа).

Также в целях существенного сокращения времени работы над созданием патентного ландшафта, в соответствии с настоящим исследованием предлагается значительно упростить наиболее трудозатратный процесс – тегирование документов. В традиционном подходе соотнесение патентных семейств с элементами модели предметной области осуществляется полностью вручную экспертной группой патентных аналитиков на основании всех полей документа [4]. Данный этап является самым продолжительным по времени, а при условиях работы со сверхбольшими коллекциями проведение тегирования в привычной форме не представляется возможным.

В рамках проведения настоящего исследования предлагается максимально оптимизировать процесс тегирования путем соотнесения патентных семейств коллекции с каждым элементом модели по следующей схеме (Рисунок 2):

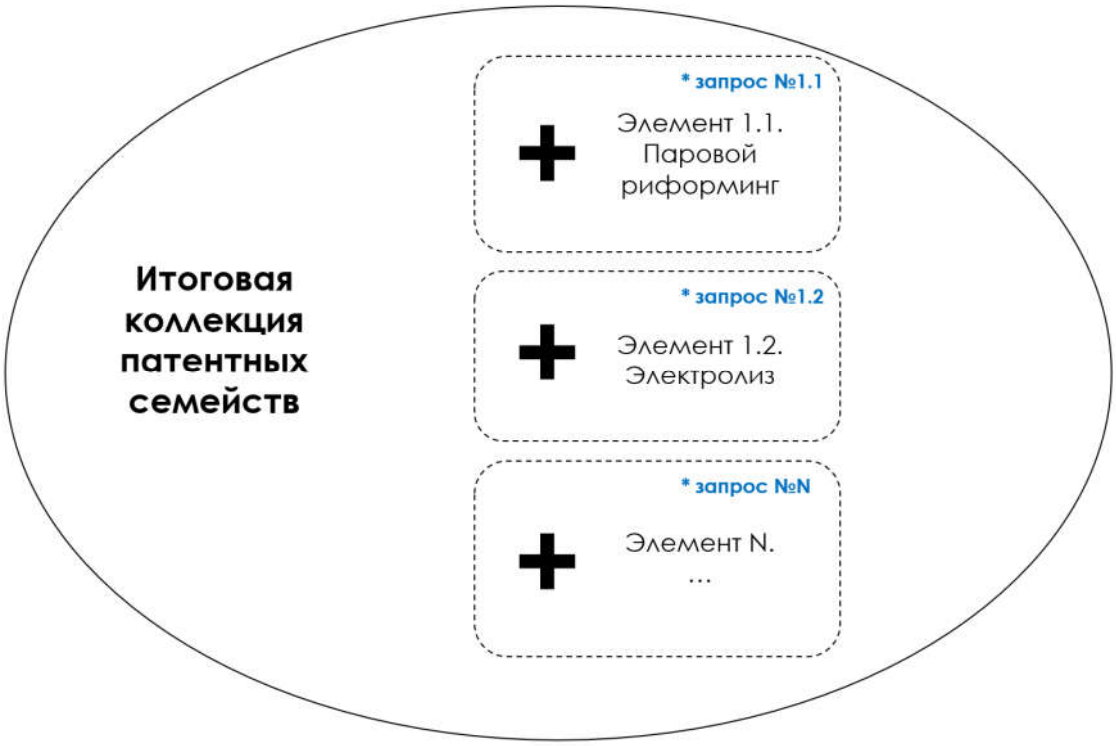

Рисунок 2 – Пример построения расшивающих запросов по элементам модели предметной области

Авторами статьи принимается во внимание, что введение принципиально новых алгоритмов работы с патентной коллекцией может нести в себе возможность допущения методологических ошибок. В связи с этим, одной из задач настоящей статьи является описание научного подхода к валидации обоих вышеописанных нововведений – подхода к чистке и тегированию сверхбольших патентных коллекций. Данный подход представлен в следующем подразделе.

Подход к валидации разработанного подхода.

При подборе необходимых вариантов апробации к настоящей статье учитывался тот факт, что патентная документация представляет собой достаточно трудный для анализа массив данных, особенно в условиях общей неоднородности данных. Соответственно, лучшим механизмом валидации о том, является ли тот или иной документ подходящим для вхождения в отчет по-

прежнему остается экспертная валидация его релевантности.

При этом данное экспертное рассмотрение не ограничивается определенными полями документа, а учитывает всю информацию в патенте в совокупности в той мере, которая позволяет сформировать о нем четкое суждение.

Тем не менее, необходимо учитывать, что такой процесс предполагает довольно тщательное ознакомление с каждым документом: в среднем полноценный анализ шести патентных семейств занимает один час работы аналитика.

При сохранении примерно тех же ограничений по времени и отсутствии в полной мере корректно работающих алгоритмов автоматизации всё острее встает проблема патентной аналитики на сверхбольших коллекциях.

Несмотря на очень глубокую проработку разных аспектов работы с данными в научной литературе в отношении патентной специфики [6, 7], проблематика масштабной обработки сверхбольших коллекций и работа с ними в рамках патентных ландшафтов освещена не так широко. Однако решение подобных проблем путем применения различных статистических методов анализа, в частности, достаточно широко освещено в ряде областей: от проведения контент-анализа в медиа-исследованиях [8] до таких профессиональных направлений, как бухгалтерский учет и аудит [9, 10].

Среди представленных областей деятельности наибольший интерес вызывает именно сфера аудита и консалтинга, стоящая на стыке с патентной аналитикой (являющейся, своего рода, научно-техническим консалтингом). И в том, и в другом роде деятельности осуществляется анализ и построение выводов на основе формализованной документации, содержащей обилие потенциально важной информации. При этом, как правило, аудит менее ограничен тематическими и временными рамками, вследствие чего анализу должно подлежать гораздо большее число самой различной документации в весьма ограниченных временных рамках.

В связи с этим, в профильной научной и профессиональной литературе уже достаточно давно широкое распространение получило упрощение рабочих процессов с помощью анализа не на всей коллекции документов, а на определенной случайной выборке в том масштабе, который необходим для проверки больших массивов документации и связанных с ней гипотез [9, 11, 12].

Целью настоящей валидации является нахождение зоны пересечения грамотного соотношения трудозатрат и приемлемой для анализа доли релевантных документов. Поэтому за основу к осуществлению анализа работы был взят подход, предложенный польским исследователем Янушем Вывилом [9], в котором методология подобных исследований представлена максимально широко. Далее будет рассмотрено применение данного подхода с определенной его доработкой в отношении работы с патентными данными, а также с использованием методологии в привязке непосредственно к патентным исследованиям [13].

В рамках данной валидации будут отдельно рассчитаны три показателя: один (релевантность коллекции) в привязке к качеству указанного подхода к чистке коллекции и два других (точность и глубина поиска) в привязке к качеству отнесения патентных семейств к тому или иному элементу модели предметной области путем проведения поисковых запросов.

Релевантность коллекции.

Общая итоговая коллекция представляет собой множество $U = (1, 2, \ldots, k, \ldots, N)$. Индекс k в этом случае относится к порядковому номеру объекта исследования, который представляет собой определенное патентное семейство. При этом размер N множества U является фиксированным, так как известно точное число документов в коллекции.

Также предположим, что z_k является фиксированной переменной, характеризующей элемент k, где $k \in U$. Установим, что $z_k = 0$, когда в элементе k наблюдается ошибка (техническое решение не является релевантным области) и $z_k = 1$, когда в элементе k ошибок нет (семейство является релевантным). В соответствии с этим, множество U делится на два подмножества $U_0 = \{k: z_k = 0, k \in U\}$ и $U_1 = \{k: z_k = 1, k \in U\}$.

Основной мерой качества обработки коллекции патентных данных в таком случае

является общая доля ошибок – e_k (соотношения нерелевантных документов к общему числу документов в выборке). Данный показатель высчитывается за счет трех других параметров: x_k, y_k и d_k, где x_k является общим числом учтённых элементов (общим числом рассматриваемых патентных семейств в выборке), y_k является числом скорректированных элементов в выборке (со встретившимися ошибками), а показатель общего числа ошибок (d_k) высчитывается путем разницы между показателями x_k и y_k ($x_k - y_k$).

Общая доля ошибок ek в таком случае высчитывается через формулу $e_k = d_k/x_k$ при условии, что $x_k \neq 0$ [9]. При этом, учитывая общий масштаб коллекции, e_k будет рассмотрен на основании сэмплирования, в рамках которого выборки будут сформированы путем систематического подхода [9, 14]. Тем не менее, также необходимо уточнить, что систематическое случайное сэмплирование предполагает наличие определенного интервала. В целях проведения более качественного анализа данный интервал определяется экспертным образом путем итерационного апробирования коллекциями выборок, сформированных на основе 4 различных перцентильных шагов.

Точность и глубина поиска.

Найденный перцентильный интервал впоследствии можно использовать без существенных потерь качества при анализе подобных сверхбольших коллекций. Следующая стадия валидации заключается в расчете показателей точности и полноты поиска для каждого из расширяющих запросов по элементам модели. В соответствии с выделенной в отчете о патентных исследованиях университетом ТвГТУ методологией расчета данных параметров, оба данных показателя определены по данным формулам [13]:

$P_i = R_{ai} / A_i$,

где P_i – точность поиска (precision) для запроса; R_{ai} – количество отобранных (правильных) документов по каждому поисковому запросу; A_i – количество найденных (общее число) документов на каждый поисковый запрос.

Отметим, что в отличие от показателя e_k, показатель Pi строится не на основе числа ошибок, а, наоборот, на основе общего числа верных патентных семейств для каждого запроса. Следующая формула относится к вычислению показателя полноты поиска и представлена ниже:

$R_i = R_{ai} / D_i$,

где R_i – глубина (полнота/recall) поиска для запроса; R_{ai} – количество отобранных (правильных) документов по каждому поисковому запросу;

D_i– общее число найденных документов (вся коллекция).

В отношении настоящего исследования данный показатель носит скорее вспомогательный характер, так как численное отношение запроса к общему числу документов в данном случае не является каким-либо систематическим отражением реальной ситуации и меняется от случая к случаю. В то же время, данный показатель может быть полезен при оценке влияния того или иного элемента модели предметной области на общую коллекцию.

Практические результаты: валидация выбранного подхода. В данном подразделе представлен механизм валидации, апробированный в рамках реального кейса патентного ландшафта Проектного офиса ФИПС в отношении технологий производства водорода без сопутствующих выбросов CO/CO_2. Объём первоначальной коллекции после поиска составлял 9 884 патентных семейства и 5 262 документа после проведения первичной чистки, что характеризует данный массив документов как сверхбольшую коллекцию.

Размер перцентильных шагов выбран экспертно на основании практического опыта при работе с патентными коллекциями и составляет: 1%, 2%, 5% и 10% от всех документов сверхбольшой коллекции (Рисунок 3). При этом очевидно, что количество трудозатрат растет по мере роста шага в большую сторону: другими словами, анализ каждого 1% коллекции занимает гораздо меньше человеко-часов, чем 10% коллекции. В то же время, по мере роста шага растет и качество проверки коллекции: то есть, на шаге в 10%, скорее всего, будет найдено больше нерелевантных документов, чем на шаге в 1%.

В этой связи, в рамках настоящего подхода выбран шаг в 10% как максимальное пороговое значение (с точки зрения трудозатрат) по качеству вычисления e_k в сверх

больших коллекциях. После чего из показателя e_k для шага в 10% вычитается e_k для каждого из трех других шагов. Данная разница (Δe_k) в таком случае отражает, по своей сути, долю упущенных из внимания нерелевантных документов по мере уменьшения шага за счёт уменьшения самой выборки. При этом Δe_k для шага в 10% приравнивается к нулю как референсное значение (квазиноль), с которым сравнивается качество проработки на остальных шагах.

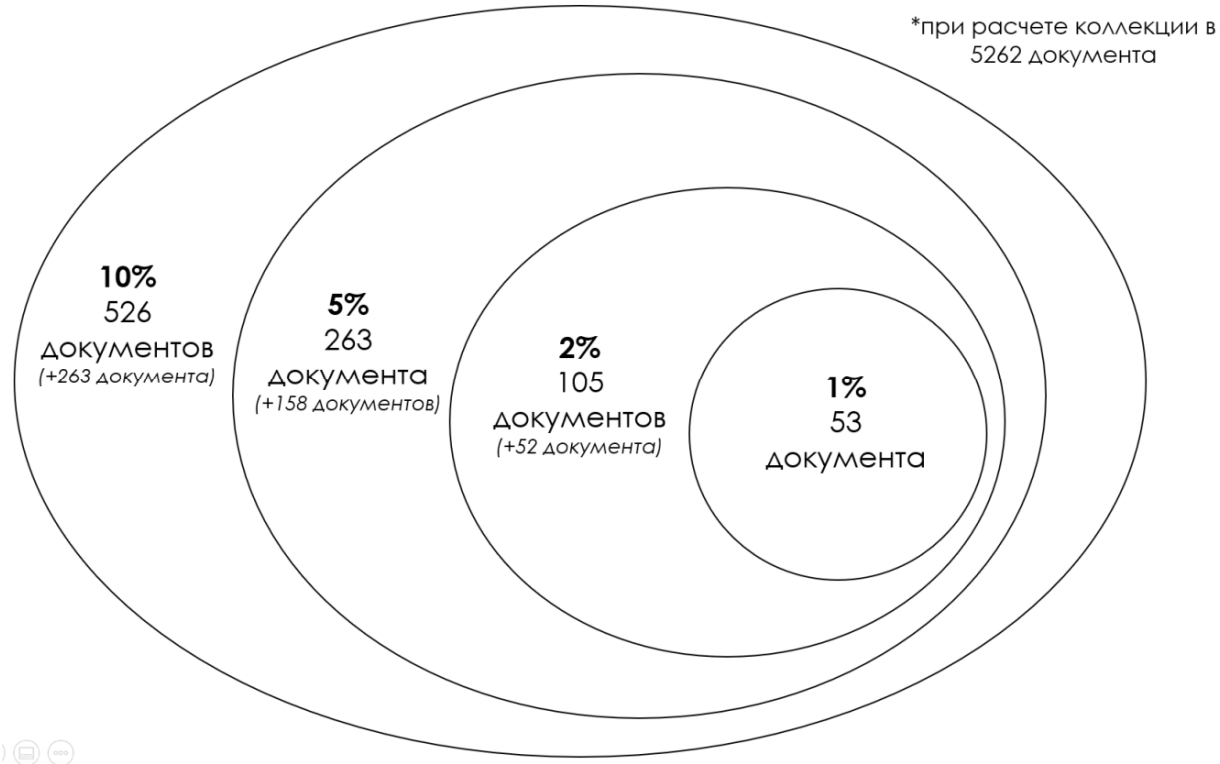

Рисунок 3 – Распределение коллекции по интервалу шага семплирования на примере рассмотренного кейса

После последовательного расчета Δe_k по всей коллекции по каждому из данных, этот показатель будет соотнесен с графиком роста трудозатрат, в результате чего экспертно определяется наиболее оптимальный интервал по соотношению доли ошибок (качества чистки коллекции) и предпринятых на анализ данного показателя трудозатрат.

Этот подход был апробирован на сверхбольшой коллекции. Основные показатели по каждой из выборок представлены в таблице ниже.

Таблица 1

Расчет основных показателей валидации чистки коллекции

Шаг	Выборка	Трудозатраты (час)	e_k (доля ошибок)	$1 - e_k$ (доля релевантов)	Δe_k
1%	53	8,83	0,0943	0,9057	0,0825
2%	105	17,50	0,1238	0,8762	0,0530
5%	263	43,83	0,1635	0,8365	0,0133
10%	526	87,67	0,1768	0,8232	0,0000

На основании полученных данных построен график соотношения показателей Δe_k и трудозатрат (Рисунок 4).

Данный график имеет в своей основе две оси ординат (Δe_k и трудозатраты), параллельно замеряемые по интервалу шага. При этом важно отметить, что в отношении Δe_k качество этого показателя растет по мере приближения к нулю (в обратную сторону). Как можно увидеть из графика, экспертно выведенное наиболее оптимальное соотношение качества проверки релевантности коллекции и трудозатрат находится в диапазоне от 2% до 5% (для данного кейса наиболее оптимальный интервал шага составил 3,6% в точке пересечения двух кривых).

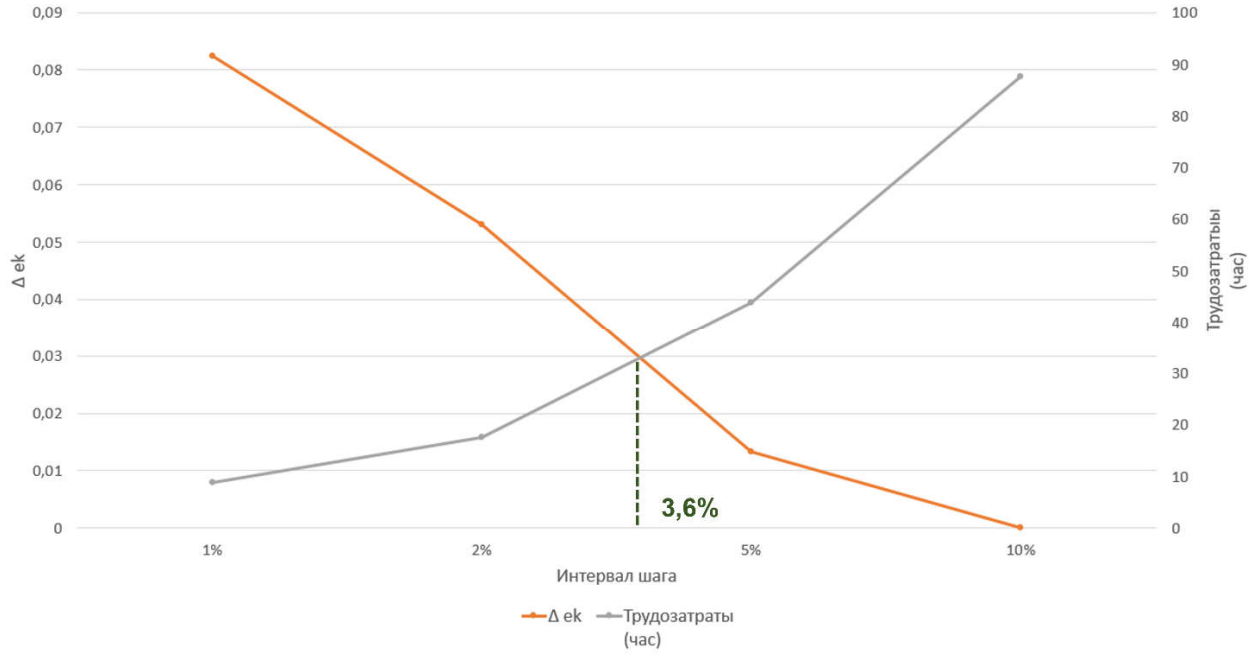

Рисунок 4 – Соотношение изменения качества оценки чистки коллекции в зависимости от потраченных на это трудозатрат

В целях экономии времени и трудозатрат расчет показателей точности (P_i) и глубины поиска (R_i) производился не по всем 79 элементам модели предметной области, а только по определенному числу элементов, достаточному для осуществления выводов о качестве отобранных коллекций и самого подхода в полной мере. Принимая во внимание тот факт, что в зависимости от основания модели может значительно меняться как тематический охват решений, так и сложность запроса (число поисковых терминов и характер взаимосвязи между ними), было принято решение рассчитать данные показатели по трём крупнейшим коллекциям по числу патентных семейств для каждого основания модели предметной области.

Таким образом, получится нивелировать существенные тематические различия между основаниями модели и сформировать достаточно значимые выводы по каждой из тематических категорий патентного ландшафта с совсем небольшим риском погрешностей (достаточных для построения полноценных выводов, систематических выводов и тенденций). При этом отметим, что, как было указано ранее, коллекциями по запросам могут существенно отличаться не только тематически или логически, но и общим масштабом коллекции. В рассмотренном кейсе присутствуют как элементы, коллекции по которым составляют несколько тысяч документов, так и те, которые имеют в своих коллекциях меньше десяти патентных семейств.

В рамках расчета точности и полноты поиска по трем наиболее масштабным элементам оснований будет предпринят схожий с описанным выше подход сэмплирования. На коллекциях свыше 600 документов будет использован уже выработанный интервал сэмплирования (3,6%), в то время как для коллекций менее 600 документов будет взят

более широкий интервал в 10%, чтобы избежать низкой репрезентативности выборок. Также среди крупных выборок для анализа были отобраны только те элементы, по которым формировались отдельные запросы: коллекции, сформированные на комбинации сразу нескольких запросов, не брались для большей чистоты исследования.

Данные расчеты будут выполнены по каждой из 82 коллекций, рассмотренных по поисковым запросам для элементов модели предметной области с учетом уже ранее выведенного интервала семплирования. В результате данной валидации будут высчитаны сразу три различных показателя, демонстрирующих качество чистки коллекции и качество соотнесения патентных семейств с элементами модели предметной области, заключающееся в точности и полноте поиска. При этом необходимо учитывать, что показатели точности поиска того или иного запроса могут вылезти за пределы корректного диапазона релевантности коллекции.

Возникновение подобных ситуаций не противоречит подходу, так как использование запросов для каждого элемента всегда может характеризоваться небольшими неточностями, некорректным составом поисковых терминов/операторов и другими неточностями, которые можно увидеть только при детальном анализе.

В соответствии с описанным выше подходом к проверке коллекции также принят за данность пороговый минимум релевантности коллекции (как генерализованной, так и по поисковым запросам) в 80% документов. Примем, что в случае подобной проверки с таким шагом сэмплирования, если коллекция по запросу имеет показатель менее 80%, такой запрос следует на процедуру корректировку группой аналитиков.

В результате образуется большая сводная таблица (Таблица 2), которую можно формировать вплоть до каждого элемента предметной области. При этом отметим, что подобная процедура является хорошим методом проверки качества запросов, которую можно проводить параллельно с предварительными стадиями формирования коллекции и корректировать точность запросов на более ранних этапах.

Таблица 2

Элемент	Число патентных семейств	Доля ошибок	Точность поиска (P_i)	Полнота поиска (R_i)
1.1 Вода как используемое сырье	4554	18/164	0,8902	0,1899
1.2.1 Природный газ как используемое сырье	673	8/24	0,6667	0,0208
1.3.1 Биомасса как используемое сырье	212	33/21	0,8571	0,0234
2.1 Тепловая энергия как источник энергии	380	2/38	0,9474	0,0468
2.2 Электрическая энергия как источник энергии	1591	2/57	0,9649	0,0715
2.3 Ядерная энергия как источник энергии	106	2/10	0,9000	0,0117
3.1 Электролиз	4413	4/159	0,9748	0,2016
3.3 Газификация твердого ископаемого топлива	213	1/21	0,9524	0,0260
3.5 Фотокаталитическое разложение воды	169	4/16	0,7500	0,0156
4.1 Транспортировка водорода в сжатом состоянии	887	3/32	0,9063	0,0377

Элемент	Число патентных семейств	Доля ошибок	Точность поиска (P_i)	Полнота поиска (R_i)
4.2 Транспортировка водорода в жидком состоянии	1637	16/59	0,7288	0,0559
4.3 Транспортировка водорода на химических носителях	34* **просмотрен полностью*	7/34	0,7941	0,0351
5.1 Топливные элементы	1002	3/36	0,9167	0,0429
5.1.4 Твердооксидные топливные элементы (SOFC)	115	0/11	1,0000	0,0143
5.2.1 Двигатели внутреннего сгорания	109	1/10	1,0000	0,0117
6.2 Выделение и концентрация CO_2	636	2/23	0,9130	0,0273
6.2.3 Сорбционные методы выделения CO_2	253	2/25	0,9200	0,0299
6.3 Хранение CO_2	294	1/29	0,9655	0,0364

В результате было выделено четыре элемента (1.2.1, 3.5, 4.2 и 4.3), по которым значение точности поиска не достигает минимального порогового значения в 80%. В таких ситуациях запрос признается некорректным и отправляется на доработку.

Итоговая групповая точность поиска P (соотношение общего числа релевантных документов по всем выборкам к общему числу замеряемых документов по выборкам) составила 89% до корректировок запросов и 91% после корректировок. Данные значения свидетельствуют, что подход к формированию расширяющих запросов в целом по коллекции собран с достаточно высокой степенью релевантности с небольшой долей некритических несоответствий, подлежащих корректировке.

Выводы, направления дальнейших исследований. В результате настоящего исследования были сформированы и апробированы рекомендации и возможные изменения в методиках работы с патентными ландшафтами на сверхбольших коллекциях (свыше 6 000 документов). В основе данных рекомендаций лежит задача снижения доли ручной обработки документов экспертами и связанных с ней трудозатрат при попытках минимизировать влияние применяемых алгоритмов автоматизации на снижение качества итогового продукта.

Наиболее приоритетным вариантом при работе с коллекциями такого рода в результате проведенного исследования представляется компромисс в виде проведения менее трудозатратной чистки с ограничением по полям документа и перевод процесса тегирования в полностью автоматизированную среду путем формирования т.н. «расширяющих» запросов.

Направление настоящего исследований представляется весьма актуальным в ближайшие годы по мере роста интереса к разработке патентных ландшафтов на более высоком стратегическом уровне с повышенной широтой тематического охвата. Дальнейшее развитие данный исследовательский вопрос может получить в работе над более глубинной проработкой контроля качества процессов разработки патентного ландшафта с использованием более продвинутой статистической базы, рассмотрением потенциального применения схожих алгоритмов автоматизации в контексте других процессов разработки патентного ландшафта (например, подготовка и интерпретация аналитических представлений) и др.

Список источников

1. Зеленкина Н.В., Павликова Д.С., Батанов Ф.А. Современная практика патентной аналитики // Интеллектуальная собственность. Промышленная собственность. 2019. № 6. С. 15–24.
2. Trippe A. Guidelines for Preparing Patent Landscape Reports // World Intellectual Property Organization (WIPO). 2015. p. 131. (In Eng.).
3. Батанов Ф.А., Зеленкина Н.В., Бачурина А.А. Углубленный анализ технологий в патентах // Интеллектуальная собственность. Промышленная собственность. 2020. № S. С. 75–81.
4. Ена О.В, Попов Н.В. Методология разработки патентных ландшафтов Проектного Офиса ФИПС // Станкоинструмент. 2019. № 1 (14). С. 28–35.
5. Ena O. «Domain-specific» Patent Analytics: Focus on Company's Technology Priorities // World Patent Information. 2021. Vol. 65. p.11. (In Eng.).
6. Aristodemou L., Tietze, F. The State-of-the-Art on Intellectual Property Analytics (IPA): A Literature Review on Artificial Intelligence, Machine Learning and Deep Learning Methods for Analyzing Intellectual Property (IP) Data // World Patent Information. 2018. Vol. 55. pp. 37–51. (In Eng.).
7. WIPO Manual on Open Source Tools for Patent Analytics // Wipo.int [Электронный ресурс]. – Режим доступа: https://www.wipo.int/publications/en/details.jsp?id=4168 (In Eng.).
8. Somer M. Media Values and Democratization: What Unites and What Divides Religious-Conservative and Pro-Secular Elites? // Turkish Studies. 2010. № 11 (4). pp. 555–577. (In Eng.).
9. Wywiał J. Contributions to Testing Statistical Hypotheses in Auditing. – Warszawa: PWN Scientific Publishers. 2016. (In Eng.).
10. Tillé Y. Algorithms of Sampling with Equal or Unequal Probabilities. 2010. (In Eng.).
11. Mucha B., Brestovanská P., Peráček T. Audit Sampling – Statistical vs. Non-statistical? // Journal of Eastern Europe Research in Business and Economics. 2018. pp.1–10. (In Eng.).
12. Yakimova V., Radomskii V. Using the Analytical Procedures to Form Audit Sampling in the Audit of Wage Settlements // International Accounting. № 20 (15). pp. 897–916. (In Eng.).
13.Тверской государственный технический университет (ТвГТУ). Интеллектуальная распределенная система информационной поддержки инноваций в науке и образовании. Отчет о патентных исследованиях. – Тверь, 2013. – 50 с.
14. Tillé Y. Sampling Algorithms. – New York: Springer, 2011. (In Eng.).

References

1. Zelenkina N.V., Pavlikova D.S., Batanov F.A. Modern Practice of Patent Analytics. *Intellektual'naya sobstvennost'. Promyshlennaya sobstvennost'*. 2019. No. 6. pp. 15–24. (In Russ.).
2. Trippe A. Guidelines for Preparing Patent Landscape Reports. *World Intellectual Property Organization (WIPO).* 2015. p. 131.
3. Batanov F.A., Selenkina N.V., Bachurina A.A. In-depth Analysis of Technologies in Patents. *Intellektual'naya sobstvennost'. Promyshlennaya sobstvennost'*. 2020. No. S. pp. 75–81. (In Russ.).
4. Ena O.B., Popov N.V. Methodology of Development of Patent Landscapes of the FIPS Project Office. *Stankoinsrtument.* 2019. No. 1 (14). pp. 28–35. (In Russ.).
5. Ena O. «Domain-specific» Patent Analytics: Focus on Company's Technology Priorities. *World Patent Information.* 2021. Vol. 65. p.11.
6. Aristodemou L., Tietze, F. The State-of-the-Art on Intellectual Property Analytics (IPA): A Literature Review on Artificial Intelligence, Machine Learning and Deep Learning Methods for Analyzing Intellectual Property (IP) Data. *World Patent Information.* 2018. Vol. 55. pp. 37–51.
7. WIPO Manual on Open Source Tools for Patent Analytics. *Wipo.int*. Available at: https://www.wipo.int/publications/en/details.jsp?id=4168
8. Somer M. Media Values and Democratization: What Unites and What Divides Religious-Conservative and Pro-Secular Elites? *Turkish Studies.* 2010. No. 11 (4). pp. 555–577.
9. Wywiał J. Contributions to Testing Statistical Hypotheses in Auditing. *Warszawa: PWN Scientific Publishers.* 2016.
10. Tillé Y. Algorithms of Sampling with Equal or Unequal Probabilities. 2010.
11. Mucha B., Brestovanská P., Peráček T. Audit Sampling – Statistical vs. Non-statistical? *Journal of Eastern Europe Research in Business and Economics.* 2018. pp.1–10.
12. Yakimova V., Radomskii V. Using the Analytical Procedures to Form Audit Sampling in the Audit of Wage Settlements. *International Accounting.* No. 20 (15). pp. 897–916.
13. Tver State Technical University (TvSTU). Intelligent Distributed System of Information Support for Innovations in Science and Education. Patent Research report. *Tver'*. 2013. 50 p. (In Russ.).
14. Tillé Y. Sampling Algorithms. *New York: Springer.* 2011.

Статья поступила в редакцию 17.05.2022; одобрена после рецензирования 25.05.2022; принята к публикации 29.05.2022.
The article was submitted 17.05.2022; approved after reviewing 25.05.2022; accepted for publication 29.05.2022.