

УДК 004.62:004.492

Е.В. Малахов, канд. техн. наук, доц. Одес. нац. политехн. ун-т

ОЦЕНКА СТЕПЕНИ АДЕКВАТНОСТИ БАЗ ДАННЫХ КАК ИНФОРМАЦИОННЫХ МОДЕЛЕЙ ПРЕДМЕТНЫХ ОБЛАСТЕЙ

Є.В. Малахов. Оцінювання ступеня адекватності баз даних як інформаційних моделей предметних областей. Розглянуто проблеми створення баз даних. Виділено класи задач, для рішення яких бази даних є основою їхньої інформаційної підтримки. Показано, що бази даних є інформаційними моделями предметних областей. Приведено метод оцінювання ступеня адекватності баз даних предметним областям як їхнім інформаційним моделям.

E.V. Malakhov. Adequacy degree estimation of databases as information models of subject domains. Problems of databases creation are considered. Classes of tasks for which decision databases are the basis of their information support are marked out. Databases are demonstrated to be information models of subject domains. The method of an adequacy degree estimation of databases to subject domains as their information models is adduced.

Объем накопленной во всем мире информации представляется числом фантастической величины. Количество информации, производимой мировым сообществом, удваивается каждые четыре года, при этом темпы ее получения непрерывно возрастают.

Информация накапливается в базах данных и информационных хранилищах различных типов. В то же время только ее незначительная доля используется для решения прикладных и научных проблем, что может быть объяснено несколькими факторами. Процессы получения информации связаны с предметными областями, для описания которых не существует как формализованного, так и общепризнанного неформализованного подхода [1]. Принято считать, что формализация понятия “предметная область” невозможна в силу того, что оно является первичным [1]. Не существует даже общепризнанных его описаний. Такая ситуация обуславливает множество неопределенностей, которые возникают при проектировании баз данных, накоплении в них информации, получении новых знаний из накопленной информации, построении процедур принятия решений в системах управления и методов их информационной поддержки.

Разработка эффективных процедур принятия решений и систем их информационной поддержки в значительной степени зависит от меры соответствия баз данных предметным областям. Базы данных должны отражать процессы информационного обмена между предметными областями и внешней средой, а также между их объектами.

Неадекватное представление информационного обмена обуславливает многообразие неопределенностей при построении процедур принятия решений. Неопределенности построения баз данных и процедур принятия решений приводят к некорректной информационной поддержке процессов управления предметными областями.

Перечисленные факты обуславливают необходимость создания методов построения процедур контроля за процессами накопления информации в базах данных с последовательным уменьшением или устранением всех форм неполноты и неопределенностей, возникающих как на этапах проектирования баз данных, так и в процессе их формирования. Необходима разработка математических методов проверки адекватности баз данных предметным областям как их информационным моделям, корректировки структур описания баз данных и инструментальных средств их наполнения выборочными данными.

Понятия “предметная область” и “объект” являются ключевыми в проблемах концептуального моделирования семантики в базах данных, создания систем искусственного интеллекта, построения процедур принятия решений в условиях неопределенности и их многоаспектной системной поддержки, построения динамических моделей предметных областей. Отсутствие строгих формальных определений этих понятий создает проблемы при построении баз данных, которые можно рассматривать как структурные, информационные, алгоритмические модели предметных областей и их объектов.

Априорная информация о предметных областях и их объектах, как правило, обладает значительной степенью неопределенности. Отсутствие точных конструктивных определений

этих понятий обуславливает неопределенность множества входных, выходных и внутренних переменных, описывающих свойства системы управления предметной областью и динамику их изменения. Еще большая неопределенность имеет место относительно структуры предметных областей и их объектов, которая отражает характер их взаимодействия и взаимосвязей. Существенная неопределенность связана с внутренними и внешними возмущающими воздействиями, которые по своей природе являются неизмеряемыми.

Одним из средств устранения неопределенностей являются базы данных, в которых накапливается информация, отражающая динамику изменения свойств предметных областей, их объектов и характер взаимодействия между ними. Основной проблемой их создания является обеспечение максимальной адекватности соответствующим предметным областям еще на первых этапах проектирования, которая зависит от полноты описания предметных областей формальными средствами.

Предлагается подход к построению формального описания предметных областей. Объект $o(t)$ будем называть элементарным, если он не может быть разложен на более простые объекты, удовлетворяющие некоторому комплексу условий C [2]. Произвольную выделяемую предметную область будем обозначать символом P , а множество всех ее элементарных объектов — $O(t)$. Предполагается, что любая предметная область P погружена в некоторую среду S [3]. Из среды S в предметную область P поступает информация, представляемая множеством $X(t) = \{x_1(t), \dots, x_n(t)\}$, каждый элемент которого представляет либо детерминированную функцию переменной t любой природы, либо случайный процесс.

Множество переменных $Y(t) = \{y_1(t), \dots, y_k(t)\}$ представляет информацию, которая передается от предметной области P в среду S . Оно может включать как детерминированные функции некоторой переменной t , так и случайные процессы, зависящие от переменных, которые могут иметь любую природу. Множество $Z(t) = \{z_1(t), \dots, z_m(t)\}$ представляет совокупность случайных процессов возмущающего типа. Они, как правило, представляют случайные процессы, недоступные измерению. Необходимо отметить, что выборочная информация о случайных многомерных процессах $X(t)$, $Y(t)$ накапливается в базах данных.

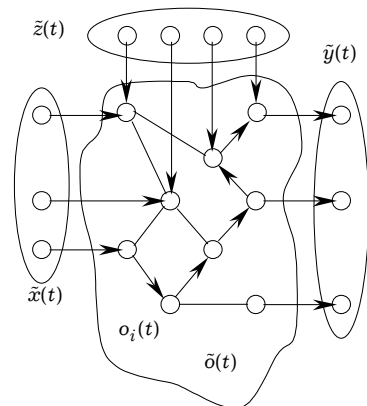
Пусть $\tilde{O}(t)$ — множество всех объектов выделенной области, которое может быть как конечным, так и бесконечным любой мощности. Каждый элементарный объект $O(t)$ содержит собственную систему входных, выходных переменных и возмущающих случайных процессов.

Пусть $\tilde{o}_i(t) \in \tilde{O}(t)$ — произвольный объект предметной области. Множество $J_i(t) = \{J_{ii}(t), J_{io}(t), J_{im}(t)\}$ состоит из подмножеств $J_{ii}(t)$, $J_{io}(t)$, $J_{im}(t)$ — соответственно входных, выходных и возмущающих переменных объекта $\tilde{o}_i(t)$. Множество $J(t)$ представляет собой объединение всех случайных переменных $\bigcup_i J_i(t)$, описывающих выделенную предмет-

ную область. Очевидно, что $X(t)$, $Y(t)$, $Z(t)$ являются собственными подмножествами $J(t)$.

Можно без потери общности предполагать, что любое подмножество элементарных объектов образует неэлементарный объект $\tilde{o}(t)$ определенной структуры, определяемой множеством элементарных объектов и оператором $Q(t)$, который отражает структуру взаимосвязи между элементарными объектами и динамику ее изменения. По существу $\tilde{o}(t)$ представляет частично ориентированный, взвешенный как по вершинам, так и по ребрам и дугам граф с выделенной системой входов $\tilde{x}(t)$, выходов $\tilde{y}(t)$ и возмущающих случайных воздействий $\tilde{z}(t)$ (см. рисунок). Множество вершин $\tilde{V}(t)$ однозначно определяется множеством элементарных объектов $O(t)$, а множество ребер и дуг — характером взаимодействия между ними, которое описывается множеством внутренних переменных $j_i(t)$.

Обозначим граф, представляющий объект $\tilde{o}_i(t)$, через $G_i(t)$, а множество всех графов, описывающих все объекты предметной области, — через $G(t)$. Каждому объекту $\tilde{o}_i(t)$ поставим в соответствие функцию $\omega_i(t)$, которая определяет степень важности объекта. Множество $W(t)$ описывает струк-



Граф неэлементарного объекта

туру распределения влияния объектов на показатели эффективности функционирования предметной области P .

Используя теоретико-множественные операции объединения и пересечения над множествами вершин графов объекта и естественного переноса ребер и дуг, определим операции *объединения* и *пересечения* над объектами, которые должны обладать свойством замкнутости.

Справедливо утверждение:

для любой предметной области множество всех ее объектов образует над множеством всех ее элементарных объектов решетку относительно обобщенных теоретико-множественных операций пересечения и объединения графов объектов.

Справедливость утверждения очевидным образом следует из однозначности обобщенных операций объединения и пересечения объектов и их эквивалентности обычным операциям объединения и пересечения подмножеств заданного множества.

Таким образом, предметная область P может быть описана набором функциональных множеств, отношений принадлежности, взаимодействия и динамической зависимости $\{X(t), Y(t), Z(t), J(t), O(t), \tilde{O}(t), G(t), W(t)\}$, которые описывают множество $\tilde{O}(t)$ всех объектов предметной области, структуру отношений между ними, закономерности их взаимодействия и динамику изменения переменных, характеризующих объекты, как во времени и пространстве, так и в зависимости от изменения переменных других типов.

При выделении предметной области и объектов, как правило, преследуется цель решить некоторое многообразие задач, которые могут носить как чисто теоретический характер, так и иметь сугубо прикладную ориентацию. Это не исключает смешанного характера проблематики формулируемых проблем. Независимо от направленности поставленных проблем априорные знания о предметной области, как правило, весьма ограничены, неполны и обладают значительной степенью неопределенности. Эти обстоятельства в значительной мере и обуславливают необходимость создания баз данных и инструментальных средств их анализа и обработки.

Ограниченность априорных знаний о предметных областях и их объектах обусловлена отсутствием точной информации о множествах $X(t), Y(t), Z(t), J(t), O(t), \tilde{O}(t)$, законах распределения многомерных случайных величин $X(t), Y(t), Z(t), J(t)$, о представлении в информационных моделях всех элементарных объектов, об операциях представления сложных объектов из элементарных объектов либо объектов меньшей степени сложности, о динамике изменения информационного обмена между объектами различной степени сложности. Термин *ограниченность* подразумевается не в смысле неполноты знаний, а в смысле ограниченности возможностей сколь угодно глубоко и детально представлять информацию о процессах, протекающих в предметной области P и внешней среде S , о структуре как самой предметной области P , так и ее внутренних и внешних возмущающих воздействиях.

Априори отсутствует информация, например, обо всех входных переменных, образующих множества $X(t), Y(t), Z(t), J(t)$, о том, конечны или бесконечны эти множества. Если онтологически хотя бы одно множество бесконечно, это может иметь решающее значение для получения знания о предельных свойствах различных его бесконечных подмножеств, т.к. базы данных имеют ограничения в плане представления некоторых объектов предметной области из-за бесконечного характера представления некоторых объектов.

Выделение предметных областей априори связано с неопределенностями различных классов. Базы данных как раз и создаются как инструмент накопления данных, анализ и обработка которых позволяет определять характер неопределенности, уменьшать ее степень либо устранять полностью.

Конструктивное проектирование баз данных и разработка стратегии их развития в значительной степени зависит от глубины знаний о характере, структуре и законах (механизмах) формирования неопределенностей в выделенной предметной области и ее объектах. Можно выделить следующие наиболее часто встречающиеся и связанные с любым вариантом выделения предметных областей классы неопределенностей:

- знаний законов изменения динамических переменных, описывающих выделенные свойства;
- механизмов взаимодействия предметных областей и среды, в которую они погружены;
- структур элементарных объектов;
- теоретико-графовых моделей предметных областей как структур, состоящих из множества объектов, взаимодействующих между собой по определенным законам;

— множества критериев, определяющих выделенную стратегию управления предметной областью;

— механизмов образования объектов из элементарных объектов, т.е. неопределенность определяющих соотношений, позволяющих создавать любой объект из некоторого множества элементарных объектов и последовательности операций;

— шкал измерений значений переменных, отражающих свойства предметной области или ее объектов.

В соответствии с предложенной системой классификации неопределенностей наиболее важным является построение баз данных, адекватно отражающих информационные процессы в предметных областях и объектах. Базы данных являются информационными моделями предметных областей в силу того, что каждый атрибут является случайным процессом, описывающим информационный обмен между предметной областью и средой, между предметной областью и управляющей системой или между объектами предметной области.

Соответственно необходимо уметь оценивать адекватность информационных моделей их предметным областям.

Можно показать, что по мере развития базы данных как информационные модели в пределе сходятся к состоянию, которое адекватно отражает информационные процессы в предметной области и ее объектах. Пусть переменная t представляет фактор времени, $r(a)$ — произвольное реляционное отношение со схемой $A = \{A_1 = X_1(t), A_2 = X_2(t), \dots, A_k = X_k(t)\}$, где вектор $X_j'(t) = (x_{j1}(t), \dots, x_{jk}(t))$ соответствует активному домену атрибута A_j . Вектор $X_i(t) = (x_{i1}(t), \dots, x_{ki}(t))$ представляет i -ю запись (кортеж) реляционного отношения.

Если все атрибуты измеряются по количественной шкале измерений, то без потери общности можно предполагать, что $x_{ji}(t)$ представляет собой выборочную функцию случайного процесса $\xi_{ji}(t)$. Записи в общем случае могут регистрироваться в банках данных с неравномерным шагом дискретизации времени. Ограничившись рассмотрением случая, когда t измеряется с шагом δ , тогда реляционное отношение $r(A)$ содержит N записей.

Предполагается, что оценивание адекватности БД предметной области как ее информационной модели осуществляется на m -м шаге последовательного анализа. В качестве критерия адекватности используется сумма квадратов разности по всем атрибутам реляционного отношения между прогнозными значениями $x_{1,N+1}(t), x_{2,N+1}(t), \dots, x_{k,N+1}(t)$ и их измеренными значениями $\tilde{x}_{1,N+1}(t), \tilde{x}_{2,N+1}(t), \dots, \tilde{x}_{k,N+1}(t)$ в момент времени $t = t_0 + \delta(N+1)$.

В качестве критерия адекватности выберем случайную величину, определяемую соотношением $\chi_{N+1}^2 = \frac{1}{K} \sum_{i=1}^K (x_{i,N+1}(t) - \tilde{x}_{i,N+1}(t))^2$, где K — степень отношения $r(A)$.

Случайная величина χ_{N+1}^2 с определенной степенью приближения подчиняется закону распределения χ^2 в том случае, когда случайные величины $\tilde{x}_1, \tilde{x}_2, \dots, \tilde{x}_k$ нормально распределены и независимы. Справедливость этого предположения просто проверяется на основе методов проверки статистических гипотез [2].

С использованием методов проверки статистических гипотез, по заданной величине α вероятности справедливости гипотезы определяется критическое значение χ_α^2 . Если $\chi_{N+1}^2 < \chi_\alpha^2$, можно утверждать, что реляционное отношение адекватно представляет (описывает) информационные процессы в соответствующем ему объекте предметной области. В противном случае реляционное отношение требует дальнейшего наполнения данными.

Литература

1. Цаленко М.Ш., Моделирование семантики в базах данных. — М.: Наука, 1989. — 274 с.
2. Гнеденко Б.Г. Курс теории вероятностей. — 7-е изд., испр. — М.: УРСС. — 2001. — 318 с.
3. Востров Г.Н., Малахов Е.В., Кулешов В.Н. Проблемы описания структуры предметных областей // Тр. Одес. политехн. ун-та. — Одесса, 2000. — Вып. 2(11). — С. 111 — 114.