ISSN 2079-3316 **ПРОГРАММНЫЕ СИСТЕМЫ: ТЕОРИЯ И ПРИЛОЖЕНИЯ** **т. 9**, №4(39), с. 219–237

ББК 5с51
ГРНТИ 76.03.59
УДК 61:004.652

В. Л. Малых, А. Е. Михеев, С. В. Рудецкий

Проблемно-ориентированная модель банка клинических данных

Аннотация. В связи с накоплением в сфере здравоохранения больших клинических данных (БКД) появляется актуальная задача построения проблемно-ориентированных информационных моделей таких данных и формирования банков клинических данных (БКД). Наибольший интерес представляют процессные модели клинических данных, в которых лечебно-диагностический процесс предлагается моделировать дискретным конечномерным управляемым процессом.

Отмечена принципиальная не наблюдаемость вектора состояния объекта управления и его большая размерность. Обоснована необходимость применения технологий Больших Данных для построения БКД. Представлены две реляционные модели клинических данных в процессной форме. Приведен авторский опыт применения одной из представленных реляционных моделей. Отмечена перспективность второй реляционной модели, основанной на реляционной СУБД с хранением данных по колонкам. Результаты работы будут полезны разработчикам информационных систем в сфере здравоохранения.

Ключевые слова и фразы: коллекции данных, банк клинических данных, реляционная база данных, база данных с хранением по колонкам, медицинские информационные системы.

Введение

Значение данных в науке невозможно переоценить, собственно наука начинается со сбора данных. В последнее время вычислительная техника обеспечила возможности накопления и обработки больших данных. И в медицине наряду с технологическим совершенствованием происходит нарастание объемов обрабатываемых и сохраняемых данных, главным образом, благодаря внедрению новых диагностических и лечебных процедур, а также новых информационных технологий, позволяющих хранить и обрабатывать большие массивы данных.

10.25209/2079-3316-2018-9-4-219-237

Современные методы искусственного интеллекта (ИИ), методы машинного обучения все чаще применяются для анализа данных различной природы, в том числе, генерируемых медицинскими информационными системами (МИС) — диагностических изображений, ЭКГ, текстов клинических документов и т.п., и все чаще позволяют получать значимый практический эффект. Можно утверждать, что объемы обрабатываемых в МИС данных будут только возрастать с расширением использования электронного документооборота.

Количество типов данных, с которыми должна работать информационная система, постоянно растет, включая новые типы данных, характерные для персонализированной медицины: результаты молекулярного анализа и параметры ДНК. Учитывая это, можно прогнозировать усиление роли систематической обработки данных, информации и знаний в деле повышения качества, доступности и эффективности медицинской помощи. Необходимым условием для встраивания технологий ИИ в повседневную жизнь, в том числе, в медицинскую практику, является наличие больших очищенных данных для проведения глубокого машинного обучения и семантического анализа.

Если в сфере потребления проблема с разной степенью успешности решается в связи с постепенной цифровизацией экономики, то барьером на пути широкого практического применения современных методов анализа больших данных в российской медицине является отсутствие доступных для исследователей банков больших обезличенных клинических данных. Большинство медицинских записей в медицинских организациях России до настоящего времени ведутся на бумаге. В то же время можно привести примеры таких банков данных за рубежом. Например, в Великобритании по инициативе и с участием 674 врачей общей практики создана база клинических данных, которая покрывает свыше 11,3 млн. пациентов Великобритании [1]. В России подобные прецеденты нам не известны.

Систематический (140 статей + 28 веб-сайтов) актуальный обзор литературы [2] приводит данные о западных проектах по формированию банков клинических данных. В этом направлении лидируют англоязычные страны: Великобритания, США, Канада, Австралия. Число пациентов, данные которых накоплены в БКД, начинается от десятков и сотен тысяч и заканчивается десятками миллионов (18 млн. в проекте Qresearch). Проект CPRD в Великобритании покрывает 20% популяции этой страны (13 млн. человек). Основными «держателями» БКД на Западе являются:

- государственные институты (Veterans Health Administration in the USA, National Health Service in the UK),

- академические институты (Nottingham University, Leuven University, Dutch College of General Practitioners),
- частные компании, специализирующиеся на разработке программного обеспечения, работающие в области извлечения данных и их анализа (IMS Health).

В обзоре [2] приводятся общие сведения об архитектуре сбора, накопления и анализа клинических данных. Основными элементами архитектуры являются централизованные официальные хранилища данных (a centralized official data warehouse), БКД в нашей терминологии. Клинические данные извлекаются из МИС, обрабатываются и помещаются в БКД. Следующими важными элементами архитектуры являются функционально интегрированные платформы (a functional integrated platform), предоставляющие готовые сервисы для отбора данных из хранилищ и сервисы анализа данных (a turnkey service). Важной особенностью архитектуры является способность интегрировать в БКД данные из разных источников (Integrating data linkage functionalities). Остается только отметить значительное превосходство развитых западных стран перед Россией в области формирования публичных БКД.

Реализуемый в России проект создания единой государственной информационной системы в сфере здравоохранения (ЕГИСЗ) пока не ставит перед собой цели создания хранилища полных детальных описаний клинических случаев. Аналогичное можно сказать и про проект московского правительства в сфере здравоохранения — ЕМИАС, а также о других проектах информатизации регионального здравоохранения. Использование зарубежных клинических данных ограничено языковым барьером и тем, что они никак не могут отражать особенности российской медицины.

Единственным источниками клинических данных в России в настоящее время являются МИС. При этом, проектировщики отечественных МИС изначально не ставили перед собой задачи представления клинических данных в форме пригодной для обработки методами искусственного интеллекта, методами машинного обучения. Не ставилась ранее и задача объединения данных из различных источников (МИС).

Следует отметить, что проблема перевода здравоохранения с бумажных технологий на современные цифровые технологии пока еще не решена в достаточной мере во всем мире, включая самые развитые страны, как и не решена проблема интероперабельности различных МИС, не решена проблема стандартизации процессов оказания медицинской помощи, в частности, стандартизации электронной медицинской карты. Но работы в этих направлениях ведутся активно.

В статье на базе доступных авторам источников клинических данных предложена проблемно-ориентированная формализация клинических данных в расчете на применение к данным банка современных методов искусственного интеллекта. Появление в России открытых БКД, предоставляющих открытый доступ к большому объему клинических данных, ускорит внедрение современных методов ИИ в российское здравоохранение, будет стимулировать профильные министерства и ведомства России к сбору и накоплению клинических данных в национальном масштабе, может способствовать реальному повышению качества и эффективности медицинской помощи в России.

1. Работы ИПС им. А.К. Айламазяна РАН в области систем поддержки принятия решений для медицины

Федеральное государственное бюджетное учреждение науки Институт программных систем им. А.К. Айламазяна Российской академии наук (ИПС им. А.К. Айламазяна РАН) почти 25 лет занимается фундаментальными исследованиями и прикладными разработками в области медицинской информатики.

Разработки и результаты научных исследований Института используются в деле охраны здоровья ведущими учреждениями здравоохранения Управления делами Президента РФ, Банка России, Федерального медико-биологического агентства, ОАО «Российские железные дороги», МВД России, ФТС России, научными медицинскими учреждениями, республиканскими и муниципальными больницами, коммерческими медицинскими организациями.

Институтом разработаны полные интегрированные решения для информатизации здравоохранения на базе МИС семейства Интерин, охватывающей все стороны деятельности учреждения, преимущества от внедрения которой получают как сами медицинские организации (МО), так и системы ведомственного, муниципального или регионального здравоохранения, в которых они работают.

Благодаря унификации автоматизируемых бизнес-процессов семейству МИС Интерин удается решить проблему «внутрикорпоративной» интероперабельности и стандартизации электронных медицинских карт. У авторов статьи — исследователей ИПС им. А. К. Айламазяна РАН — имеется доступ к достаточному объему верифицированных обезличенных клинических данных с требуемой полнотой и детализацией. Нельзя упускать эту уникальную как для академического сообщества, так и для России в целом, возможность.

Идеи анализа клинических данных и процессов лечебно-профилактических учреждений с целью выявления закономерностей, выделения

устойчивых последовательностей (прецедентов) и, в конечном счете, моделирования предметной области возникли в ИПС им. А.К. Айламазяна РАН еще в конце 90-х годов прошлого века. Применение прецедентного подхода к анализу и моделированию различных процессов в медицинских организациях, использование прецедентного подхода в системах поддержки принятия врачебных решений (СППВР), стали воплощаться в результаты конкретных исследованиях еще в начале 2000-х годов. В 2009 г. вышли две публикации, посвященные изложению концепции прецедентного подхода и его практического использования в МИС [3, 4].

Полученные результаты оказались столь плодотворны, что было решено расширить сферу применения прецедентного подхода, распространив его на проблему поддержки принятия врачебных решений. В обзоре [5] по теме поддержки принятия врачебных решений также отмечается плодотворность и важность прецедентного подхода: «СППР в сфере здоровья часто основаны на анализе прецедентов и принципах доказательной медицины (case-или evidence-based подходы), при этом используются сведения как из практики, так и из результатов научных исследований».

В ходе решения поставленных задач по СППВР стало ясно, что клинические данные нуждаются в отдельной проблемно-ориентированной формализации в интересах применения методов машинного обучения. Сам же прецедентный подход к СППВР предполагал создание представительного банка клинических данных, но подобных банков в России не было. Все эти проблемы нашли свое отражение в публикациях и докладах 2013–2017 годов [6–12]. На представительных международных конгрессах и конференциях были представлены идеи по целесообразности создания банков больших клинических данных, вплоть до национального масштаба.

Параллельно в 2010-2017 годах шла работа над процессной, основанной на понятии состояния, и событийной формализации клинических данных в форме, подходящей для применения методов машинного обучения [9, 10]. Проводилась оценка эффективности и точности СППВР, построенной на основе различных методов машинного обучения (прецедентный подход, нейронные сети) [11, 12]. В ходе проведенных исследований прецедентный подход доказал свою эффективность и тем самым еще раз подтвердил научную и практическую целесообразность в сборе и накоплении больших клинических данных. Одновременно были выявлены и проблемы при решении задач проектирования и создания хранилищ для больших клинических данных.

2. Постановка задачи

Существуют различные подходы к проектированию хранилищ данных, выбору систем управления базами данных [13–16]. В качестве СУБД для МИС хорошо себя зарекомендовали реляционные СУБД (Oracle, Microsoft, IBM и др.). В рамках реляционных СУБД удается применять не только реляционный, но и объектный подход к проектированию и хранению данных [13, 14], удается сочетать возможности реляционных и NO SQL баз данных [15]. Однако, когда мы переходим к Большим Данным, все значительно усложняется, и «традиционные» реляционные СУБД уже не являются лидерами для построения хранилищ Больших Данных [17–19].

Еще одной проблемой является необходимость перевода данных в проблемно-ориентированную модель. Основой для МИС является понятие медицинского документа. Данные, как правило, возникают в контексте документов: осмотров, дневников, диагностических протоколов и т.п. Формализация документов, особенно в части семантики содержащихся в них данных, вызывает много проблем.

Разработаны различные стандарты для представления моделей документов и для обмена медицинскими данными: HL7, CDA (Clinical Document Architecture), Open EHR, и др. Но все эти представления мало подходят для решения задач интеллектуального анализа данных и машинного обучения. Поэтому, задача построения проблемно-ориентированных хранилищ больших клинических данных, предназначенных для интеллектуальной обработки данных, представляется актуальной.

Далее в работе будут предложены две различные реляционные проблемно-ориентированные модели данных, опирающиеся на понятия управления, процесса и состояния.

3. Лечебно-диагностический процесс (ЛДП) как управляемый процесс смены состояний

Традиционный широко применяемый подход к моделированию процессов опирается на понятие состояния. Процесс рассматривается как смена состояний. Часто удается рассматривать процесс как дискретный, когда состояние меняется в дискретные моменты времени, а само состояние описывается конечномерным вектором характеристик состояния. Применительно к лечебно-диагностическому процессу подобное представление развивалось нами в работах [6, 7, 10, 11, 20]. Там же мы отмечали сложности, с которыми сталкивается этот подход. Основные проблемы связаны с тем, что вектор состояния человека может потенциально иметь очень большую размерность

(10^3–10^4), а сами характеристики состояния не могут быть наблюдаемы одновременно ни в какой выделенный дискретный момент времени.

Сегодня все большее распространение получают устройства для повседневного мониторинга состояния здоровья и физической активности (показатели работы сердца, содержание сахара в крови, двигательная активность и т.п.). Возникают новые формы организации медицинской помощи как, например, удаленный персонализированный мониторинг здоровья, личный кабинет и т.п., которые могут быть более удобны для пациента, позволяя ему вести обычный образ жизни.

Это становится особенно актуальным в условиях относительного постарения населения. В докладе объединенного экономического комитета Конгресса Президенту США 2018 года отмечается перспективность использования мобильных медицинских устройств для помощи при хронических заболеваниях (сердечно сосудистые заболевания, диабет). Предполагается широко применять инсулиновые насосы, имплантированные дефибрилляторы и кардиостимуляторы. Такие инструменты окажут врачам серьезную помощь в составлении полного списка медицинских проблем пациента в динамике за счет более полного сбора и учета данных, необходимых для тщательного и координированного медицинского наблюдения, а также для улучшения первичной профилактики путем выявления групп пациентов с различными категориями риска.

Мобильная медицина и мобильные медицинские устройства (mHealth and mHealth devices) обещают очень многое, они позволяют практически непрерывно наблюдать отдельные характеристики состояния человека, но в целом не отменяют отмеченную выше неполноту данных о состоянии человека в дискретные моменты времени, когда принимаются решения о том или ином медицинском мероприятии – воздействии на состояние. Несмотря на сложности, связанные с наблюдением состояния человека, мы считаем оправданным и перспективным применение модели ЛДП, основанной на понятии состояния пациента. Вследствие неизбежной в будущем виртуализации ключевых клинических процессов взаимодействия врача и пациента, клиницистам потребуются такие средства и инструменты, которые помогут правильно обработать данные, поступающие от множества источников информации, объединить их в единое процессное представление, и выделить те данные, которые необходимые для принятия решений.

В принятой сегодня (традиционной) модели ЛДП пребывание пациента в стационаре с ежедневным обходом лечащего врача и ежедневной констатацией состояния пациента (исключая нахождение пациента в реанимации, где требуется постоянный мониторинг состояния), амбулаторное и санаторно-курортное лечение, диспансерное

наблюдение, профосмотры – все эти процессы имеют ярко выраженную дискретизацию с моментами наблюдения состояния человека. Поэтому мы можем с уверенностью утверждать, что сегодня ЛДП можно достаточно адекватно представлять в виде дискретного управляемого процесса с неполнотой данных о состоянии управляемого объекта.

Если темп наблюдения какой-либо характеристики состояния превышает темп дискретизации процесса, то необходимо решать частную задачу осреднения (интегрирования) наблюдаемой характеристики с целью приведения ее значений к темпу всего процесса. Возможно применение метода генерализации значений наблюдаемой характеристики. Для медленно меняющихся характеристик возможно применение методов экстраполяции их значений на моменты времени, когда эти характеристики непосредственно не наблюдались. Эти особенности работы с характеристиками обсуждались в работе [**10**].

Обосновав выбор модели, можно перейти к описанию основных концептов модели. К таковым, по нашему мнению, относятся: Нозология (класс заболевания), Процесс, Состояние, Характеристика состояния, Управляемая Характеристика, Наблюдаемая Характеристика, Класс генерализации Характеристики. К Концептам предметной области следует добавить информационные концепты, связанные с представлением и обработкой БКД: Модель, Подмодель, Словарь характеристик, Селектор характеристик подмодели, Граф тесного мира, Матрица переходов Марковского процесса, Хэш-функция состояния и т.п.

Не вдаваясь в детали, мы приведем две модели БКД. Первая модель – это традиционная реляционная модель данных, а вторая модель – модель, рассчитанная на специфическую СУБД, предоставляющую возможности хранения данных таблиц по столбцам.

4. Реляционная модель БКД

Реляционная модель БКД разрабатывалась традиционным способом «без оглядки» на мощность данных и применяемую СУБД. Модель была развернута в СУБД Oracle 10g. Частично модель представлена на рис. 1. Опущена часть модели, связанная с обработкой данных, моделированием нейронных обучаемых сетей, сбором и обработкой статистики в вычислительных экспериментах. Основные концепты отражены на рис. 1.

Дадим краткое описание модели. БД хранит реализации лечебно-диагностических процессов (таблица `MDP_PROCESS`). Каждый ЛДП имеет дату начала и завершения, определена длительность процесса, указан код основного диагноза по МКБ-10, указан клинический исход

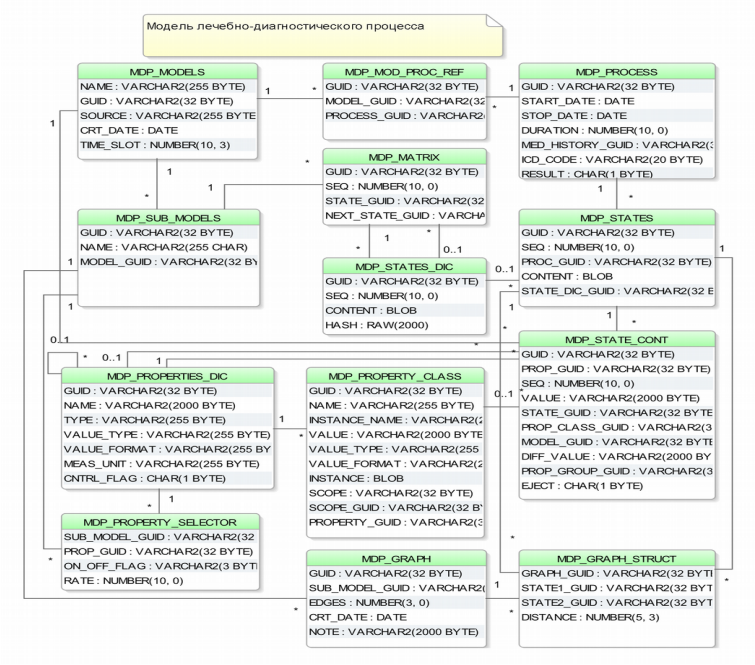

Рисунок 1. Реляционная модель БКД

процесса. Процесс дискретный и заключается в смене состояний (таблица `MDP_STATES`). Состояния в реализации процесса упорядочены путем присвоения им последовательного целочисленного номера, имеющего смысл длительности госпитализации пациента к данному дискретному моменту времени. С каждым состоянием связано множество наблюдаемых в этом состоянии характеристик (таблица `MDP_STATE_CONT`). Описание характеристик дается в словаре характеристик (таблица `MDP_PROPERTIES_DIC`). Определяется тип характеристики, тип значения характеристики, формат значения, единица измерения, выделяются характеристики - наблюдения и характеристики -управления.

В модель заложена возможность генерализации значений характеристики (таблица `MDP_PROPERTY_CLASS`). Сохраняется как оригинальное значение характеристики, так и ссылка значения характеристики на свой обобщающий класс. Над одними и теми же данными возможно построение многих моделей (таблицы `MDP_MODELS` и `MDP_SUB_MODELS`). Присутствует селектор свойств (таблица `MDP_PROPERTY_SELECTOR`), которая позволяет в рамках данной подмодели включать и исключать

Таблица 1. Мощность данных в БД

№	Концепт	Таблица	Число записей
1	Моделей	MDP_MODELS	16
2	Подмоделей	MDP_SUB_MODELS	23
3	Процессов	MDP_PROCESS	17877
4	Состояний	MDP_STATES	204920
5	Характеристик	MDP_STATE_CONT	10104151

характеристики из рассмотрения. В том случае, когда применяется модель марковского управляемого процесса [7], рассчитывается матрица переходов (таблица MDP_MATRIX). Широко применялась структуризация данных на графах тесного мира (таблицы MDP_GRAPH и MDP_GRAPH_STRUCT).

На представленной модели удалось провести целый комплекс исследований, результаты которых представлены в работах и докладах на представительных международных конференциях [8, 11, 12, 20]. Было проведено моделирование ЛДП для 8 широко распространённых нозологий. В таблице 1 приведены данные о количестве записей. Представленные данные соответствовали 8 различным нозологиям, наблюдавшимся в одном лечебном учреждении г. Москвы в течение 10 лет, с 2006 по 2016 годы. Если бы были взяты данные по всем нозологиям, то количество процессов выросло бы примерно на порядок. Соответственно мощности таблицы состояний и характеристик также увеличились бы на порядок. При попытке создать представительный банк клинических данных, объединив данные 10 аналогичных лечебных учреждений, мы бы имели мощность таблицы значений характеристик уже в миллиард записей, что позволяет нам относить этот случай к большим данным.

СУБД Oracle 10g была развернута на одном сервере. Выборки данных по одной нозологии для обработки могли выполняться в течение 5-10 минут. Структуризация данных по одной нозологии (выполнение сохраненной процедуры на языке PL/SQL) могла потребовать до 10-20 часов работы с БД. Обучение нейронных сетей для отдельной нозологии, содержащей всего 266 процессов, длилось от 1 до 2 недель [19].

Все это ясно свидетельствовало о том, что переход к большим данным невозможен в рамках имеющихся ограниченных вычислительных мощностей. И одним из лимитирующих производительность ресурсов оказалась реляционная СУБД.

Встал вопрос о применении специализированных СУБД, хорошо себя зарекомендовавших в работе с большими данными.

5. Реляционная модель БКД в СУБД с хранением по колонкам

Проблему производительности при работе с большими данными можно решать: 1) путем наращивания вычислительных ресурсов, что потребует затрат на оборудование и на используемое проприетарное ПО; 2) путем надлежащего выбора СУБД, рассчитанных именно на большие данные; 3) путем улучшения самой модели данных.

Попробуем исследовать пути 2 и 3. В работах [17–19] обсуждаются вопросы выбора СУБД для работы с большими данными. Нас в особенности заинтересовала работа [18], поскольку в ней приводились данные о сравнении производительности 5 различных СУБД. Победителем среди них оказалась СУБД Vertica. Особенностью этой СУБД является хранение данных по колонкам (Columns storage). Майкл Стоунбрейкер, продвигающий СУБД с хранением данных по колонкам, в статье «Один размер пригоден для всех»: идея, время которой пришло и ушло» обращает внимание на то, что специализированные СУБД могут работать на два порядка быстрее универсальных [17]. В [18] также обсуждаются проблемы, с которыми столкнулась компания, при обработке больших данных в СУБД Oracle.

Принцип хранения данных по колонкам позволил нам выработать другую перспективную модель БКД. Особенностью решаемых нами задач (поддержка принятия врачебных решений) является необходимость выборки обучающих и контрольных данных по отдельным нозологиям в заданном временном срезе процесса (например, на конкретный день госпитализации). Необходимо выбирать множество векторов состояний, относящихся к заданному нозологическому и временному контексту.

Все ЛДП имеют ограниченную длительность. Госпитализация зачастую длится не более 2–4 недель. Имеющиеся ограничения (1600 колонок в одной таблице для Vertica Analytics Platform 9.0) на число колонок в таблице в этом случае не являются критическими и позволяют «укладывать» процесс в одну таблицу.

Можно спроектировать таблицу, в которой колонки будут соответствовать дискретным шагам процесса (например, дням госпитализации), а упорядоченные должным образом строки сформируют в колонках последовательности векторов состояний. Если для каждой нозологии спроектировать отдельную таблицу с колонками, соответствующими шагам дискретного ЛДП, то выборка векторов состояния в заданном нозологическом и временном контексте сведется к чтению колонки (колонок) из отдельной таблицы. В этом случае Columns storage СУБД должны будут показать свое явное преимущество в решении данной задачи перед такими универсальными реляционными СУБД, как Oracle.

Таблица 2. Унифицированная модель БКД в Columns Storage DB

Таблица БКД по нозологии NOSOLOGY_ICD_CODE					
PrGuid Процесс	PropGuid Свойство	Step_1 Значение 1	Step_2 Значение 2	..	Step_N Значение n
ЛДП_1	Пол	Мужской	Мужской		Мужской
ЛДП_1	Возраст	27	27		27
ЛДП_1	Темпер.,°C	38,7	38,5		36,6
ЛДП_1	Аспирин, мг.	1	2		15
...	...	...	...	...	...
ЛДП_2	Пол	Женский	Женский	...	...
ЛДП_2	Возраст	33	33	...	...
ЛДП_2	Темпер.	36,7	36,7	...	...
...	...	...	...	...	...
ЛДП_NN	Пол	...	...	...	...
...	...	...	...	...	...

Следует отметить, что в каждой строке таблицы мы будем иметь последовательность значений одной характеристики состояния, развернутой по шагам процесса (график отдельной характеристики). Итак, предлагается дизайн модели БКД, при котором оптимизируется выборка векторов состояний в заданном контексте, а в строках отражается динамика каждой характеристики состояния. В этом случае модель данных основывается на множестве унифицированных таблиц, см. условный пример в таблице 2.

Были проведены сравнения производительности СУБД Oracle и СУБД Vertica на небольшом объеме первичных клинических данных. Клинические данные были представлены 1403 завершенными клиническими процессами по нозологии I67.9 Цереброваскулярная болезнь неуточненная. Массив клинических данных заключал в себе 24875 состояний и 655543 значений характеристик состояния. Обе СУБД были развернуты на двух отдельных серверах с сопоставимыми характеристиками. В СУБД Oracle для оптимизации выполнения запросов широко применялись индексы. В СУБД Vertica, как уже было подчеркнуто выше, применялась модель данных, ориентированная на хранение данных по колонкам.

Эксперименты действительно подтвердили заметное преимущество Vertica по скорости выборки данных (времени выполнения SQL запроса). При выборке данных из одного временного слоя (одного столбца в модели Vertica) скорость выборки Vertica превышала скорость выборки Oracle примерно в 15 раз. При выборке для указанной

нозологии всех данных, отсортированных по временным слоям (дням госпитализации) Vertica выбрала все данные за 222 мс, а Oracle для формирования той же выборки потребовалось 460 с. Скорость выборки для Oracle удалось существенно повысить путем создания соответствующих материализованных представлений (materialized view). Из представления, содержащего клинические данные по всем нозологиям, указанным в таблице 1, данные по нозологии I67.9 выбирались в 40 раз медленнее за 8,81 с против 0,222 с у Vertica.

При создании в Oracle материализованного представления с данными только по одной нозологии I67.9 скорость выборки возросла до 0,96 сек., что в 4,3 медленнее, чем у Vertica. Необходимо отметить, что в СУБД Oracle в одних таблицах смешивались данные по всем нозологиям, см. таблицу 1. А в СУБД Vertica согласно принятой модели данные расщепляются по нозологиям по отдельным таблицам, что ускоряет выборку данных по каждой нозологии в отдельности.

Авторы понимают всю условность приведенных результатов и не претендуют на исчерпывающую оценку производительности этих двух СУБД, т.к. организация данных была различна и это существенно влияло на производительность. Однако, можно с уверенностью сделать вывод о том, что вторая проблемно-ориентированная модель клинических данных в СУБД с хранением данных по колонкам имеет преимущества перед первой моделью, а СУБД Vertica на второй модели данных имеет преимущества по скорости выборки данных перед СУБД Oracle с первой моделью данных на сопоставимых вычислительных мощностях.

Заключение

В работе рассмотрены две модели клинических данных. Модели ориентированы на применение методов ИИ и машинного обучения.

Изложен многолетний опыт по моделированию лечебно-диагностических процессов, формированию базы клинических данных, решению задач поддержки принятия врачебных решений. Отмечены недостатки модели данных, разработанной для универсальной реляционной СУБД. Отмечены проблемы, возникающие при наращивании мощности данных и переходе к большим клиническим данным.

Предложен новый дизайн модели клинических данных, ориентированный на СУБД с хранением данных по колонкам. В качестве такой перспективной СУБД рассмотрена СУБД Vertica. Отмечены преимущества специализированных СУБД перед универсальными при работе с большими данными.

Полученные результаты будут практически использованы в дальнейших исследованиях проблем проектирования хранилищ для больших данных и поддержки принятия врачебных решений на основе представительных банков клинических данных.

Благодарности. Авторы искренне благодарны М. И. Хаткевичу за интерес к работе и конструктивную критику.

Список литературы

[1] E. Herrett et al. “Data resource profile: Clinical Practice Research Datalink (CPRD)”, *International Journal of Epidemiology*, **44**:3 (2015), pp. 827–836. ↑220

[2] M. L. Gentil et al. “Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature”, *BMC Medical Informatics and Decision Making*, **17**:1 (2017), pp. 139. ↑220,221

[3] В. Л. Малых, Я. И. Гулиев. «Прецеденты в медицинских информационных системах», *Программные продукты и системы*, 2009, №2 (86), с. 19–27. ↑223

[4] В. Л. Малых, Я. И. Гулиев, А. И. Крылов, Е. В. Рюмина. «Проблемы автоматизации учета прямых материальных затрат в медицине. Архитектура прецедентного материального учета», *Аудит и финансовый анализ*, 2009, №2, с. 465–471. ↑223

[5] И. В. Ефименко, В. Ф. Хорошевский. «Интеллектуальные системы поддержки принятия решений в медицине: ретроспективный обзор состояния исследований и разработок и перспективы», *Открытые семантические технологии проектирования интеллектуальных систем*, Материалы международной конференции, OSTIS-2017, БГУИР, Минск, 2017, с. 251–260. ↑223

[6] В. Л. Малых, Я. И. Гулиев. «Моделирование лечебно-диагностического процесса в классе управляемых стохастических процессов с памятью», *Врач и информационные технологии*, 2013, №2, с. 6–15. ↑223,224

[7] В. Л. Малых, Я. И. Гулиев. «Управляемый стохастический прецедентный процесс с памятью как математическая модель лечебно-диагностического процесса», *Информационные технологии и вычислительные системы*, 2014, №2, с. 60–72. ↑223,224,228

[8] V. L. Malykh, Y. I. Guliev. “Precedent approach to decision-making in clinical processes”, MEDINFO 2015: eHealth-enabled Health, Studies in Health Technology and Informatics, vol. **216**, eds. I. N. Sarkar et al., IMIA and IOS Press, 2015, pp. 957. ↑223,228

[9] В. Л. Малых, Я. И. Гулиев, Д. В. Белышев. «Построение банка клинических данных на основе унифицированной модели лечебно-диагностического процесса», *Труды XVII международной конференции DAMDID/RCDL’2015* (Обнинск, 13–16 октября 2015). ↑223

[10] V. L. Malykh, D. V. Belyshev. “Case-based reasoning in clinical processes using clinical data banks”, *Proceedings of 2015 International Conference on Biomedical Engineering and Computational Technologies*, SIBIRCON (Russia, Technopark of Novosibirsk Akademgorodok, 28–30 October 2015), IEEE Conference Publications, pp. 211–216. ↑223,224,226

[11] V. L. Malykh, I. N. Kononenko, S. V. Rudetskiy. “Estimation of accuracy of recommended diagnostic and treatment actions based on precedent approach”, *Proceedings of the International Conference e-Health 2016* (Madeira, Portugal, July 1–4, 2016), pp. 52–58. ↑223,224,228

[12] В. Л. Малых, С. В. Рудецкий. «Сравнительный анализ различных подходов к поддержке принятия врачебных решений на основе больших клинических данных», Московская научно-практическая конференция по Искусственному интеллекту в медицине MosCAI’17 (12 октября 2017 года, Москва, Конгресс-центр гостиницы Космос). ↑223,228

[13] Д. В. Белышев, Е. В. Кочуров. «Анализ методов хранения данных в современных медицинских информационных системах», *Программные системы: теория и приложения*, **7**:2(29) (2016), с. 85–103. ↑224

[14] Д. В. Белышев, Е. В. Кочуров. «Перспективные методы работы с данными в медицинских информационных системах», *Программные системы: теория и приложения*, **7**:3(30) (2016), с. 79–97. ↑224

[15] В. Л. Малых, А. Н. Калинин, Т. Ш. Юсуфов. «Объектно- реляционный подход к построению хранилища данных», *Программные системы: теория и приложения*, **8**:3(34) (2017), с. 167–185. ↑224

[16] *Modernizing IBM i applications from the database up to the user interface and everything in between*, IBM Redbooks publication, 2014. ↑224

[17] К. Вахрамеев. «СУБД для анализа Больших Данных», *Открытые системы*, 2011, №10. ↑224,229

[18] А. Зайцев. *Эволюция аналитической инфраструктуры*, 2012. ↑224,229

[19] «Mail.Ru и «АйТеко» представили платформу CoIIoT», *Открытые системы*, 2017, №3. ↑224,228,229

[20] V. L. Malykh, S. V. Rudetskiy. “Approaches to medical decision-making based on big clinical data”, *Journal of Healthcare Engineering*, **2018** (2018), 3917659, 10 pp. ↑224,228

Поступила в редакцию 29.08.2018
Переработана 03.09.2018
Опубликована 28.11.2018

Рекомендовал к публикации *к.т.н. Я.И. Гулиев*

Пример ссылки на эту публикацию:

В. Л. Малых, А. Е. Михеев, С. В. Рудецкий. «Проблемно-ориентированная модель банка клинических данных». *Программные системы: теория и приложения*, 2018, **9**:4(39), с. 219–237.

10.25209/2079-3316-2018-9-4-219-237

http://psta.psiras.ru/read/psta2018_4_219-237.pdf

Об авторах:

Владимир Леонидович Малых

Зав. лабораторией Института программных систем им. А. К. Айламазяна

0000-0002-0072-0724

e-mail: mvl@interin.ru

Александр Евгеньевич Михеев

С.н.с. Института программных систем им. А. К. Айламазяна

e-mail: miheev@interin.ru

Сергей Владимирович Рудецкий

М.н.с. Института программных систем им. А. К. Айламазяна

e-mail: rsv@interin.ru

CSCSTI 76.03.59
UDC 61:004.652

Vladimir Malykh, Aleksandr Mikheyev, Sergey Rudetskiy. *Problem-oriented model of clinical data Bank.*

Abstract. The paper considers the problem of building a problem-oriented information model of the clinical data Bank (CDB). The treatment and diagnostic process is proposed to be modeled by a discrete finite-dimensional controlled process.

The principal nonobservability of the state vector of the control object and its large dimension are noted. The necessity of application of Big Data technologies for the construction of BCD is grounded. Two relational models of clinical data are presented in a process form. The author's experience of using one of the presented relational models is given. The perspective of the second relational model, based on a relational DBMS with data storage by columns, is noted. The results of the work will be useful to developers of health information systems.

Key words and phrases: data collection, clinical data bank, relational database, columns storage database, medical information system.

2010 *Mathematics Subject Classification*: 68P15; 68P20, 92C50

10.25209/2079-3316-2018-9-4-219-237

References

[1] E. Herrett et al. "Data resource profile: Clinical Practice Research Datalink (CPRD)", *International Journal of Epidemiology*, **44**:3 (2015), pp. 827–836. ↑220

[2] M. L. Gentil et al. "Factors influencing the development of primary care data collection projects from electronic health records: a systematic review of the literature", *BMC Medical Informatics and Decision Making*, **17**:1 (2017), pp. 139. ↑220,221

[3] V. L. Malykh, Ya. I. Guliyev. "Precedents in medical information systems", *Programmnyye produkty i sistemy*, 2009, no.2 (86), pp. 19–27 (in Russian). ↑223

[4] V. L. Malykh, Ya. I. Guliyev, A. I. Krylov, Ye. V. Ryumina. "Problems of automation of calculation of direct material expenditures in medicine. Precedent-based architecture of materials accounting", *Audit i finansovyy analiz*, 2009, no.2, pp. 465–471 (in Russian). ↑223

[5] I. V. Yefimenko, V. F. Khoroshevskiy. "Intelligent decision support systems in medicine: state of the art and beyond", *Otkrytyye semanticheskiye tekhnologii proyektirovaniya intellektual'nykh sistem*, Materialy mezhdunarodnoy konferentsii, OSTIS-2017, BGUIR, Minsk, 2017, pp. 251–260 (in Russian). ↑223

[6] V. L. Malykh, Ya. I. Guliyev. "Modeling of medical-diagnostic process in the class of controlled stochastic processes with memory", *Vrach i informatsionnyye tekhnologii*, 2013, no.2, pp. 6–15 (in Russian). ↑223,224

[7] V. L. Malykh, Ya. I. Guliyev. "Controlled stochastic precedent process with memory as a mathematical model of the diagnostic and treatment process", *Informatsionnyye tekhnologii i vychislitel'nyye sistemy*, 2014, no.2, pp. 60–72 (in Russian). ↑223,224,228

[8] V. L. Malykh, Y. I. Guliev. "Precedent approach to decision-making in clinical processes", MEDINFO 2015: eHealth-enabled Health, Studies in Health Technology and Informatics, vol. **216**, eds. I. N. Sarkar et al., IMIA and IOS Press, 2015, pp. 957. ↑223,228

[9] V. L. Malykh, Ya. I. Guliyev, D. V. Belyshev. "The Bank of clinical data based on the unified model of treatment and diagnostic process", *Trudy XVII mezhdunarodnoy konferentsii DAMDID/RCDL'2015* (Obninsk, 13–16 oktyabrya 2015) (in Russian). ↑223

[10] V. L. Malykh, D. V. Belyshev. "Case-based reasoning in clinical processes using clinical data banks", *Proceedings of 2015 International Conference on Biomedical Engineering and Computational Technologies*, SIBIRCON (Russia, Technopark of Novosibirsk Akademgorodok, 28–30 October 2015), IEEE Conference Publications, pp. 211–216. ↑223,224,226

[11] V. L. Malykh, I. N. Kononenko, S. V. Rudetskiy. "Estimation of accuracy of recommended diagnostic and treatment actions based on precedent approach", *Proceedings of the International Conference e-Health 2016* (Madeira, Portugal, July 1–4, 2016), pp. 52–58. ↑223,224,228

[12] V. L. Malykh, S. V. Rudetskiy. "Comparative analysis of different approaches to developing medical decision-making systems based on large clinical data", Moskovskaya nauchno-prakticheskaya konferentsiya po Iskusstvennomu intellektu v meditsine MosCAI'17 (12 oktyabrya 2017 goda, Moskva, Kongress-tsentr gostinitsy Kosmos) (in Russian). ↑223,228

[13] D. V. Belyshev, Ye. V. Kochurov. "Analysis of data storage methods for modern healthcare information systems", *Program Systems: Theory and Applications*, **7**:2(29) (2016), pp. 85–103 (in Russian). ↑224

[14] D. V. Belyshev, Ye. V. Kochurov. "Advanced methods of data management in healthcare information systems", *Program Systems: Theory and Applications*, **7**:3(30) (2016), pp. 79–97 (in Russian). ↑224

[15] V. L. Malykh, A. N. Kalinin, T. Sh. Yusufov. "Object-relational approach to building a data storage", *Program Systems: Theory and Applications*, **8**:3(34) (2017), pp. 167–185 (in Russian). ↑224

[16] *Modernizing IBM i applications from the database up to the user interface and everything in between*, IBM Redbooks publication, 2014. ↑224

[17] K. Vakhrameyev. "DB for Big Data analysis", *Otkrytyye sistemy*, 2011, no.10 (in Russian). ↑224,229

[18] A. Zaytsev. *Evolution of analytical infrastructure*, 2012 (in Russian). ↑224,229

[19] "Mail.Ru and «ITex» presented the platform CoIIoT", *Otkrytyye sistemy*, 2017, no.3 (in Russian). ↑224,228,229

[20] V. L. Malykh, S. V. Rudetskiy. "Approaches to medical decision-making based on big clinical data", *Journal of Healthcare Engineering*, **2018** (2018), 3917659, 10 pp. ↑224,228

Sample citation of this publication:

Vladimir Malykh, Aleksandr Mikheyev, Sergey Rudetskiy. "Problem-oriented model of clinical data Bank". *Program Systems: Theory and Applications*, 2018, **9**:4(39), pp. 219–237. (*In Russian*).

10.25209/2079-3316-2018-9-4-219-237

http://psta.psiras.ru/read/psta2018_4_219-237.pdf