УДК 681.3.016

Г.Н.Востров, канд. техн. наук, доц.,
Е.В.Малахов, канд. техн. наук, доц.,
В.В.Мороз, канд. техн. наук, доц.

ПРОБЛЕМЫ СОЗДАНИЯ БАЗ ДАННЫХ И ИНФОРМАЦИОННЫХ ХРАНИЛИЩ

Г.М. Востров, Є.В. Малахов, В.В. Мороз. Концептуальні проблеми створення баз даних та інформаційних сховищ. Розглядаються проблеми створення баз даних та досліджуються умови переходу їх в клас сховищ даних.

G.M. Vostrov, E.V. Malakhov, V.V. Moroz. Conceptual problems of creation of databases and warehouses. Consideration is given to the problems of databases creation and investigated are conditions of their transfer to class of data warehouses.

В настоящее время в области информационных технологий объектом серьезного исследования становятся информационные хранилища или хранилища данных (datawarehouse). Их можно рассматривать как развитие конструкции баз данных в направлении ориентации на конечного потребителя информационной продукции. Особую важность приобретает обеспечение эффективного использования данных с применением интеллектуальных средств их анализа и представления, алгоритмов многомерного моделирования и других технологий анализа и обработки информации. Использование информации конечным потребителем предполагает, что базы данных удовлетворяют определенному множеству концептуальных предположений, знание которых важно для него. Если потребитель информации заранее не знает, какому набору свойств удовлетворяет база данных, то это может создать для него ряд проблем на этапе использования информации. В частности, повторный одинаковый запрос к базе данных через небольшой промежуток времени может давать различные результаты. Такая ситуация зачастую оказывается неожиданной для конечного потребителя как с точки зрения использования полученной информации, так и методов ее обработки. В тех случаях, когда известно, что предметная область находится в стационарном состоянии, база данных, входящая в ее информационную модель, не должна включать быстро изменяющиеся данные, она должна обладать свойством стабильности. Базы данных не всегда удовлетворяют условию стабильности. В более общем случае при обращении к базам данных важно заранее знать полный спектр свойств которыми они обладают. К таким свойствам относятся адекватность, полнота, семантика баз данных, используемые форматы и др.

В процессе сопровождения баз данных и обработки информации они могут, начиная с некоторого момента, переходить в состояние, когда либо вся совокупность накопленной информации практически не изменяется, либо некоторая ее часть или в любой части базы данных в любой момент времени могут происходить значительные изменения. Для всех пользователей информации и особенно тех, кто обращается к ней через международные информационные сети, важно знать характер стабильности накопленных данных. В классе баз данных с четко определенными концептуальными свойствами важный подкласс образуют те, которые ориентированы лишь на хранение данных, а приложения носят внешний характер. Это означает, что пользователь таких баз данных только по своему запросу получает всю информацию, согласующуюся с запросом, а ее обработка не может изменять хранящейся информации.

На поддержку решения задач подобного рода ориентированы хранилища данных. Создание информационных хранилищ требует полного концептуального анализа баз данных и условий превращения их в эту новую категорию.

Результаты исследования всех вариантов развития баз данных в специализированные хранилища данных зависят от базовых концептуальных предположений определяющих разработку, создание и сопровождение реляционных, сетевых и многомерных баз данных. Рассмотрим класс реляционных баз данных [1] и концептуальные предположения, определяющие их свойства как основных компонент информационных моделей предметных областей [2].

Приведенные классы предположений в общем случае не исчерпывают всего многообразия возможных характерных особенностей реляционных баз данных. Приведенные предположения являются основными в том смысле, что они определяют закономерности проектирования, создания и развития баз данных в объекты, которые содержат достаточно полную информацию о предметной области.

Ослабляя либо усиливая эти предположения, будем получать реляционные базы данных, которые могут быть переведены в категории носителей информации, обращение к которым может носить принципиально другой характер.

На первых этапах создания баз данных используются операции включения и удаления кортежей, изменения кортежей как по набору атрибутов, так и по их содержанию. В процессе развития баз данных включаются новые кортежи, создаются реляционные отношения и функциональные зависимости, конструируются и программно реализуются новые операции над реляционными отношениями,

модифицируются исчисление кортежей и атрибутов, изменяется структура базы данных. Для сетевых баз данных может изменяться и топологическая структура графа, определяющего каналы обмена данными между ее компонентами. Развитие баз данных включает операции удаления записей и их преобразования. Изменение множества атрибутов, удаление и включение новых кортежей, изменение реляционных отношений меняет внутреннюю структуру базы данных. В процессе развития база данных стремится к некоторому предельному состоянию, которое является основной компонентой информационной модели выделенной предметной области с заданной степенью приближения, определяемой некоторой мерой адекватности.

Ранее предпринималась попытка концептуального моделирования объектов информационных технологий, которые авторами относятся к классу баз данных [3]. При этом формулировались, в основном, концептуальные предположения, которые позволяли моделировать семантику баз данных. Предположения, сформулированные разными авторами, весьма отличаются, но в то же время в большинстве признается, что любые современные базы данных должны удовлетворять определенному общепризнанному набору свойств.

Для одной и той же предметной области может быть построено многообразие информационных моделей. Они могут отличаться не только степенью адекватности, но и используемым математическим аппаратом, структурой, схемами отношений, полными множествами атрибутов, реляционными отношениями и другими характеристиками. Выбор конкретной информационной модели определяется концептуальными предположениями, сформулированными как на этапе проектирования базы данных для выделенной предметной области, так и на этапе развития базы данных. На формулировку концептуальных предположений существенное влияние оказывают, с одной стороны бизнес-приложения, а с другой стороны интересы конечного потребителя информации. Они определяются классами задач, которые решаются над базой данных. В процессе решения задач могут возникать новые классы задач, которые изменяют концептуальные предположения. При этом могут изменяться структура базы данных, реляционные отношения, наборы операций над реляционными отношениями, множество атрибутов полного описания. Концептуальные предложения формируют направленность создания баз данных.

Базы данных являются основными компонентами информационных моделей предметных областей. Информационные модели отражают концептуальную сущность объектов предметной области [2]. Получение новой информации о предметной области приводит к появлению новых кортежей без прямого построения новых реляционных отношений, но с появлением новых результатов применения операций реляционной алгебры. Из данных определенной модели могут быть выделены зависимости между атрибутами, кортежами. Характер представления зависимостей определяется инструментальными средствами, разрабатываемыми на уровне приложений базы данных. Независимо от формы представления данных, предполагается, что развитая база данных содержит основной базис данных информационной модели предметной области. На этапе перехода базы данных в стационарное состояние получаем для объекта в предметной области законченное информационное описание. В этом состоит одна из основных задач создания баз данных.

Семантика баз данных определяется процедурами прямого и обратного отображения между объектами предметной области и конструкциями баз данных. При рассмотрении предметных областей как объектов информационного моделирования всегда используется определенная степень идеализации содержательных представлений, знаний о структуре предметной области и отношениях между ее объектами. Уровень идеализации не имеет точной формализации и методов оценивания его величины в процессе проектирования и построения баз данных. Предполагается, что по мере развития баз данных строящаяся информационная модель в пределе стремится к некоторой величине степени идеализации.

Эквивалентом содержательного смысла отношений, зависимостей между объектами предметной области является семантика баз данных. Семантические отношения, зависимости между атрибутами, кортежами базы данных являются отображением содержательного смысла отношений, зависимостей в предметной области. Очевидно, что атрибуты, кортежи, реляционные отношения и функциональные зависимости между ними являются информационной формализацией объектов, структурных образований, отношений и зависимостей между ними в предметной области.

Предположение, связанное с семантикой баз данных, состоит в том, что существует механизм содержательной интерпретации реляционных отношений, функциональных зависимостей, которые строятся с помощью формального аппарата реляционной алгебры, системы функциональной зависимости. Это предложение влечет за собой требование полноты реляционных отношений, исчислений атрибутов, кортежей и других формальных систем выделения и представления информации из совокупностей данных, накапливаемых в базе данных.

Решение задачи содержательной интерпретации предложений, высказываний, описанных в терминах реляционных отношений, функциональных зависимостей в каждой базе данных осуществляется путем процедур отображения предметной области на базу данных и обратного отображения базы данных в предметную область. Базы данных как развивающиеся системы стремятся к предельному состоянию, удовлетворяющему требованию информационной модели предметной области.

База данных является полным информационным представлением предметной области, если добавление любых кортежей не приводит к построению новых реляционных отношений средствами

реляционной алгебры, исчислений кортежей, атрибутов и приложений. Ядром базы данных будем называть то минимальное подмножество кортежей, которое обеспечивает полноту. Очевидно, что любая база данных имеет не одно ядро. Ядро будет называться устойчивым, если в процессе сопровождения базы данных его записи не изменяются.

Предположение о полноте базы данных отражает тот факт, что построенная информационная модель предметной области может быть использована для построения бизнес-приложений, поддерживающих эффективные процедуры решений, либо адаптирована для решения задач конечного пользователя, не связанного с поддержкой процедур принятия решений. Семантическая полнота баз данных является важной при содержательной интерпретации функциональных реляционных зависимостей, которые строятся средствами реляционной алгебры. Качество семантической интерпретации реляционных отношений и функциональных зависимостей влияет и на полноту баз данных.

Предметную область можно декомпозировать на элементарные объекты, каждый из которых описывается совокупностью атрибутов, образующих полную схему базы данных. Объекты предметной области связаны между собой определенными отношениями, которые можно в совокупности представить в виде взвешенного по ребрам частично ориентированного графа [4]. Структура графа представляет структуру предметной области. Подграфы графа представляют сложные объекты или подсистемы предметной области. Вместо графов для представления структуры предметной области можно использовать язык теории множеств и решеток их разбиений. Каждый кортеж базы данных является описанием состояния некоторого элементарного объекта. Подмножество всех кортежей, сходных с данным кортежем, относительно выбранной меры сходства, является представлением элементарного объекта. В качестве таких подмножеств можно выбрать кластеры [5]. Применение методов иерархической классификации [6] дает представление структуры баз данных, которое так же может иметь вид частично ориентированного взвешенного графа, либо подрешетки разбиения множества кластеров. Если структуры предметной области и базы данных изоморфны, то ясно, что база данных адекватно отражает предметную область. Предположение об адекватности означает, что база данных как развивающаяся система стремится к состоянию, в котором она адекватно представляет в информационном плане предметную область.

База данных содержит стационарное ядро, если не изменяется схема базы данных, выделенное подмножество кортежей, реляционных отношений, функциональных зависимостей, исчислений кортежей и атрибутов. Изменяющуюся часть базы данных будем называть нестационарной компонентой. База данных обладает свойством полноты, если выделенное множество реляционных отношений стационарного ядра позволяет вывести средствами реляционной алгебры все реляционные отношения, которые могут быть построены над ядром. Предположения о полноте и адекватности означают, что база данных является основной компонентой информационной модели предметной области, которая содержит всю необходимую информацию о предметной области как в количественном, так и в качественном отношениях.

Стационарные состояния баз данных являются их предельным состоянием как развивающейся системы. В процессе создания баз данных динамически изменяются их схемы, удаляются и обновляются кортежи, изменяются реляционные отношения, функциональные однозначные и многозначные зависимости, исчисление атрибутов и кортежей. Динамические изменения баз данных носят стохастический характер, однако они всегда находятся под влиянием человеческого фактора. В конечной цели база данных должна содержать всю необходимую информацию для решения выделенного класса задач. Это обусловливает целенаправленный характер их развития. Результатом развития является переход в стационарное состояние. В этом состоянии все характеристики баз данных колеблются вокруг некоторых значений. Колебания носят черты стационарности, присущие стационарным стохастическим процессам. Данное предположение означает, что стационарное состояние является предельным состоянием развивающейся системы.

Во всех реально разработанных реляционных базах данных информация в кортежах представляется с использованием ограниченного набора стандартизированных форматов. Однако при проектировании и создании баз данных, ориентированных на конкретные предметные области, разные авторы используют для атрибутов различные единицы измерения, способы кодирования и, как следствие, разработчики создают собственные форматы. Свобода в выборе форматов удобна при проектировании и создании баз данных, но очевидно создает серьезные проблемы при поиске и передаче информации конечным потребителем, не связанных с созданием клиентских приложений.

Базы данных создаются с целью полного информационного обеспечения всех классов задач, связанных с предметной областью. Для решения задач разрабатываются приложения и функции. Математическое и программное обеспечение приложений и функций разрабатывается различными авторами. На него не накладываются жесткие ограничения. Это обусловливает многообразие методов анализа и обработки информации и интерпретации их результатов.

В частности, создание баз данных ориентировано на информационное обеспечение процедур поддержки принятия решений. Разработка эффективных управленческих решений основана на использовании методов группового выбора, экспертных систем, прогнозирования оптимального управления, дискретной оптимизации, автоматической классификации, статистических алгоритмов анализа и обработки данных. Все методы, применяемые для принятия решений, позволяют пользователям достигать желаемых

результатов при условии, что выборочные данные обладают свойством представительности. Выполнение условия репрезентативности не зависит от стабильности хранящейся в базах данных информации, оно опирается на свойства адекватности и полноты баз данных как основных компонент информационных моделей предметных областей. Поэтому базы данных, обладающие такими свойствами, как правило, удовлетворяют всем требованиям приложений и функций, разрабатываемых системными программистами и прикладными математиками. Однако существует обширный класс потребителей информации баз данных, которых не устраивает среднестатистическая их устойчивость, так как данные, размещенные в них, рассматриваются ими не только в плане адекватности и полноты.

Систематическое обращение к базам данных, обладающих свойствами полноты и адекватности, часто предполагает, что при каждом выполнении одного и того же запроса будет получена та же информация. Требование стабильности информации связано с тем, что объекты данных выходят на первый план, а приложения и функции становятся второстепенными или вообще не представляющими интереса. Первостепенная важность данных проявляется тогда, когда обработка и анализ накопленной в них информации осуществляется за их пределами. Данные, получаемые при обработке запросов пользователей, могут комбинироваться ими с данными, получаемыми из других объектов данных, для решения задач, лежащих за пределами предметной области, связанной с базой данных. В таких случаях условие стабильности информации является первостепенным.

Информация в базах данных создается группами авторов под одним центром управления, но так бывает не всегда. Когда используется информация из приложений, созданных разными авторами, не согласующими друг с другом свои разработки, возникает необходимость приведения данных из различных объектов данных к общему знаменателю.

Истинность информации в базах данных носит условный характер. Она проявляется как среднестатистическая характеристика. Истинность не сохраняется для элементов данных, так как они могут модифицироваться в результате выполнения транзакций. Эта степень свободы информации в базах данных становится не приемлемой, когда элементы данных приобретают другую ценность и могут использоваться многими потребителями, связанными с различными другими предметными областями.

Во временном интервале, фиксируемом обществом внешних пользователей, базы данных не сохраняют стабильность накопленной информации, что может оказаться неприемлемым для них. Это особенно важно в тех случаях, когда для них представляют интерес стационарные свойства предметных областей, а не быстро изменяющаяся внешняя оболочка.

Из приведенных соображений следует, что базы данных в некоторых случаях обладают слишком большим количеством степеней свободы. Предположения о свойствах баз данных и методах накопления в них информации не всегда согласуются со всеми возможными схемами ее потребления. Это приводит к созданию новых классов объектов данных. Одними из них являются информационные хранилища, идея создания которых возникла в фирме IBM более десяти лет тому назад, но стала интенсивно развиваться только в последние два года. Поэтому возникают две важные проблемы, связанные с разработкой аксиоматики и базовых предположений, определяющих информационные хранилища и с определением условий, при которых базы данных переходят в класс хранилищ данных. На решение этих проблем и направлены исследования в области перспективных информационных технологий.

Литература

1. Мейер Д. Теория реляционных баз данных. —М.: Мир, 1988.
2. Шлеер С., Меллор С. Объектно-ориентированный анализ: моделирование мира в состояниях. —К.: Диалектика, 1993.
3. Цаленко М.Ш. Моделирование семантики в базах данных. —М.: Наука, 1989.
4. Кристофидес Р. Теория графов. Алгоритмический подход. —М.: Мир, 1982.
5. Классификация и кластер. —М.: Мир, 1980.
6. Жамбю Ф. Иерархический кластерный анализ. —М.: Мир, 1980.