УДК 330.1:332.1:614.8:658.5 DOI: https://doi.org/10.25728/pu.2018.6.5

ФОРМАЛИЗОВАННЫЕ МОДЕЛИ И МЕТОДЫ АНАЛИЗА И ОЦЕНКИ ПОЛНОТЫ ПАТЕНТНЫХ ИНФОРМАЦИОННЫХ ФОНДОВ (НА ПРИМЕРЕ МЕЖДУНАРОДНОЙ ПАТЕНТНОЙ ОРГАНИЗАЦИИ)

В.О. Сиротюк

Сформулированы требования к структуре и полноте патентного информационного фонда. Дано определение показателя полноты фонда и предложены формализованные модели и методы ее анализа и оценки. Полученные результаты применены для оценки полноты и совершенствования баз данных патентного информационного фонда международной патентной организации — Евразийского патентного ведомства.

Ключевые слова: патентный информационный фонд, патентный поиск международного типа, базы данных патентной и непатентной информации, тематическая база данных, показатель полноты патентного информационного фонда, эталонная база данных патентной и непатентной информации.

ВВЕДЕНИЕ

Главная функция патентного ведомства заключается в проведении патентной экспертизы, в ходе которой заявка на изобретение оценивается на новизну, изобретательский уровень и промышленную применимость. Важное место в экспертизе заявок занимает патентный поиск, цель которого состоит в выявлении предшествующего уровня техники. Патентный поиск проводится на основании формулы изобретения с учетом описания и чертежей в объеме, соответствующем международным требованиям и правилам, предусмотренными Договором о патентной кооперации (PCT) [1]. Патентный поиск должен охватывать максимально возможное число релевантных источников информации безотносительно к языку, на котором издан документ, срока давности и типа документа.

Эффективность и качество патентного поиска определяются, в первую очередь, полнотой ПИФ, который представляет собой совокупность классифицированной, структурированной и систематизированной по областям знаний (тематикам) патентной документации, непатентной литературы и патентно-ассоциированной документации, снабженной справочно-поисковым аппаратом и инструментально-программными средствами доступа к данным и их обработки [2].

Принятые сокращения

БД — база данных
ВОИС — Всемирная организация интеллектуальной собственности
ЕАПАТИС — Евразийская патентно-информационная система
ЕАПВ — Евразийское патентное ведомство
ЕПВ — Европейское патентное ведомство
МПК — Международная патентная классификация
НПБД — БД непатентной литературы
ПБД — БД патентной документации
ПИФ — патентный информационный фонд
PCT — Patent Cooperation Treaty
ТБД — тематическая БД
ТЭБД — тематическая эталонная БД
УДК — Универсальная десятичная классификация
ЦБИС — цифровая библиотека интеллектуальной собственности
ЭБД — эталонная БД

Патентное ведомство несет ответственность за полноту патентного поиска, поэтому при его проведении должно быть использовано максимально возможное число патентных документов соответствующих классификационных рубрик МПК, независимо от языка публикации документов и глубины их ретроспективы, научно-техническая (непатентная) литература, отобранная из внутренних и внешних источников.

Для проведения полноценного патентного поиска международного типа ведомству рекомендуется иметь в своем распоряжении или иметь доступ к патентной и непатентной документации, определенной в рамках договора PCT, просмотр которой обязателен при рассмотрении заявок [1, 3]. Современные информационно-телекоммуникационные технологии и инфраструктуры, в том числе облачные, обеспечивают доступ к ПБД, выкладываемым патентными ведомствами в сети Интернет, как правило, на условиях бесплатного пользования. Кроме того, региональные и международные патентные организации (ВОИС, ЕАПВ, ЕПВ) предоставляют свободный доступ к своим информационно-поисковым системам Patentscop [3], ЕАПАТИС [4] и Espacenet [5] соответственно, содержащим ПБД региональной и национальной патентной документации. Активно разрабатываемые и внедряемые в информационных системах ведомств средства машинного перевода позволяют преодолеть также «языковый барьер» и делают «читаемыми» патентную документацию, представленную на разных языках.

Базы данных непатентной литературы, как правило, предоставляются за определенную плату. Источниками НПБД служат крупные издательства, библиотеки, научные и исследовательские центры и организации, занимающиеся оцифровкой хранящихся у них материалов и переводом их в электронную форму. Например, порталы научно-технической литературы такие, как Science Direct (http://www.sciencedirect.com), High Wire Press (http://highwire.stanford.edu/lists/freeart.dtl), American Chemical Society (http://pubs.acs.org/about.html), IEEE (http://ieeexplore.ieee.org/browse/periodicals/title/); сайты непатентной литературы в области химии Pharmaceutical Research (http://www.springerlink.com/content/1573-904X/), Chemistry Journals (University of Cambridge) (http://www.ch.cam.ac.uk/c2k/cj), ChemMedChem (http://www3.interscience.wiley.com/cgi-bin/jtoc/110485305/), в области медицины — European Journal of Cancer (http://www.sciencedirect.com/science/journal/09598049), Русский медицинский журнал (http://rmj.ru), American Journal of Hematology (http://www3.interscience.wiley.com/cgi-bin/jhome/112621403), в области физики — European Journal of Physics (http://www.iop.org), Health Physics (http://www.health-physics.com) и др.

Таким образом, в современных условиях для обеспечения проведения полноценных патентных поисков ПИФ патентного ведомства должен иметь распределенную информационно-управляющую структуру и содержать локальные (внутренние) БД патентной и непатентной информации ведомства, средства доступа к внешним распределенным патентно-информационным ресурсам, средства поиска и обработки данных, а также сервисные средства [2, 6]. Исходя из этого, полнота ПИФ определяется не только и не столько числом хранимых в локальных БД документов в электронной форме, но также структурой и составом образующих их информационных элементов, наличием, структурой и возможностью реализации путей доступа к требуемым при проведении патентных поисков внешним БД патентной и непатентной информации и реализованным в них сервисным средствам.

Отсутствие четкого определения полноты ПИФ, имеющего в современных условиях распределенную информационно-управляющую структуру, а также формализованных моделей и методов анализа и оценки показателя полноты ПИФ не позволяет ведомствам объективно оценивать свое состояние в области информационного обеспечения экспертизы, вырабатывать на этой основе стратегию и планы мероприятий по развитию информационно-технологической поддержки процессов проведения полноценных патентных поисков и принятия экспертами эффективных и качественных решений по патентоспособности заявок на изобретения.

Настоящая работа направлена на решение данной проблемы. В работе предложена единая методология анализа и оценки полноты ПИФ, охватывающая этапы анализа структур и содержания локальных баз данных ПИФ ведомства и доступных внешних БД, расчета численных значений показателя их полноты, анализа и оценки данных показателей, подготовки предложений и рекомендаций по повышению эффективности и качества информационных фондов, в том числе патентных. Предлагаемая методология базируется на комплексе разработанных формализованных моделей, методов и алгоритмов анализа и оценки структурной, функциональной, структурно-функциональной и информационной полноты фондов, а также методов и процедур построения канонических структур тематических БД патентной и непатентной информации, используемых для оценки полноты фондов.

1. ТРЕБОВАНИЯ К СТРУКТУРЕ ПАТЕНТНОЙ И НЕПАТЕНТНОЙ ДОКУМЕНТАЦИИ

Рассмотрим общие требования, предъявляемые к структуре патентных документов, составляющих основу ПИФ.

Патентный документ в общем виде содержит следующие составные части: титульный лист, на котором приводятся библиографические данные, формула и/или реферат изобретения, а также, при наличии, основной чертеж; полный текст описания изобретения, оформленный в соответствии со стандартами ВОИС; графические материалы (чертежи, рисунки, схемы и т. п.); дополнительные материалы, например, отчет о результатах международного поиска по заявке, данные о правовом статусе патента, данные об изменениях и другие сведения. Отметим, что патентная документация достаточно хорошо и строго формализована. Существует ряд международных стандартов ВОИС на представление патентных документов, например, стандарт ST16 определяет допустимые коды видов документов, стандарт ST3 — средства дли идентификации патентных ведомств или организаций, ST10/C — допустимые правила задания номеров приоритетных заявок и др. [7].

Структура описания непатентной документации в соответствии со стандартом ST14 ВОИС включает в себя:

— имя автора;
— название статьи, монографии;
— аннотацию (реферат);
— номер издания;
— место публикации и имя издателя;
— год публикации;
— полный текст описания.

Для классификации непатентной документации применяется УДК.

Патентно-ассоциированные документы в основном неструктурированные и содержат законодательные, нормативно-правовые и справочные сведения. По сути, эти документы не играют существенной роли при проведении патентных поисков и могут быть отнесены для упрощения дальнейшего изложения к одной из разновидностей непатентной документации.

2. ОПРЕДЕЛЕНИЕ ПОКАЗАТЕЛЯ ПОЛНОТЫ ПАТЕНТНОГО ИНФОРМАЦИОННОГО ФОНДА

Введем определение полноты ПИФ.

Показатель полноты ПИФ характеризует соответствие состава, структуры, содержания, эксплуатационных и сервисных характеристик локальных (внутренних) и внешних распределенных, доступных патентному ведомству, ПБД и НПБД спецификациям требований к составу и структурам данных, контенту, поисковым, сервисным и функциональным требованиям эталонных БД фондов патентной документации и непатентной литературы.

В дальнейшем для удобства изложения под базами данных ПИФ будем понимать как ПБД, так и НПБД, а под ЭБД — эталонные ПБД и эталонные НПБД.

Составляющими элементами показателя полноты ПИФ служат структурная, функциональная, структурно-функциональная и информационная полнота ПИФ. Введем их определения.

Структурная полнота определяется отношением числа входящих в базы данных ПИФ с учетом классификационных индексов МПК структурных элементов (объектов данных и информационных элементов), связей между элементами и путей доступа к ним, в том числе виртуальных, реализуемых посредством организации доступа к внешним распределенным БД патентной и непатентной документации, к их общему числу, зафиксированному в ЭБД. Оценка данного показателя характеризует полноту информационного фонда «вширь». Отсутствие каких-либо информационных элементов, например, данных о приоритете заявки, номера международной публикации и др., а также отдельных частей документов в структуре БД, например, реферата или полного описания заявки (патента) делает невозможным проведение полноценных патентных поисков, а также получение полной и достоверной информации о найденных документах. Это не позволяет экспертам принимать качественные решения относительно патентоспособности заявок на изобретения. Структурная полнота является интегральным показателем полноты по элементам (элементной полноты) и полноты по связям (отношениям) между элементами.

Функциональная полнота определяется отношением числа методов и процедур поиска данных (включая средства нумерационного, именного, тематического поиска), обработки баз данных ПИФ (пополнения, обновления, контроля, верификации и др.), удаленного доступа к внешним источникам патентной и непатентной информации и метапоиска в них, машинного перевода, наглядного визуального отображения документов, и других сервисных процедур к их общему числу, реализованному в ЭБД. Данный показатель позволяет оценить технологическую оснащенность и функциональные возможности баз данных ПИФ в распределенной информационно-управляющей инфраструктуре ПИФ.

Структурно-функциональная полнота определяется отношением числа классифицированных в соответствии с МПК объектов данных, информационных элементов, связей и путей доступа, а также процедур и методов доступа, поиска, обработки, отображения данных и сервисных процедур, реализованных на сформированной информационно-управляющей структуре ПИФ, к их общему числу в ЭБД. Данный показатель является интегральным показателем структурной и функциональной полноты ПИФ.

Информационная полнота характеризует отношение числа хранимых в локальных базах данных ПИФ ведомства и в доступных внешних БД документов к их общему числу, определенному в ЭБД. Показатель информационной полноты характеризует полноту ПИФ «вглубь», т. е. по числу хранимых в базах данных ПИФ документов (записей). Его оценка важна в смысле числа находимых при реализации поисковых запросов патентных и непатентных документов. Например, требованиями к минимуму документации PCT установлена глубина ретроспективы для патентной документации СССР с 1924 г., для патентной документации России с 1991 г., для патентной документации ВОИС и ЕПВ с 1978 г. и т. д. [3, 4].

3. ОСНОВНЫЕ ПОЛОЖЕНИЯ ПРЕДЛАГАЕМОЙ МЕТОДОЛОГИИ АНАЛИЗА И ОЦЕНКИ ПОЛНОТЫ ПАТЕНТНОГО ИНФОРМАЦИОННОГО ФОНДА

Разработанная методология анализа и оценки полноты ПИФ базируется на комплексе формализованных моделей, методов и алгоритмов анализа и оценки структурной, функциональной, структурно-функциональной и информационной полноты фондов, а также методов и процедур построения канонических структур ТБД патентной и непатентной информации, используемых для оценки полноты фондов. Разработанные модели, методы и алгоритмы обеспечивают:

— построение (или формализацию существующих) канонических структур ЭБД и баз данных ПИФ;

— анализ общности (сходства) канонических структур баз данных ПИФ и ЭБД;

— расчет численных значений составляющих показателя полноты ПИФ;

— анализ и оценку составляющих показателей;

— определение общего показателя полноты ПИФ;

— подготовку предложений и рекомендаций по повышению эффективности и качества баз данных ПИФ.

Оценка структурной и функциональной, а также структурно-функциональной полноты ПИФ осуществляется с помощью методов анализа общности (сходства) канонических структур баз данных ПИФ и ЭБД, формируемых на этапе концептуального проектирования БД [8, 9]. Как известно, на данном этапе создается независимая от системы управления базой данных и среды реализации безызбыточная структура данных, наиболее полно и адекватно описывающая свойства и характеристики предметной области пользователей БД. Проектирование БД на других уровнях представления (логическом, физическом) связано, как правило, с введением некоторой избыточности по данным и связям с целью возможности их реализации в выбранной среде [9]. Поэтому определение структурной и функциональной, а также структурно-функциональной полноты ПИФ целесообразно осуществлять на уровне концептуального представления данных.

Разработанные методы анализа общности основаны на оценке функций подобия спецификаций структурных и функциональных требований к ЭБД и базам данных ПИФ путем их сравнения. Они обеспечивают оценку структурной и функциональной полноты ПИФ путем вычисления функций подобия баз данных ПИФ и ЭБД по хранящимся в них объектам данных, информационным элементам, связям и процедурам поиска и обработки данных, а также структурно-функциональной полноты ПИФ. Информационная полнота ПИФ рассчитывается как отношение количественных характеристик (экземпляров) классов, объектов данных и информационных элементов, зафиксированных в локальных и доступных внешних базах данных ПИФ, к количественным характеристикам ЭБД.

В целях сокращения размерности решаемых задач, а также возможности применения разработанной методологии, моделей, методов и алгоритмов для оценки полноты информационных фондов другого назначения (например, библиотек, научных и исследовательских организаций, архивов и др.), предложены методы построения канонических структур ТЭБД и ТБД патентного информационного фонда, создаваемых для определенных областей знаний, например, для отдельных классификационных рубрик МПК и связанных с ними таблицами соответствия индексов УДК. В качестве тематических предметных областей могут выступать, например, документы и информация, относящиеся к разделам МПК «А. Удовлетворение жизненных потребностей человека» или «С. Химия, металлургия», или «Н. Электричество», или к более дробным внутри разделов классам, подклассам, рубрикам и связанным с ними таблицами соответствия непатентные документы и информация из разделов УДК, например, «0. Общий раздел», «3. Общественные науки», «5. Математика. Естественные науки» соответственно. Формируемые для этих предметных областей ТБД могут быть определены как специальным образом отобранные для решения определенной научной или прикладной проблемы (в том числе для определения предшествующего уровня техники) из патентных и непатентных первоисточников и организованные коллекции тематических связанных данных (патентных и непатентных).

Для построения канонической структуры ТЭБД применяются методы объектно-ориентированного анализа и проектирования БД, так как они наибо-

лее адекватно отражают технологию формирования и хранения ТЭБД в распределенной информационно-управляющей структуре и предоставляемых ТЭБД услуг. С этой целью модифицированы и развиты методы объектно-ориентированного анализа и проектирования БД [9, 10], учитывающие особенности предметной области ТЭБД, в частности, как отмечалось выше, требования и рекомендации стандартов ВОИС и цифровых библиотек интеллектуальной собственности, а также экспертов, проводящих патентные поиски международного типа. В результате их применения формируется объектная каноническая структура ТЭБД.

Канонические структуры ТБД строятся с помощью методов и процедур (структурных или объектно-ориентированных), рассмотренных в работах [8, 9].

Канонические структуры находящихся в эксплуатации (функционирующих) ЭБД и баз данных ПИФ известны из их описания в технической документации на БД. Для расчета численных значений составляющих показателя полноты ПИФ согласно предложенной методологии требуется их представление в формализованном виде согласно методикам, приведенным в работах [8, 9].

На завершающем этапе предложенной методологии на основе полученных количественных значений оценок показателей полноты баз данных ПИФ (структурной, функциональной, структурно-функциональной и информационной) патентным ведомством разрабатываются планы мероприятий по совершенствованию и развитию конкретных БД (их состава, структуры, технологического и инструментально-программного окружения), что позволяет в целом повысить эффективность и качество ПИФ.

4. ОСОБЕННОСТИ ПОСТРОЕНИЯ И ФОРМАЛИЗОВАННОЕ ОПИСАНИЕ КАНОНИЧЕСКИХ СТРУКТУР ТЕМАТИЧЕСКИХ БАЗ ДАННЫХ

Рассмотрим особенности построения эталонных БД. Эталонные БД разрабатываются на основе общих и специальных требований, предъявляемых стандартами ВОИС к составу, содержанию и структуре патентной документации и непатентной литературы, требований и рекомендаций к поисковым, сервисным и эксплуатационным характеристикам доступа к данным и их использования, предъявляемых к современным цифровым библиотекам интеллектуальной собственности (или IPDL — Intellectual Property Digital Library) [3], а также информационных и функциональных требований, предъявляемых к проведению экспертами патентных поисков международного типа. В качестве ЭБД патентной документации могут служить ПБД либо источника информации (патентная БД национального патентного ведомства, например, ПБД USPTO или ПБД JPO, или ПБД Роспатента и др.), либо ПБД всемирных патентных информационных систем типа Patentscope или Espacenet [3, 5], содержащих коллекцию ПБД стран, входящих в минимум документации PCT. В качестве ЭБД непатентной литературы могут служить НПБД порталов или сайтов непатентной литературы, рассмотренные выше. Рабочее место эксперта, проводящего патентные поиски международного типа в ЭБД, оснащается современным инструментарием для доступа к данным и обработки патентной и непатентной документации, сервисными средствами работы с ней, включая машинный перевод.

Построение объектной канонической структуры ТЭБД осуществляется в четыре этапа [9, 10].

На *первом этапе* формируется объектная модель предметной области ТЭБД, формализованно описываемой системой множеств и булевых матриц смежности между компонентами предметной области. Описание предметной области ТЭБД включает в себя:

— пользователей ТЭБД, к которым относятся специалисты — эксперты, проводящие патентные поиски международного типа по определенным областям знаний (тематикам), и сотрудники служб администраторов БД патентной и непатентной документации, отвечающие за создание, сопровождение и администрирование ТЭБД, предоставление доступа к данным и сервисам;

— объекты данных и их характеристики;

— информационные элементы и их характеристики;

— процедуры поиска и обработки данных, последовательность их выполнения и характеристики, включая процедуры виртуального доступа к данным внешних источников и метапоиска в них, а также средства машинного перевода и другие сервисные операции и процедуры;

— отношения между объектами данных, информационными элементами и процедурами поиска и обработки данных.

На *втором этапе* на основе сформированной объектной модели предметной области формируются модели спецификаций информационных и функциональных требований пользователей, представляемых, соответственно, в виде множества n-арных (в частном случае, бинарных) отношений между структурными элементами и элементами множества предикатов. Для предметной области ТЭБД, основу которой составляют патентные документы, имеющие формализованный, стандартизованный вид, характерно наличие в основном двуместных предикатов.

Полученные на первом и втором этапах результаты используются на *третьем этапе* для формирования объектных моделей требований пользова-

телей и обобщенной объектной структуры требований пользователей.

На *четвертом этапе* строится безызбыточная (минимальная) объектная каноническая структура ТЭБД путем сведения многообразия объектных моделей требований пользователей, зафиксированных в обобщенной объектной структуре пользователей ТЭБД, к базовым и специфическим классам объектов.

Для реализации процедур каждого этапа модифицированы и развиты с учетом особенностей предметной области ТЭБД модели и методы, рассмотренные в работах [9, 10]. В результате их выполнения формируется объектная каноническая структура ТЭБД, формализованно представляемая в виде графа $G_{\text{к.с}}^{\text{об}}(O, \Delta)$, вершинами которого $O = \{O_\varepsilon / \varepsilon = \overline{1, \varepsilon_0}\}$ служат классы и объекты данных предметной области, а дугами $\Delta = \{\delta_{\varepsilon\varepsilon'} / \varepsilon, \varepsilon' = \overline{1, \varepsilon_0}\}$ — связи (или отношения) между классами и объектами данных. Характеристиками графа $G_{\text{к.с}}^{\text{об}}$ являются интегральные характеристики классов (объектов) и связей между ними. При этом свойства класса (объекта) определяются включенными в его состав агрегатами (группами) данных и информационными элементами, а поведение — методами (функциями) поиска и обработки данных, среди которых выделено подмножество интерфейсных и подмножество реализационных процедур. Каждый класс O_ε формально описывается множеством информационных элементов $D_\varepsilon = \{d_l^\varepsilon\}$, матрицей смежности информационных элементов $B_\varepsilon = \| b_{ll'}^\varepsilon \|$, множеством процедур поиска и обработки данных $P_\varepsilon = \{p_r^\varepsilon\}$ и матрицей технологии обработки информационных элементов $W_\varepsilon = \| w_{rl}^\varepsilon \|$.

Рассмотрим особенности формирования ТБД.

Как отмечалось ранее, ТБД представляет собой коллекцию (подборку) отобранных из разных источников (ПБД и НПБД) патентных и непатентных документов определенной (заданной) тематики. Она формируется в результате проведения тематических поисков в базах данных ПИФ с помощью индексов МПК, УДК и ключевых слов (выражений), связываемых в запросе посредством логических функций (И/ИЛИ/НЕТ и др.). Учитывая, что характеристики баз данных ПИФ, как правило, разные и зависят от используемой при их создании информации (состава и структуры файлов и массивов данных, получаемых патентным ведомством по обмену из других организаций и патентных ведомств), то и отбираемые в ТБД из этих БД данные имеют разный состав, структуру и характеристики. Поэтому ТБД на физическом уровне можно представить в виде мультисписка, состоящего из разделов, число которых соответствует числу используемых в тематическом запросе баз данных ПИФ. При этом каждый раздел содержит записи определенной структуры и содержания, соответствующей структуре той базе данных ПИФ, из которой они были отобраны, и содержит информацию о найденных тематически связанных патентных и/или непатентных документах.

Единая интегрированная каноническая структура ТБД строится путем объединения канонических структур баз данных ПИФ с помощью предложенных в работе [8] методов и процедур. В дальнейшем данная структура используется для определения только одной составляющей показателя полноты ПИФ — информационной полноты ПИФ. Использование единой канонической структуры ТБД для оценки других показателей полноты ПИФ (структурной и функциональной) невозможно из-за различий, как было отмечено выше, в структурах данных. Так, например, если структура одной из баз данных ПИФ является подмножеством структуры другой БД, что характерно для тех БД, которые не в полной мере удовлетворяют требованиям ВОИС и ЦБИС, то при их объединении в каноническую структуру ТБД первая структура «поглотится» второй и создастся ложное представление, что структурная полнота этой БД достаточна и не требуются дальнейшие действия по ее совершенствованию. Поэтому структурная, функциональная и структурно-функциональная полнота ТБД рассчитываются отдельно для каждой базы данных ПИФ, используемой при формировании ТБД, путем попарного сравнения и анализа канонических структур ТЭБД и баз данных ПИФ. Это позволяет в дальнейшем разработать мероприятии по совершенствованию и развитию каждой базы данных ПИФ в отдельности и тем самым повысить полноту ПИФ в целом. В частном случае, при проведении патентных исследований по заданной тематике по фонду патентной документации только одной страны (например, России или Германии, или США и др.) формируется одна ТБД.

Для построения канонической структуры баз данных ПИФ могут применяться методы и процедуры (структурные или объектно-ориентированные), рассмотренные в работах [8, 9]. Дополнительная особенность применения данных процедур состоит в возможности их применения для построения баз данных ПИФ, характеризующихся содержанием разнородной мультимедийной информации из патентных и непатентных источников. Обозначим $S = \{s_k\}$, $k = \overline{1, K_0}$ — множество локальных баз данных ПИФ. Каноническая структура k-й БД представляется в виде графа $G_{\text{к}}(D, U)$, где $D = \{d_l\}$, $l = \overline{1, L}$ — полное множество структурных

элементов, $D = D^{об} \cup D^{зп}$, где $D^{об} = \{d_l\}$, $l = L_{об}$ — подмножество объектов (групп) данных, $D^{зп} = \{d_l\}$, $l \in L_{зп}$, — подмножество информационных элементов (ключей и атрибутов объектов (групп) данных), $L = L_{об} \cup L_{эл}$; $F = \{f_r\}$, $r = \overline{1, R}$, — множество процедур поиска и обработки данных, доступа к внешним БД, машинного перевода, метапоиска и других сервисных процедур; $U = U_1 \cup U_2 \cup U_3$ — полное множество взаимосвязей между структурными элементами, где $U_1 = \{d_l, d_{l'}\}$, $l, l' \in L_{об}$ — множество взаимосвязей между объектами (группами) данных, $U_2 = \{d_l, d_{l'}\}$, $l, l' \in L_{эл}$ — множество взаимосвязей между ключами и атрибутами объектов (групп) данных, $U_3 = \{f_r, d_l\}$, $l \in L_{эл}$, $r = \overline{1, R}$ — множество взаимосвязей между процедурами поиска и обработки данных и используемыми ими информационными элементами. Для каждого объекта (группы) данных определен состав образующих его информационных элементов (ключей и атрибутов) $H_l^{об} = \{d_l\}$ и процедур (методов) поиска и обработки данных $H_l^{об} = \{f_r\}$.

Тематическую базу данных ПИФ, формируемую в результате выполнения тематического поискового запроса к базам данных ПИФ, формализованно представим в виде мультисписка (множества) $T = \{t_1, t_2, ..., t_k, ..., t_{K_0}\}$, где t_k — раздел ТБД, содержащий записи k-й базы данных ПИФ, отобранные в соответствии с классификационными индексами МПК и УДК.

5. ФОРМАЛИЗОВАННЫЕ МЕТОДЫ РАСЧЕТА ПОКАЗАТЕЛЕЙ ПОЛНОТЫ ТЕМАТИЧЕСКИХ БАЗ ДАННЫХ

Оценка полноты ТБД патентной документации и непатентной литературы ПИФ патентного ведомства осуществляется с помощью разработанных методов анализа общности канонических структур баз данных ПИФ, используемых для формирования ТБД и ТЭБД. Эти методы основаны на оценке функций подобия спецификаций структурных, функциональных и информационных требований к ТЭБД и базам данных ПИФ путем их сравнения и вычисления соответствующих значений показателя полноты ТБД.

Разработанные методы обеспечивают определение элементной полноты, полноты по связям (отношениям), структурной полноты, функциональной полноты и структурно-функциональной полноты баз данных, а также информационной полноты ТБД.

Для определения полноты баз данных ПИФ по информационному составу (элементной полноты) определим числа:

— общих в множествах O и D структурных элементов (объектов (групп) данных, информационных элементов): $p_{11} = |O \cap D|$;

— структурных элементов, присутствующих в множестве O, но отсутствующих в множестве D: $p_{10} = |O| - p_{11}$;

— структурных элементов, присутствующих в множестве D, но отсутствующих в множестве O: $p_{01} = |D| - p_{11}$.

Для оценки элементной полноты баз данных ПИФ с учетом особенностей рассматриваемых множеств может быть применен нормированный показатель подобия, предложенный в работах [9, 10] и вычисляемый по формуле:

$$\varepsilon_{зп} = \frac{1}{2}\left(\frac{p_{11}}{p_{11}+p_{10}} + \frac{p_{11}}{p_{11}+p_{01}}\right) = \frac{1}{2}\left(\frac{p_{11}}{|O|} + \frac{p_{11}}{|D|}\right) = \frac{\alpha+\beta}{2}. \quad (1)$$

Показатель $\varepsilon_{зп}$:

— учитывает взаимную меру общности элементов множеств O и D и принимает значения $0 \le \varepsilon_{зп} \le 1$;

— монотонно возрастает с увеличением числа p_{11}.

Нормированный показатель подобия принимает максимальное значение ($\varepsilon_{зп} = 1$), когда информационные составы сравниваемых БД (базы данных ПИФ и ТЭБД) полностью совпадают. При $\varepsilon_{эл} = 0$ сравниваемые информационные структуры не имеют ни одного общего элемента.

Далее определим степень общности БД и ТЭБД по взаимосвязям между структурными элементами. Поскольку общие взаимосвязи между элементами структур $G_{к.с}^{об}(O, \Delta)$ и $G_k(D, U)$ могут существовать только между элементами множества пересечения $D_{ko} = O \cap D$, анализ общности структур взаимосвязей проводится на множестве элементов D_{ko}. При этом разработанные процедуры анализа позволяют учитывать не только наличие общих дуг (отношений) между парами структурных элементов множества D_{ko}, но и наличие общих путей доступа между элементами.

Данный анализ проводится для ограниченного (зафиксированного) множества элементов, не превышающего мощности множества элементов структуры $G_{к.с}^{об}(O, \Delta)$. Поэтому для оценки полноты баз данных ПИФ по связям (отношениям) между структурными элементами будем пользоваться мерой подобия, вычисляемой с помощью показателя Жаккарда:

$$\varepsilon_{св} = \frac{p_{11}}{p_{11}+p_{10}+p_{01}}. \quad (2)$$

Числа p_{11}, p_{10}, p_{01}, входящие в выражение (2), вычисляются, исходя из специфики рассматриваемой задачи анализа с учетом наличия общих взаимосвязей и (или) путей доступа между элементами множества D_{ko}, следующим образом:

$$p_{11} = \sum_i |F(d_i) \cap F(d_{i'})|; \quad p_{10} = \sum_i |F(d_i)| - p_{11};$$

$$p_{01} = \sum_i |F(d_{i'})| - p_{11}, \quad \forall d_i, d_{i'} \in D_{ko},$$

где $F(d_i) = \{d_j\}$, $F(d_{i'}) = \{d_{j'}\}$ — множества достижимости, соответственно, для элементов $d_i \in O$ и $d_i \in D$ ($d_i, d_{i'} \in D_{ko}$).

Анализ выражения (2) показывает, что показатель подобия $\varepsilon_{св}$ принимает значения в интервале $0 \le \varepsilon_{св} \le 1$. При $\varepsilon_{св} = 0$ анализируемые структуры баз данных ПИФ и ТЭБД не имеют общих взаимосвязей или путей доступа, а $\varepsilon_{св} = 1$ означает тождественность структур БД по их взаимосвязям между элементами.

Исходя из полученных выражений для вычисления мер подобия (1) и (2) структурная полнота БД ПИФ $\varepsilon_{стр}$ определится как $\varepsilon_{стр} = \varepsilon_{эл} + \varepsilon_{св}$.

Показатель структурной полноты $\varepsilon_{стр}$ принимает значения в интервале $0 \le \varepsilon_{стр} \le 2$. Максимальное значение, равное 2, показатель $\varepsilon_{стр}$ принимает, если структура баз данных ПИФ (по составу образующих ее элементов и связей между ними) полностью соответствует структуре ТЭБД по соответствующей тематике (области знаний).

Функциональную полноту баз данных ПИФ целесообразно оценивать с помощью показателя меры подобия Сокала и Мичинера, учитывающего число схожих элементов (процедур поиска, доступа и обработки данных) в канонических структурах баз данных ПИФ и ТЭБД:

$$\varepsilon_{пр} = \frac{2P'_{11}}{2P'_{11} + P'_{10} + P'_{01}}, \qquad (3)$$

где P'_{11} — число общих процедур (методов) в требованиях пользователей баз данных ПИФ и ТЭБД, P'_{10} — число процедур, присутствующих в требованиях пользователей ТЭБД, но отсутствующие в требованиях пользователей баз данных ПИФ; P'_{01} — число процедур, присутствующих в требованиях пользователей баз данных ПИФ, но отсутствующие в требованиях пользователей ТЭБД.

Анализ выражения (3) показывает, что показатель подобия $\varepsilon_{пр}$ принимает значения в интервале $0 \le \varepsilon_{стр} \le 1$. При $\varepsilon_{пр} = 0$ анализируемые структуры не имеют общих процедур обработки данных, а $\varepsilon_{пр} = 1$ означает тождественность набора процедур в сравниваемых структурах.

Интегрированный показатель структурно-функциональной полноты баз данных ПИФ ε вычисляется по формуле:

$$\varepsilon = \varepsilon_{стр} + \varepsilon_{пр} = \varepsilon_{эл} + \varepsilon_{св} + \varepsilon_{пр}. \qquad (4)$$

Интегрированный показатель подобия, вычисляемый по формуле (4), принимает максимальное значение $\varepsilon = 3$ в случае, когда структурная полнота и функциональная полнота баз данных ПИФ максимальны и тождественны полноте ТЭБД.

Для определения информационной полноты ТБД, характеризуемой числом отобранных из баз данных ПИФ в ТБД документов заданной тематики, введем следующие параметры и характеристики тематических поисковых запросов, реализуемых на канонических структурах ТЭБД и баз данных ПИФ.

Каждый тематический поисковый запрос $q_\mu \in Q$, где $Q = \{q_\mu / \mu = \overline{1, \mu_0}\}$ — множество тематических запросов, формализовано описывается поисковым предписанием π_k. Поисковое предписание π_k задается в виде множества пар $\{(d_l = d_l^{зн}) R (d_{l'} = d_{l'}^{зн})\}$ ({«поисковый признак = значение» R «поисковый признак = значение»}), где $d_l, d_{l'} \in D^{зп}$ — множество информационных элементов, R — логическая операция (AND/OR/NOT и др.), например, МПК = G01N3/32 AND KW (ключевое слово) = материал. При поиске в ТЭБД тематический запрос адресуется во все БД патентной документации стран, входящих в минимум документации PCT, и во все БД непатентной литературы, определенные списком ВОИС.

Число документов (записей, экземпляров объектов данных), находимых в результате реализации множества тематических запросов $Q = \{q_\mu\}$, $\mu = \overline{1, \mu_0}$ на графе объектной канонической структуры ТЭБД $G^{об}_{к.с}(O, \Delta)$, формализовано представим в виде множества $W_\mu = \{w_\mu\}$, $\mu = \overline{1, \mu_0}$, где w_μ — число документов (экземпляров объектов), находимых в БД стран минимума PCT и непатентной литературы по списку ВОИС при реализации μ-го тематического запроса.

Число документов (записей), находимых в результате реализации множества тематических запросов на графах канонических структур баз данных ПИФ $G_k(D, U)$, $k = \overline{1, K_0}$ формализовано представим в виде множества $V_\mu = \{v^k_\mu\}$, $\mu = \overline{1, \mu_0}$, $k = \overline{1, K_0}$, где v^k_μ — число документов (экземпляров объектов), находимых в k-й базе данных ПИФ при реализации μ-го тематического запроса.

Тогда информационная полнота тематической базы данных ПИФ при реализации μ-го тематического запроса определится из выражения:

$$\varepsilon_{\text{инф}}^{\mu} = \frac{1}{w_{\mu}} \sum_{k=1}^{K_0} v_{\mu}^{k}.$$

Вычисляемые по приведенным выше формулам показатели элементной полноты, полноты по связям (отношениям), функциональной и информационной полноты являются относительными величинами и могут представляться в процентах.

ЗАКЛЮЧЕНИЕ

Предложенные методология, модели и методы анализа и оценки полноты ПИФ применялись для формирования, совершенствования и развития евразийского патентного информационного пространства, создаваемого в рамках деятельности международной патентной организации — Евразийского патентного ведомства (ЕАПВ) [4]. На основе полученных количественных значений оценок показателей полноты баз данных ПИФ этого ведомства (структурной, функциональной, структурно-функциональной и информационной) разработаны и внедрены планы мероприятий по повышению их качества и эффективности. В результате их реализации ПИФ пополнился патентной документацией отдельных организаций и стран (Европейского патентного ведомства (ЕПВ), ВОИС, США, Японии, России, документацией стран Евразийской патентной конвенции, Украины, Узбекистана, Грузии). Структуры баз данных ЕПВ, ВОИС, России (а также СССР), стран СНГ дополнились рефератами. Расширены библиографические описания отдельных БД путем включения в них таких элементов, как, например, номер международной заявки, код вида документа, регистрационные номера приоритетных заявок и др. Поисковые средства информационной системы ЕАПВ (ЕАПАТИС) дополнились средствами виртуального доступа и метапоиском во внешних БД патентной документации и непатентной литературы, а также средствами машинного перевода. В настоящее время в ЕАПАТИС поддерживается более 25-ти постоянно пополняемых локальных тематических баз данных патентного информационного фонда ЕАПВ. Общий объем содержащейся в БД информации превышает 70 млн. документов [4].

Выполненные мероприятия позволили не только повысить эффективность и качество ПИФ, но и обеспечить возможность проведения в ЕАПВ полноценных патентных поисков международного типа по евразийским заявкам по ряду тематик, относящимся к компетенции отделов физики и механики, а также по отдельным областям знаний (в частности, пищевая, сельскохозяйственная, медицинская биотехнология, лекарства и медикаменты для терапевтических, стоматологических или гигиенических целей, удобрения, смеси удобрений, ациклические и карбоциклические соединения; химические или физические процессы, аппараты для их проведения), относящимся к компетенции отдела химии и медицины Управления экспертизы ЕАПВ. При этом результаты поиска в тематических базах данных патентного информационного фонда ЕАПВ не уступают, а в ряде случаев превосходят результаты, содержащиеся в отчетах о поиске, представленных международными поисковыми органами.

Самостоятельное проведение патентных поисков по евразийским заявкам позволило значительно сократить расходы ЕАПВ по соответствующим статьям бюджета, повысить эффективность и качество принимаемых экспертами решений по заявкам, а также квалификацию экспертов.

Полученные результаты использовались при формировании и развитии евразийского патентного информационного пространства [4, 6].

ЛИТЕРАТУРА

1. *Patent* Cooperation Treaty (PCT). — Geneva: WIPO, 2000.
2. *Кульба В.В., Сиротюк В.О., Косяченко С.А.* Информационная безопасность патентных ведомств: теория и практика. — М.: ИПУ РАН, 2017. — 166 с.
3. *Материалы* сайта ВОИС. — URL: www.wipo.int (дата обращения: 3.09.2018).
4. *Фаязов Х.Ф., Сиротюк В.О., Овчинников А.В., Бурцев А.Б.* Евразийская патентно-информационная система. — М.: ОАО ИНИЦ «Патент», 2006. — 108 с.
5. *Материалы* сайта Европейской патентной организации. — URL: www.epo.org (дата обращения: 3.09.2018).
6. *Фаязов Х.Ф., Сиротюк В.О., Овчинников А.В., Бурцев А.Б.* Формирование и развитие евразийского патентно-информационного пространства. — М.: ИНИЦ «Патент», 2010. — 124 с.
7. *Стандарты* Всемирной организации интеллектуальной собственности. — М.: ВНИИПИ, 1996.
8. *Мамиконов А.Г., Ашимов А.А., Кульба В.В.* и др. Оптимизация структур данных в АСУ. — М.: Наука, 1988. — 256 с.
9. *Кульба В.В., Ковалевский С.С., Косяченко С.А., Сиротюк В.О.* Теоретические основы проектирования оптимальных структур распределенных баз данных. Сер. «Информатизация России на пороге XXI века». — М.: СИНТЕГ, 1999. — 660 с.
10. *Кульба В.В., Микрин Е.А., Сиротюк В.О., Сиротюк О.В.* Модели и методы проектирования оптимальных структур объектно-ориентированных баз данных в автоматизированных информационно-управляющих системах. Научное издание. — М.: ИПУ РАН, 2005. — 103 с.

Статья представлена к публикации членом редколлегии В.М. Вишневским.

Сиротюк Владимир Олегович — д-р техн. наук, вед. науч. сотрудник, Институт проблем управления им. В.А. Трапезникова РАН, г. Москва, ✉ vsirotyuk@ipu.ru.