DOI:

ФОРМАЛИЗОВАННАЯ МЕТОДОЛОГИЯ АНАЛИЗА И СИНТЕЗА ОПТИМАЛЬНЫХ СТРУКТУР ТЕМАТИЧЕСКИХ ПАТЕНТНЫХ БАЗ ДАННЫХ

Сиротюк В.О.

Институт проблем управления им. В.А. Трапезникова РАН, Россия, г. Москва
ул. Профсоюзная д.65
vsirotyuk@ipu.ru

Аннотация: Сформулированы требования к тематическим патентным базам данных (ТПБД) патентного информационного фонда. Рассмотрены особенности проектирования ТПБД в архитектуре федеративных баз данных. Предложены формализованная методология, модели и методы анализа и синтеза объектных канонических структур патентных баз данных, обобщенной канонической структуры ТПБД и оптимальных по различным критериям эффективности логических структур ТПБД. Полученные результаты использовались при создании ТПБД международной патентной организации.

Ключевые слова: патентный информационный фонд (ПИФ); патентная база данных (ПБД); тематическая патентная база данных (ТПБД); федеративная база данных; тематический патентный поиск; объектная каноническая структура ПБД; обобщенная каноническая структура ТПБД; логическая структура ТПБД.

Введение

Тематические патентные базы данных (ТПБД) играют важную роль при проведении патентных исследований. Они формируются на основе тематических подборок патентной документации, извлекаемой по запросам пользователей из локальных и внешних источников патентной документации (патентных баз данных - ПБД) патентного информационного фонда (ПИФ). ТПБД представляет собой упорядоченную, структурированную и систематизированную совокупность взаимосвязанных данных (файлов) патентных документов ПИФ определенной тематики (области знаний). В общем случае в ТПБД может включаться также научно-техническая информация, отбираемая из соответствующих источников (из электронных библиотек, с сайтов научно-технических изданий, из БД журналов и т.п.- в дальнейшем из БД НТИ). ПБД являются первичными, в ТПБД - вторичными массивами патентной документации [1].

Пользователями ТПБД являются как эксперты и сотрудники патентных ведомств, осуществляющие рассмотрение заявок на изобретения и выдачу патентов, так и многочисленная аудитория внешних пользователей – хозяйствующих субъектов (ученых, исследователей, организаций, предприятий и т.д.), проводящих патентные исследования при выполнении НИР и ОКР.

Несмотря на гетерогенность источников патентной информации, сведения из ТПБД должны предоставляться пользователям в унифицированном виде. Это вызывает необходимость интеграции данных, извлекаемых из ПБД. Особую актуальность при этом приобретает интеграция данных на логическом уровне представления ТПБД, обеспечивающем возможность доступа к данным, содержащимся в ПБД, в нотациях единой логической структуры (схемы) ТПБД, которая описывает их совместное представление с учетом структурных и поведенческих свойств и ограничений данных.

В работе предложена формализованная методология оптимального проектирования структур ТПБД, создаваемых в архитектуре федеративных БД. Предложенная методология базируется на комплексе взаимосвязанных моделей и методов построения канонических структур ПБД и обобщенной канонической структуры ТПБД, синтеза оптимальных логических структур ТПБД.

1 Архитектуры, методы и технологии создания, организации и реализации ТПБД

ТПБД формируются в результате проведения тематических поисков в отобранных в соответствии с регламентом поиска ПБД и БД НТИ.

ТПБД могут создаваться централизовано в виде сетевых БД (СБД) (в соответствии с классификацией БД, приведенной в [2]) на основе виртуализации рабочих станций (серверов) или в составе облачных сервисов [3], а также децентрализовано в виде федеративных БД [4]. Централизация предполагает физическую интеграцию данных в едином хранилище данных с использованием методов и ИТ консолидации данных, извлекаемых при поиске из гетерогенных источников (ПБД и БД НТИ), а децентрализация - логическую интеграцию данных, физически хранимых в ПБД и БД НТИ источников, с использованием методов и ИТ федерализации данных. Независимо от выбранной информационно-технологической инфраструктуры (платформы) ТПБД основным их назначением является полное и оперативное удовлетворение информационных потребностей пользователей и приложений в патентной и научно-технической информации определенной тематики. Ввиду этого, особую важность

для ТПБД имеют как эксплуатационные характеристики их использования (время доступа к ПБД и БД НТИ, поиска данных, объемы хранимой информации и т.п.), так и вопросы повышения их эффективности и качества (полноты, достоверности, актуальности, защиты и т.п.) [1].

Решение задачи выбора среды формирования централизованных БД (создаваемых в составе ЛВС в виде СБД или «вынесение» БД в облако, в частности, создание БД в составе «частного облака») с использованием методов кластерного анализа предметных областей пользователей, рассмотрено в работах [3,5,6]. Там же приводятся методы построения канонической структуры облачной БД (ОБД), а также модели и методы разработки логической структуры базы метаданных (БмД) репозитория ОБД и оптимального управления изменениями ОБД. Эти методы могут также использоваться при выборе среды формирования и проектировании структур ТПБД в облачной среде. Однако, при этом следует учитывать, что для создания облачных ТПБД потребуется консолидация данных на серверах провайдера облачных сервисов, что не всегда осуществимо с точки зрения возможности переноса данных ПБД и БД НТИ из соответствующих источников информации, владельцами которых в силу принятой у них политики информационной безопасности данные процедуры либо запрещены, либо ограничены по составу, структуре и объему предоставляемых для копирования данных.

Что касается проектирования ТПБД в виде СБД, то в зависимости от выбранной топологии ЛВС ТПБД могут создаваться в различных архитектурах. Одним из вариантов ТПБД являются СБД, создаваемые для многотерминальных сетей, в которых все информационное и программное обеспечение полностью размещается на выделяемых для этих целей мини-ЭВМ, а пользовательские места оснащаются цифровыми и графическими терминалами. Этот вариант характеризуется жесткой централизацией хранения информационных ресурсов и программных средств и ограниченностью использования вычислительных ресурсов, которые обусловлены только мощностью центральной мини-ЭВМ. Для ЛВС с выделенным сервером ТПБД размещается на компьютере-сервере и к ней обеспечивается одновременный доступ с компьютеров-клиентов. При этом ТПБД имеет архитектуру типа "Клиент-Сервер", которая в свою очередь, может представляться в виде 3-х моделей: модели доступа к удаленным данным (т.н. RDA-Remote Date Access), модели сервера базы данных (т.н. DBS-Data Base Server) и модели сервера приложений (т.н. AS-Application Server). Каждая из вышеупомянутых моделей имеет свои особенности, связанные, в основном, с распределением компонентов приложений (компонента представления, прикладного компонента и компонента доступа к информационным ресурсам) между компьютерами-клиентами и компьютером (компьютерами)-сервером (серверами), но основополагающим моментом во всех моделях является то, что сама ТПБД централизовано хранится на компьютере-сервере, оказывающем услуги по обслуживанию запросов пользователей.

Модели и методы проектирования оптимальных логических структур сетевых БД рассмотрены в работах [2,8]. Данные модели и методы целесообразно использовать при синтезе логических структур ТПБД, создаваемых на основе только локальных (внутренних) ПБД ПИФ, например, при проектировании ТПБД, формируемых на основе проведения поиска по одной или нескольким локальным ПБД и БД НТИ, полнота которых и ретроспектива удовлетворяют требованиям регламента поиска. Для использования данных методов и алгоритмов на начальном этапе потребуется объединение канонических структур отобранных в соответствии с регламентом поиска локальных ПБД и БД НТИ в единую интегрированную (каноническую) структуру ТПБД. Для этого могут быть применены методы, рассмотренные в [2].

Рассмотрим особенности проектирования ТПБД в архитектуре федеративных БД (ФБД) [4,7]. Федеративный подход к созданию ТПБД позволяет повысить полноту, эффективность и качество информационных поисков, поскольку создаваемая на его основе ТПБД включает данные, получаемые не только из локальных, но и из внешних удаленных БД, входящих в состав ПИФ.

Формирование ТПБД в архитектуре ФБД имеет ряд преимуществ по сравнению с централизованным подходом к организации и хранении ТПБД. Во-первых, созданные на основе этой технологии ТПБД всегда обращаются за данными только к первоисточникам патентной и научно-технической информации, что гарантирует согласованность и непротиворечивость данных, полноту и актуальность данных, минимизацию количества ошибок в данных, т.к. данные из источников никуда не перемещаются, и за их качество отвечает владелец ПБД (БД НТИ). Во-вторых, федеративный подход позволяет легко расширять состав используемых первоисточников патентной и научно-технической информации, что важно с точки зрения полноты информационного поиска, эффективности и качества ПИФ. В-третьих, ввиду гетерогенности источников патентной и научно-технической информации создание ТПБД в архитектуре ФБД не требует от разработчика использования сложных процедур преобразования БД к единому формату и структуре (т.е.

консолидации данных) и позволяет предоставлять данные пользователю в формате и структуре первоисточника. Основным недостатком федеративного подхода при создании ТПБД является увеличение времени ее формирования в связи с необходимостью доступа к многочисленным ПБД и БД НТИ и их обработки с целью извлечения требуемой в запросах информации. Также считается, что федерализация данных, как метод интеграции данных, не очень хорошо подходит при извлечении и согласовании больших массивов данных [7]. Однако этот фактор не критичен для ТПБД, т.к. количество отбираемых записей из ПБД и БД НТИ должно быть обозримым для пользователя, принимающего решения, а это, в свою очередь, обеспечивается уточнением и последующей корректировкой поисковых запросов. В то же время, ТПБД, создаваемые в архитектуре ФБД, позволяют выявить проблемы с качеством данных отдельных ПБД и БД НТИ ПИФ, а также обеспечить соблюдение политики безопасности данных и лицензионных ограничений, установленных их владельцами, запрещающих (ограничивающих) копирование данных из БД.

Создание ТПБД в архитектуре ФБД предполагает проектирование виртуальной ТПБД, логическая структура (глобальная схема) которой обеспечивает единый интерфейс доступа пользователей к отобранным в соответствии с регламентом поиска локальным и внешним удаленным ПБД и БД НТИ, скрывающий от пользователей особенности обращения к каждому источнику данных. Глобальная схема ТПБД интегрирует схемы ПБД и БД НТИ (канонические и логические структуры), из которых извлекаются данные для ТПБД. Виртуализация данных обеспечивает представление данных в абстрактном виде, независимо от структуры, характеристик и систем управления ПБД (БД НТИ), что позволяет унифицировать данные из нескольких источников в рамках логической структуры ТПБД. С учетом этих особенностей логическая структура ТПБД должна проектироваться на основе обобщенной канонической структуры ТПБД.

Основные положения формализованной методологии анализа и синтеза оптимальных структур ТПБД

Разработанная методология анализа и синтеза оптимальных структур ТПБД в архитектуре ФБД базируется на комплексе формализованных моделей, методов и алгоритмов анализа предметной области ПИФ, построения канонических структур ПБД и обобщенной канонической структуры ТПБД, синтеза по заданным критериям эффективности логических структур ТПБД. Разработанные модели, методы и алгоритмы обеспечивают:

- проектирование канонических структур ПБД ПИФ;
- проектирование обобщенной канонической структуры ТПБД;
- синтез оптимальной логической структуры ТПБД.

Рассмотрим особенности предлагаемой формализованной методологии. Логическая структура ТПБД с учетом отмеченных выше особенностей проектирования ТПБД в архитектуре ФБД должна проектироваться на основе обобщенной объектной канонической структуры ТПБД, обеспечивающей логическую интеграцию данных ПБД. Для ее построения используются методы объектно-ориентированного анализа и проектирования [2,8], т.к. они наиболее адекватно отражают технологию формирования и хранения патентной информации в ТПБД и предоставляемых на их основе услуг. Проектирование канонических структур ПБД и обобщенной канонической структуры (ОКС) ТПБД осуществляется на основе формализованного описания объектной модели предметной области ПИФ и формализованных объектных моделей требований пользователей, методы построения которых рассмотрены в работах [1,2,8]. Для построения ОКС ТПБД используется подход «от предметной области», суть которого заключается в том, что проектирование ТПБД на концептуальном уровне осуществляется на основе анализа общих системных требований. Например, для ПБД таковыми являются требования, предъявляемые стандартами ВОИС к структуре патентной документации, требования и рекомендации цифровых библиотек интеллектуальной собственности (IPDL-Intellectual Property Digital Library) к инструментально-программным средствам поиска, доступа к данным и сервисным средствам, требуемым для проведения пользователями эффективных патентных поисков соответствующих требованиям поисков международного типа [1]. Данные требования являются общими (типовыми) для разных областей знаний (тематик), по которым проводятся поиски международного типа. Поэтому сформированная ОКС ТПБД является типовой для разных тематик.

Синтез оптимальных логических структур ТПБД рассматривается как процесс поиска оптимального варианта отображения свойств и характеристик объектной модели предметной области ПИФ, ОКС ТПБД, а также характеристик канонических структур отобранных для проведения тематического поиска ПБД в логические структуры ТПБД. Основными критериями эффективности синтеза оптимальных логических структур ТПБД являются минимум суммарного времени обслуживания множества тематических запросов пользователей ПИФ, минимум суммарной длины

путей доступа к данным. Ограничениями задач синтеза являются структурные ограничения и ограничения по эффективности использования вычислительных ресурсов при эксплуатации ТПБД.

Модели и методы построения канонических структур ПБД и обобщенной канонической структуры ТПБД

Построение объектных канонических структур ПБД осуществляется с использованием методов и алгоритмов, предложенных в работах [2,8]. Кратко изложим их суть. Исходными данными для построения канонической структуры отдельной v-й ПБД являются:

- формализованное описание модели предметной области ПИФ,
- формализованные описания информационных требований пользователей v-й ПБД, задаваемые в виде в виде матриц смежности $\{B_k\}$ и графов $G_k^{\text{in}}(D_k, R_k)$, $k = \overline{1, K_0}$;
- формализованные описания функциональных требований пользователей v-й ПБД, задаваемые в виде в виде матриц технологии обработки структурных элементов $\{W_k = \|w_{ij}^k\|\}$, $k = \overline{1, K_0}$.

Построение канонической структуры ПБД осуществляется в следующем порядке.

На первом этапе производится нормализация информационных структур пользователей. Процедуры данного этапа обеспечивают минимальную избыточность и дублируемость данных и связей, выделение ключей и атрибутов объектов данных и приведение информационных структур пользователей $G_k^{\text{in}}(D_k, R_k)$ к нормализованному виду – графам $G_k^{nor}(D_k, R_k)$. Для выполнения процедур данного этапа применяются методы и алгоритмы, предложенные в [2].

На втором этапе для каждого объекта данных $d_\varepsilon^k \in D_k$ графа $G_k^{nor}(D_k, R_k)$ на основании технологической матрицы $W_k = \|w_{ij}^k\|$ определяется принадлежность процедур поиска и обработки патентной информации объектам данных, т.е. определяются их информационный $H(d_\varepsilon^k) = \{d_l^k / l \in L_{el}\}$ и функциональный $H(d_\varepsilon^k) = \{h_j^k / j \in J\}$ составы, где $H = \{h_j / j = \overline{1, J}\}$ - множество процедур, $D = \{d_l / l = \overline{1, L}\}$ - полное множество структурных элементов предметной области (объектов данных и информационных элементов).

На третьем этапе осуществляется построение канонической структуры ПБД путем последовательного объединения нормализованных информационных структур пользователей $G_k^{nor}(D_k, R_k)$.

В результате выполнения данных процедур каноническая структура v-й ПБД представляется в виде графа $G_v(D_v, R_v)$, где $D_v = \left\{d_\varepsilon / \ \varepsilon \in L_v^{ob}, L_{ob}^v \subseteq L_v\right\}$ – множество классов (объектов) данных v-*й* ПБД, R_v– множество взаимосвязей (отношений) между элементами. Каждый объект $d_\varepsilon \in D_v$ характеризуется множеством образующих его информационных элементов (ключей и атрибутов данных) $D_l = \left\{d_l / \ l \in L_v^{el}\right\}$ и функций (процедур (методов) поиска и обработки данных) $H_l^{pr} = \{h_j / \ j \in J_v\}$; $R_v = R_v^1 \cup R_v^2 \cup R_v^3$ - полное множество взаимосвязей между структурными элементами канонической структуры v-*й* ПБД где $R_v^1 = \left\{\left(d_\varepsilon, d_{\varepsilon'}\right) / \ \varepsilon, \varepsilon' \in L_v^{ob}\right)\right\}$ –множество взаимосвязей между объектами данных, $R_v^2 = \left\{\left(d_l, d_{l'}\right) / \ l, l' \in L_v^{el}\right\}$ – множество взаимосвязей между ключами и атрибутами объектов данных, $R_v^3 = \left\{\left(h_j, d_\varepsilon\right) / \ \varepsilon \in L_v^{ob}, j \in J_v\right\}$ - множество взаимосвязей между процедурами поиска и обработки данных и используемыми ими объектами данных. Формализовано граф $G_v(D_v, R_v)$ описывается матрицей смежности $W_v = \|w_{\varepsilon\varepsilon'}^v\|$, элементы которой $w_{\varepsilon\varepsilon'}^v = 1$, если между объектами d_ε и $d_{\varepsilon'}$ имеется информационная или функциональная взаимосвязь и $w_{\varepsilon\varepsilon'}^v = 0$, в противном случае.

Рассмотрим методы и алгоритмы построения обобщенной канонической структуры ТПБД.

Под обобщенной (типовой) канонической структурой ТПБД будем понимать минимальную, не содержащую дублируемых элементов и избыточных взаимосвязей структуру данных, описывающую

предметную область ТПБД и представляемую в виде классов и объектов данных предметной области и отношений между ними.

Построение ОКС ТПБД осуществляется в 4 этапа.

На первом этапе производится построение модели предметной области ПИФ, моделей спецификаций информационных и функциональных требований пользователей и объектно-ориентированных (объектных) моделей требований пользователей ПБД. Методы их построения рассмотрены в [2,8]. Особенностью данного этапа в отличие от процедур, рассмотренных для построения канонических структур ПБД, является учет и анализ общесистемных требований, предъявляемых соответствующими стандартами к структуре патентной и непатентной документации, а также требований пользователей, проводящих патентные поиски международного типа, к инструментально-программным средствам поиска, доступа к данным и сервисным средствам [1].

На втором этапе производится объединение объектных моделей требований пользователей в единую обобщенную структуру ТПБД.

На третьем этапе производится нормализация обобщенной структуры ТПБД, в результате которой осуществляется построение безызбыточной (минимальной) объектной канонической структуры ТПБД путем сведения многообразия объектных моделей требований пользователей, зафиксированных в обобщенной объектной структуре пользователей ТПБД, к базовым и специфическим классам объектов. Для выполнения процедур данного этапа используется методы, предложенные в [2,8].

На четвертом этапе формируется ОКС ТПБД, формализовано представляемая в виде графа $G_{кс}^{об}(O,\Delta)$, вершинами которого $O=\{O_{\varepsilon} / \varepsilon=\overline{1,\varepsilon_{ob}}\}$ являются классы и объекты данных предметной области, а дугами $\Delta=\{\delta_{\varepsilon\varepsilon'} / \varepsilon,\varepsilon'=\overline{1,\varepsilon_{ob}}\}$ - связи (или отношения) между классами и объектами данных. ОКС ТПБД описывается матрицей смежности $B_{кс}^{об}=\left\|b_{\varepsilon\varepsilon'}\right\|$ между объектами, составами объектов $H(O_{\varepsilon})=\{d_l,...,d_L,(d_l,d_j),...,(d_k,d_l),\{f_r^{\varepsilon}\}\}$. На ОКС ТПБД выделены множества базовых $O_{баз}$ и специфических $O_{спец}$ объектов. Характеристиками графа $G_{кс}^{об}$ являются интегральные характеристики классов (объектов) и связей (отношений) между ними. При этом свойства класса определяются включенными в его состав объектами данных и информационными элементами, а поведение - методами (функциями) поиска и обработки данных, среди которых выделено подмножество интерфейсных и подмножество реализационных процедур.

Следует отметить, что сформированная объектная каноническая структура ТПБД покрывает канонические структуры ПБД ПИФ, т.е. $G_v(D_v,R_v) \subseteq G_{кс}^{об}(O,\Delta), v=\overline{1,V_0}$. Это означает, что ОКС ТПБД описывает все поведенческие, структурные и информационные аспекты и характеристики, присущие каноническим структурам ПБД ПИФ (локальным и внешним).

Модели и методы синтеза оптимальных логических структур ТПБД

Рассмотрим модели и задачи синтеза оптимальных логических структур ТПБД, создаваемых в архитектуре ФБД, возникающие на этапе их технического проектирования. На данном этапе на основе характеристик объектной модели предметной области ПИФ, обобщенной канонической структуры ТПБД, а также канонических структур подмножества ПБД, отобранных в соответствии с регламентом патентного поиска, характеристик и параметров тематических запросов формируется эффективная объектно-ориентированная (объектная) логическая структура ТПБД, которая должна удовлетворять следующим требованиям:

- обеспечить неизменность свойств и характеристик классов и объектов данных, методов и процедур предметной области ПИФ;
- обеспечить сохранение семантических свойств информационных элементов предметной области, информационных и функциональных связей между ними, зафиксированных в ОКС ТПБД и объектных канонических структурах ПБД;
- учитывать возможности и ограничения информационной инфраструктуры ПБД, интеграционного ПО ФБД, требования различных режимов и возможностей функционирования распределенной информационно-управляющей структуры ПИФ;
- обеспечивать удобство и простоту формирования тематических запросов к ПБД;
- являться оптимальной по заданному критерию эффективности.

Синтез логических структур ТПБД рассматривается как процесс поиска оптимального варианта отображения подструктур (подсхем) ОКС ТПБД, соответствующих отобранным для поиска каноническим структурам ПБД, характеристик объектной модели предметной области ПИФ в такие

структуры, которые обеспечивают оптимальное значение заданного критерия эффективности их использования и удовлетворяют основным требованиям и ограничениям, накладываемым на логическую структуру. Создаваемая при этом логическая схема ТПБД является для пользователей виртуальной структурой, оперирующей выборками данных из ПБД. Она формируется из элементов ОКС ТПБД с учетом метаданных о канонических структурах ПБД и модели предметной области ПИФ при выполнении тематических запросов и обеспечивает сокрытие от пользователей свойств, структур и характеристик самих ПБД. Логическая структура ТПБД представляет собой интерфейс для доступа к ПБД. При этом физический доступ к ПБД обеспечивается с сервера ФБД, хранящего информацию (базу метаданных) о логической структуре ТПБД, свойствах и характеристиках используемых ПБД и предметной области ПИФ, запросах пользователей.

Следует отметить, что ОКС ТПБД используется при создании ТПБД разных тематик. При этом отдельная логическая структура ТПБД формируется в соответствии с требованиями k-го тематического запроса. На физическом уровне ТПБД представляется в виде витрины данных и содержит информацию по определенной тематике (области знаний).

При синтезе логических структур ТПБД в архитектуре ФБД требуется определить отображение:

$$\{G(D,R), G_{кс}^{об}(O,\Delta), \{G_v(D_v,R_v), v \in V_k \subseteq V_0\}\} \xrightarrow{\theta^*} G(N,L),$$

где $V_k \subseteq V_0$ - подмножество ПБД, отобранных в соответствии с регламентом патентного поиска *k-го* тематического запроса; *G(D,R)* - граф объектной модели предметной области ПИФ ($D=\{d_i/i=\overline{1,n_0}\}$- полное множество структурных элементов предметной области ПИФ (объектов данных и информационных элементов), $R=\{(d_i,d_{i'})/i,i'=\overline{1,n_0}\}$- полное множество информационных и функциональных взаимосвязей (отношений) между элементами); $G_{кс}^{об}(O,\Delta)$-граф ОКС ТПБД ($O=\{o_\varepsilon / \varepsilon=\overline{1,\varepsilon_0}\}$- множество классов объектов, $\Delta=\{\delta_{\varepsilon\varepsilon'} / \varepsilon,\varepsilon'=\overline{1,\varepsilon_0}\}$ - множество связей между классами); $G_v(D_v,R_v)$- граф объектной канонической структуры v-*й* ПБД ($D_v=\{d_\varepsilon / \varepsilon \in L_v\}$ – множество классов (объектов) данных v-*й* ПБД, $R_v=\{(d_\varepsilon,d_{\varepsilon'})/\varepsilon,\varepsilon' \in L_v\}$ - множество взаимосвязей между классами (объектами) данных v-*й* ПБД), *G(N,L)* - граф логической структуры ТПБД, где $N=\{n_j/j=\overline{1,J}\}$ - множество логических записей, $L=\{(n_j,n_{j'})/j,j'=\overline{1,J}\}$ - множество взаимосвязей между записями. Отображение предполагает объединение элементов множества $O=\{o_\varepsilon / \varepsilon=\overline{1,\varepsilon_0}\}$ в элементы множества $N=\{n_j/j=\overline{1,J}\}$ и множества связей $\Delta=\{\delta_{\varepsilon\varepsilon'} / \varepsilon,\varepsilon'=\overline{1,\varepsilon_0}\}$ в множество связей (внешних ключей) $L=\{(n_j,n_{j'})/j,j'=\overline{1,J}\}$, при котором достигается оптимальное значение некоторой целевой функции Ω на множество альтернативных вариантов отображения θ. При этом должны учитываться параметры и характеристики объектной модели предметной области ПИФ *G(D,R)*, запросов пользователей, а также выполняться ограничения и обеспечиваться получение оптимального значения заданного критерия эффективности.

Основными критериями эффективности синтеза оптимальных логических структур ТПБД являются: минимум суммарного времени обслуживания множества тематических запросов пользователей ПИФ, минимум суммарной длины путей доступа к данным.

Ограничениями задач синтеза являются ограничения двух видов - структурные ограничения и ограничения, связанные с необходимостью эффективного использования вычислительных ресурсов при эксплуатации создаваемой ТПБД.

К первому типу относятся ограничения на число и состав логических записей; на структуру связей между ними; на число точек входа в канонические структуры ПБД; на число объектов в составе отдельной записи. Ко второму типу относятся ограничения на сложность связей между записями; на время поиска по отдельным запросам и др.

Результатом решения задач синтеза оптимальных логических структур ТПБД является определение числа и состава логических записей, выбор структуры связей между записями, определение оптимальной структуры тематических запросов к ТПБД.

Для формализации задач синтеза рассмотрим логическую схему выполнения тематического запроса. Логическая схема реализации отдельного тематического запроса пользователя включает следующую последовательность операций:

- конструирование с учетом требований регламента поиска и поисковых предписаний структуры запроса и описание его на языке запросов выбранной СУБД (ООСУБД) ФБД,

- обработка запроса соответствующими методами и средствами репозитария ПИФ: выбор из БмД репозитория требуемых в запросе метаданных; декомпозиция запроса на подзапросы (задания), количество которых определяется количеством отобранных для проведения патентного поиска ПБД; определение маршрутов доступа к серверам ПБД; установление центральным сервером ФБД логических соединений с серверами ПБД ПИФ,

- передача требований запроса по каналам связи на сервер (серверы) ПБД (доступ к ПБД),

- обслуживание подзапросов серверами ПБД,

- передача по каналам связи отобранных с серверов ПБД блоков данных на центральный сервер ФБД,

- сборка блоков и логическая интеграция данных в ТПБД на центральном сервере ФБД.

Исходя из логической схемы выполнения запроса, общее время реализации отдельного k-го тематического запроса T_k складывается из следующих основных составляющих:

$$T_k = t_k^{rep} + t_k^{ac} + t_k^{ser} + t_k^{tr} + t_k^{as},$$

где t_k^{rep} - время обработки запроса средствами репозитария ПИФ; t_k^{ac} - время доступа к ПБД; t_k^{ser} - время обслуживания запроса на сервере ПБД (поиска и выборки данных); t_k^{tr} - время передачи отобранных блоков данных на центральный сервер ФБД; t_k^{as} - время сборки блоков данных на сервере ФБД, логической интеграции данных и формирования ТПБД.

Как видно из представленной схемы выполнения запросов, основой их реализации являются структуры поиска запрашиваемых данных. Структура запроса к ТПБД формально представляется деревом поиска, задаваемом на графе ОКС ТПБД $G_{кс}^{об}(O,\Delta)$, алгоритмы построения которых рассмотрены в работе [2]. Прохождение каждого пути, входящего в состав дерева поиска, из l-й точки входа в ОКС ТПБД сопровождается просмотром указателей связей (внешних ключей) между объектами и выдачей информации из объектов классов, задаваемых условиями поиска. Задание условий поиска определяет направление и стратегию обхода вершин дерева поиска.

Тематические запросы множества $З = \{з_k / k = \overline{1, K_0}\}$ характеризуются составами запрашиваемых ими классов (объектов данных) и связей ОКС ТПБД и формально описываются матрицами $A = \|a_{\varepsilon k}\|$ и $F = \left\| f_{\varepsilon\varepsilon'}^{k} \right\|$ соответственно, а также множеством частот использования $Q = \{q_k / k = \overline{1, K_0}\}$.

Количественными характеристиками реализации отдельно *k-го* тематического запроса, формально представленного в виде дерева поиска на графе $G_{кс}^{об}(O,\Delta)$ и реализуемого через *l-ю* точку входа, являются: множество средних значений суммарного числа выбираемых при реализации k-го запроса из l-й точки входа объектов классов $P^{з} = \{p_{\varepsilon}^{kl} / a_{\varepsilon k} = 1\}$, где p_{ε}^{kl} – число выбираемых экземпляров объекта o_{ε}, входящего в путь доступа из *l-й* точки входа *k-го* запроса; множество средних значений суммарного числа просматриваемых при реализации k-го запроса из l-й точки входа внешних ключей $\beta^{з} = \{\beta_{\varepsilon\varepsilon'}^{kl} / f_{\varepsilon\varepsilon'}^{k} = 1\}$, где $\beta_{\varepsilon\varepsilon'}^{kl}$ – среднее суммарное число просматриваемых указателей ($\varepsilon\varepsilon'$)-й связи, входящей в путь доступа из *l-й* точки входа *k-го* запроса.

Временными параметрами запроса являются: $t_{\varepsilon\varepsilon'}$ - среднее время просмотра указателей ($\varepsilon\varepsilon'$)-й связи; t_{ε} - среднее время поиска и выбора объекта ε-го класса из ПБД; $t_v^{д}$ - среднее время доступа к *v-й* ПБД; $t^{сб}$ - среднее время сборки блока данных при формировании логической записи на сервере ФБД; $t_v^{пер}$ - среднее время передачи блока данных с сервера *v-й* ПБД на сервер ФБД; $t^{экз}$ - среднее время загрузки в ТПБД экземпляра объекта данных; $t^{рет}$ - среднее время обработки запроса методами и средствами репозитария. Методы расчета количественных и временных характеристик запросов, зависящие от выбранной стратегии и варианта поиска, организации указателей связей и других характеристик, приведены в [2].

Для постановки и решения задач синтеза оптимальных логических структур ТПБД дополнительно используется следующая информация:

- сведения о принадлежности объектов данных канонических структур ПБД классам (объектам) ОКС ПИФ. Данная информация формализуется с помощью матрицы смежности $M = \|m_{\varepsilon\varepsilon'}\|$, $\varepsilon = \overline{1,\varepsilon_0}$, $\varepsilon' \in L_v^{ob}$, $v = \overline{1,V_0}$, элементы которой $m_{\varepsilon\varepsilon'} = 1$, если объект $d_{\varepsilon'} \in D_v$ описывается на графе ОКС ПИФ классом (объектом) $o_\varepsilon \in O$, $m_{\varepsilon\varepsilon'} = 0$, в противном случае;

- информация об используемых при проведении тематических патентных поисков ПБД, формализуемая с помощью матрицы смежности $N = \|n_{kv}\|$. Элементы матрицы $n_{kv} = 1$, если в соответствии с регламентом поиска для выполнения *k-го* запроса требуется обращение к *v-ой* ПБД, и $n_{kv} = 0$, в противном случае;

- сведения о зафиксированных в ПБД классах (объектах) данных, формализуемые с помощью матрицы $\Delta = \|\delta_{\varepsilon v}\|$, элементы которой $\delta_{\varepsilon v} = 1$, если объект $d_\varepsilon \in D_v$ и $\delta_{\varepsilon v} = 0$, если $d_\varepsilon \notin D_v$;

- информация о количестве экземпляров классов (объектов) ПБД - элемент π_ε, который характеризует количество экземпляров класса (объекта) d_ε

Рассмотрим модели синтеза оптимальной логической структуры ТПБД, формируемой для *k-го* тематического запроса.

Для формализации задачи синтеза введем следующие переменные:

$x_{\varepsilon j}$=1, если ε-й класс (объект) данных ОКС ТПБД используется при формировании *j*-й логической записи ТПБД; $x_{\varepsilon j}$=0, в противном случае.

$$z_{kj} = 1, \text{ если } \sum_{\varepsilon=1}^{\varepsilon_0} x_{\varepsilon j} a_{\varepsilon k} \geq 1;\ z_{kj} = 0, \text{ если } \sum_{\varepsilon=1}^{\varepsilon_0} x_{\varepsilon j} a_{\varepsilon k} \geq 0.$$

Переменная z_{kj} определяет необходимость обращения *k-го* тематического запроса к *j-й* записи логической структуры ТПБД.

$$z_{jv} = 1, \text{ если } \sum_{\varepsilon=1}^{\varepsilon_0} x_{\varepsilon j} \delta_{\varepsilon v} \geq 1'\ z_{jv} = 0, \text{ если } \sum_{\varepsilon=1}^{\varepsilon_0} x_{\varepsilon j} \delta_{\varepsilon v} = 0$$

Переменная z_{jv} определяет формирование *j-й* логической записи ТПБД из данных *v-й* ПБД.

$$y_{jj'}^{(\varepsilon\varepsilon')} = 1, \text{ если } x_{\varepsilon j} x_{\varepsilon' j'} b_{\varepsilon\varepsilon'} = 1;\ y_{jj'}^{(\varepsilon\varepsilon')} = 0, \text{ если } x_{\varepsilon j} x_{\varepsilon' j'} b_{\varepsilon\varepsilon'} = 0.$$

Переменная $y_{jj'}^{(\varepsilon\varepsilon')}$ определяет необходимость использования внешних ключей (связей) между формируемыми логическими записями n_j и $n_{j'}$ ТПБД на основе информации о связности объектов o_ε и $o_{\varepsilon'}$ ОКС ТПБД, вошедших в запись.

$$y_{jj'} = 1, \text{ если } \sum_{\varepsilon,\varepsilon'=1}^{\varepsilon_0} y_{jj'}^{(\varepsilon\varepsilon')} \geq 1';\ y_{jj'} = 0, \text{ если } \sum_{\varepsilon,\varepsilon'=1}^{\varepsilon_0} y_{jj'}^{(\varepsilon\varepsilon')} \geq 0.$$

Переменная $y_{jj'}$ определяет наличие или отсутствие связей между записями логической структуры ТПБД.

$z_{kl} = 1$, если для k-го запроса выбирается l-я точка входа в ОКС ТПБД; $z_{kl} = 0$ в противном случае.

Общая задача синтеза оптимальной логической структуры ТПБД, создаваемой в архитектуре ФБД, по критерию минимума общего суммарного времени обслуживания тематического запроса формулируется следующим образом:

$$(1)\quad \min_{\{x_{\varepsilon j},z_{kl}\}}\sum_{j=1}^{J}\{q_k[(\sum_{v\in V_k}z_{jv}(t_v^{\partial}+t_v^{nep})+z_{kj}t^{pen})+\sum_{l\in L_k}z_{kl}\sum_{\varepsilon=1}^{\varepsilon_0}(\sum_{j'\neq j}^{J}\sum_{\varepsilon'\neq\varepsilon}y_{(jj')}^{(\varepsilon\varepsilon')}\beta_{\varepsilon\varepsilon'}^{kl}t_{\varepsilon\varepsilon'}+x_{\varepsilon j}p_{\varepsilon}^{kl}t_{\varepsilon})]+$$

$$+(\sum_{\varepsilon=1}^{\varepsilon_0}x_{\varepsilon j}-1)t^{cб}+\sum_{\varepsilon=1}^{\varepsilon_0}x_{\varepsilon j}\pi_{\varepsilon}t^{экз}\}$$

при ограничениях:

- на однократность включения классов (объектов) в логическую запись

$$(2)\quad \sum_{j=1}^{J}x_{\varepsilon j}=1'\quad \varepsilon=\overline{1,\varepsilon_0}$$

- на число классов (объектов) в составе логической записи

$$(3)\quad \sum_{\varepsilon=1}^{\varepsilon_0}x_{\varepsilon j}\le N'\ j=\overline{1,J}$$

где E - максимально допустимое число классов (объектов) в логической записи;

- на общее число типов логических записей в структуре

$$(4)\quad J\le H\ ,$$

где H - максимально допустимое число записей в логической структуре ТПБД;

- на допустимость включения некоторых классов (объектов) канонических структур ПБД в состав одной логической записи ТПБД

$$(5)\quad x_{\varepsilon j}+x_{\varepsilon' j}\le 1\ \textit{для заданных}\ \varepsilon,\varepsilon',\ j=\overline{1,J}\ ;$$

- на количество внешних ключей в логической структуре ТПБД

$$(6)\quad \sum_{j'\neq j}^{J}y_{jj'}\le N,\qquad j=\overline{1,J}\ ,$$

где N- максимально допустимое число внешних ключей в логической структуре ТПБД;

- на время поиска данных по тематическому запросу

$$(7)\quad \sum_{l\in L_k}z_{kl}\sum_{j,j'=1}^{J}(\sum_{\varepsilon,\varepsilon'=1}^{\varepsilon_0}y_{(jj')}^{(\varepsilon\varepsilon')}\beta_{\varepsilon\varepsilon'v}^{kl}t_{\varepsilon\varepsilon'}+x_{\varepsilon j}p_{\varepsilon}^{kl}t_{\varepsilon})\le T_k\ ,$$

где $T_к$- регламентное время, отводимое для поиска данных по k -му тематическому запросу;

- на единственность точки входа в ОКС (ПБД) по каждому запросу

$$(8)\quad \sum_{l\in L_k^v}z_{kl}=1,\ \textit{для заданных}\ k=\overline{1,K_0}$$

При использовании одной точки входа в каждую каноническую структуру ПБД критерий (3.3.1) примет вид:

$$(9)\quad \min_{\{x_{\varepsilon j}\}}\sum_{j=1}^{J}\{q_k[(\sum_{v\in V_k}z_{jv}(t_v^{\partial}+t_v^{nep})+z_{kj}t^{pen})+\sum_{j'\neq j}^{J}(\sum_{\varepsilon,\varepsilon'=1}^{\varepsilon_0}y_{(jj')}^{(\varepsilon\varepsilon')}\beta_{\varepsilon\varepsilon'}^{kl}t_{\varepsilon\varepsilon'}+x_{\varepsilon j}p_{\varepsilon}^{kl}t_{\varepsilon})]+$$

$$+(\sum_{\varepsilon=1}^{\varepsilon_0}x_{\varepsilon j}-1)t^{cб}+\sum_{\varepsilon=1}^{\varepsilon_0}x_{\varepsilon j}\pi_{\varepsilon}t^{экз}\}$$

Если существуют трудности при определении временных параметров запросов целесообразна постановка задачи синтеза оптимальной логической структуры ТПБД по критерию минимума суммарной длины путей доступа к данным. В этом случае критерий эффективности имеет вид:

$$(10)\ \min_{\{x_{\varepsilon j}\}}q_k\sum_{j=1}^{J}\ (1+\sum_{j'\neq j}^{J}\sum_{\varepsilon'\neq\varepsilon}y_{(jj')}^{(\varepsilon\varepsilon')}\beta_{\varepsilon\varepsilon'}^{k})$$

Поставленные задачи синтеза логических структур ТПБД являются задачами нелинейного целочисленного математического программирования с булевыми переменными. Для их решения могут быть использованы методы и алгоритмы, рассмотренные в [2,8].

В результате решения задач синтеза определяются следующие характеристики ТПБД и тематических запросов пользователей:

- состав типов логических записей ТПБД, формализовано описываемый в виде матрицы смежности $C = \left\| c_{\varepsilon j'} \right\|$, проиндексированной по строкам классами (объектами) множества $O = \{o_\varepsilon / \varepsilon = \overline{1, \varepsilon_0}\}$, а по столбцам – множеством сформированных логических записей $N=\{n_j/j=\overline{1, J}\}$. Элемент $c_{\varepsilon j'}$ =1, если $\varepsilon -$ й объект включен в состав *j-ой* логической записи;

- структура взаимосвязей (отношений) между записями, формализуемая в виде матрицы смежности $B = \left\| b_{jj'} \right\|$, элемент которой $b_{jj'}$ =1, если $c_{\varepsilon j'} \wedge c_{\varepsilon' j'} \wedge b_{\varepsilon\varepsilon'} = 1$, иначе $b_{jj'}$ =0, если $c_{\varepsilon j'} \wedge c_{\varepsilon' j'} \wedge b_{\varepsilon\varepsilon'} = 0$;

- перечень объектов-точек входа в ОКС ТПБД, обеспечивающих кратчайшие пути доступа к данным и минимальное время обработки запросов в ПБД. Данная информация формализуется подмножеством O_{lk}, состоящим из элементов $o_\varepsilon \in O$;

- структура *k-го* тематического запроса пользователей, формализуемая вектором $W_k = \{w_{kj}\}$, элемент которого w_{kj} =1, если *j-я* логическая запись используется в *к-м* запросе;

- источники пополнения логических записей ТПБД. Данная информация формализуется матрицей $M = \left\| m_{jv} \right\|$, элемент которой m_{jv} =1, если для пополнения информацией *j-й* логической записи используются данные *v-й* ПБД, m_{jv} =0, если *v-я* ПБД не участвует в формировании *j-й* записи.

Заключение

Патентная информация является ключевым стратегическим ресурсом высокотехнологических предприятий и организаций XXI века. Централизация хранения патентно-информационных ресурсов в соответствующих патентных БД и БД НТИ ПИФ и децентрализация их использования коллективами пользователей выдвигают проблему построения эффективных структур тематических ПБД, формируемых при проведении патентных поисков для решения определенных научных или прикладных проблем.

В работе предложены формализованная методология, модели и методы анализа и построения канонических структур ПБД и обобщенной (типовой) канонической структуры ТПБД, синтеза оптимальных логических структур ТПБД. Модели и методы синтеза разработаны применительно к ТПБД, создаваемым в архитектуре федеративных БД (ФБД), что обеспечивает ряд преимуществ по сравнению с централизованным подходом к их организации и хранению. Предложенные методология, модели и методы использовались при проектировании ряда ТПБД международной патентной организации - Евразийского патентного ведомства [9]. В частности, при формировании ТПБД в области медицины (ТПБД «Устройства для испытания остроты зрения» (МПК A61B 3/00), ТПБД «Аппаратура для лучевой диагностики» (МПК A61B 6/00), ТПБД «Лекарства и медикаменты для терапевтических, стоматологических или гигиенических целей» (МПК A61K) и др.); в области химии (ТПБД «Фосфорные удобрения» (МПК C05B), ТПБД «Обработка воды, промышленных и бытовых сточных вод» (МПК C02)); в области пищевых продуктов (ТПБД «Хлебопекарные печи» (МПК A21B), ТПБД «Способы приготовления теста и выпечки изделий» (МПК A21D 8/00)) и в других областях знаний. В качестве ПБД при формировании ТПБД использовались локальные и внешние удаленные ПБД евразийского патентного информационного пространства (локальные ПБД ЕАПВ, ПБД РФ, ПБД стран СНГ, ПБД международных заявок, ПБД ЕПВ, ПБД США и др.; внешние ПБД – ИПС Espacenet, PatentScop, ПБД USPTO (патенты), ПБД USPTO (заявки) и др.)[1,9].

Использование предложенных моделей и методов позволило значительно сократить время обслуживания тематических запросов экспертов ЕАПВ и хозяйствующих субъектов при проведении ими патентных поисков, повысить полноту, эффективность и качество ПБД ПИФ.

Литература

1. *Кульба В.В., Сиротюк В.О.* Формализованная методология повышения эффективности и качества патентных информационных фондов и опыт ее использования при формировании и

развитии евразийского патентно-информационного пространства. - М.:ИПУ РАН. Монография, 2019. - 236с.
2. *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.
3. *Kul'ba V.V., Sirotyuk V.O., Sirotyuk O.V.* Models and methods for analyzing and structuring the domains of cloud technologies users. М.: Интернет-журнал «Науковедение», т.9, вып.№5,2017
4. *Бегг К., Коннолли Т.* Базы данных. Проектирование, реализация и сопровождение. Теория и практика, 3-е изд. М: Вильямс, 2017. - 1440 с.
5. *Сиротюк В.О., Косяченко С.А.* Моделирование предметных областей пользователей при использовании облачных технологий. *Вестник* РГГУ. Серия «Экономика. Управление. Право». 2017. №4(10).с.74-87.
6. *Лепе Н.Л., Сиротюк В.О.* Модели и методы управления изменениями облачных баз данных. - Вестник РГГУ. Серия «Экономика. Управление. Право». 2018. №2 (12). С.81-98.
7. *David Waddington* An Architected Approach to Information Integration – Federated Enterprise Data Warehousing Overview-Kalido, 2004. - 23p.
8. *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.
9. Материалы сайта Евразийской патентной организации: www.eapo.org/ru