

УДК 004.62:004.492

Е.В. Малахов, Г.Н. Востров, М.Г. Микулинская  
Одесский Национальный Политехнический Университет 65044, г. Одесса, проспект Шевченко, 1

## МЕТОДЫ ОПРЕДЕЛЕНИЯ СТЕПЕНИ ВАЖНОСТИ СВОЙСТВ СУЩНОСТЕЙ ПРЕДМЕТНЫХ ОБЛАСТЕЙ.

*Рассмотрены методы определения важности (информативности) свойств объектов предметных областей с целью построения критериев оценки адекватности моделей самих объектов и, в дальнейшем, оценки адекватности моделей предметных областей в целом.*

**Ключевые слова:** База данных — Модель предметной области — Оценка адекватности — Информативность свойства

*Methods of properties importance (comprehension) definition of objects of subject domains for the purpose of adequacy estimation criteria construction of objects models and an estimation of models adequacy of subject domains as a whole further are considered.*

**Keywords:** Database — Subject domain model — Adequacy estimation — Property comprehension

### I. ВВЕДЕНИЕ

Современные системы организационного управления создаются для информационной поддержки управленческих решений в различных областях науки и техники. Построение системы организационного управления, как и любой информационной системы (ИС), начинается с создания модели предметной области (ПрО). Как правило, выбор сущностей предметной области, которые необходимо ввести в информационную модель, осуществляется разработчиком интуитивно. Кроме того, ПрО являются динамичными системами, поэтому в процессе работы уже построенной ИС возникает необходимость уточнения или расширения модели ПрО, расширения информации об объектах, которые отражены в модели. Благодаря этому и при проектировании, и при сопровождении ИС постоянно возникает целый ряд сложных проблем, которые приводят к нарушению адекватности модели ПрО, лежащей в основе этой ИС. Поэтому уже после первой итерации построения баз данных (БД) или информационных хранилищ (ИХ) возникает необходимость оценить степень адекватности модели ПрО на основе спроектированной системы БД.

Если при построении системы БД следовать математической теории исключительно реляционных БД, то ряд проблем построения математических моделей ПрО останутся нерешенными [1]. Как следствие, система БД будет неадекватной математической моделью ПрО. Кроме того, полученная информационная или структурная модель ПрО имеет довольно высокие шансы быть плохо интерпретируемой с содержательной точки зрения. Для такого пессимистического взгляда на систему БД, полученных на первой итерации, существуют серьезные факторы.

Прежде всего, отсутствуют методы оценивания степени отображения в системе БД информационной полноты, определенности, непротиворечивости представления самих ПрО. Данные, полученные на первой итерации, несложно организовать в виде реляционных отношений. Однако такая система реляционных отношений, как правило, неполна. Для нее характерны многочисленные неопределенности, размытости и противоречия. Кроме того, информация, которая хранится в них, во многом носит отражение субъективных и, следовательно, неточных выводов и выводов на основе неполных, нечетких, а временами, противоречивых соображений экспертов, пользователей и лиц, принимающих решения. Поэтому построение математической модели ПрО является необходимой мерой, способствующей обеспечению необходимого уровня адекватности [2].

Второй важный фактор связан с отсутствием в теории БД математических средств упорядочения ПрО, экземпляров и свойств универсальных сущностей по степени их важности при построении математических моделей ПрО. Причем речь идёт о свойствах именно универсальных сущностей, определённых в [3], т.к. существует ещё одна проблема: при выполнении математических операций над метамоделями ПрО необходимо иметь объективные факторы или средства идентификации объектов ПрО как проекций одной и той же или разных универсальных сущностей на различные ПрО.

Решение этих проблем предлагается выполнять на основе прогнозирования значений свойств универсальных сущностей или их проекций на ПрО и оценки их важности.

## II. ПОЛУЧЕНИЕ МНОЖЕСТВА НАИБОЛЕЕ ВАЖНЫХ СВОЙСТВ СУЩНОСТЕЙ

Предположим, что для реляционной схемы  $R_i(a_{1i}(t), a_{2i}(t), \dots, a_{mi}(t))$  построена таблица, значения которой получены в результате  $K$  статистических испытаний, т.е. содержат  $K$  кортежей (рис. 1).

В этой таблице сумма  $\sum_{j=1}^q K_j = K$ , где  $q$  соответствует количеству проведенных итераций и определяет количество проведенных над ПрО испытаний. При этом предполагается, что  $j$ -я итерация содержит  $K_j$  испытаний. Предполагается также, что с увеличением объема выборки растет ее репрезентативность.

На первой итерации выполняется  $K_1$  статистических испытаний над ПрО, которые содержат данные, полученные от пользователей информационной системы, создаваемой на основе выделенной ПрО.

Следует отметить, что для различных универсальных сущностей, выделенных в качестве наиболее важных с одновременным выделением наиболее важного конечного множества их свойств, количество проведенных испытаний может различаться.

$R_i$	$a_{1i}(t)$	$a_{2i}(t)$	...	$a_{mi}(t)$	
$K_1$	1	$x_{1i}^1(t)$	$x_{2i}^1(t)$	...	$x_{mi}^1(t)$
	2	$x_{1i}^2(t)$	$x_{2i}^2(t)$	...	$x_{mi}^2(t)$
	3	$x_{1i}^3(t)$	$x_{2i}^3(t)$	...	$x_{mi}^3(t)$
	⋮	⋮	⋮	...	⋮
	$K_1$	$x_{1i}^{K_1}(t)$	$x_{2i}^{K_1}(t)$	...	$x_{mi}^{K_1}(t)$
$K_2$	$K_1 + 1$	$x_{1i}^{K_1+1}(t)$	$x_{2i}^{K_1+1}(t)$	...	$x_{mi}^{K_1+1}(t)$
	$K_1 + 2$	$x_{1i}^{K_1+2}(t)$	$x_{2i}^{K_1+2}(t)$	...	$x_{mi}^{K_1+2}(t)$
	⋮	⋮	⋮	...	⋮
	$K_1 + K_2$	$x_{1i}^{K_1+K_2}(t)$	$x_{2i}^{K_1+K_2}(t)$	...	$x_{mi}^{K_1+K_2}(t)$
	$K_1 + K_2 + 1$	$x_{1i}^{K_1+K_2+1}(t)$	$x_{2i}^{K_1+K_2+1}(t)$	...	$x_{mi}^{K_1+K_2+1}(t)$
⋮	⋮	⋮	...	⋮	
$K_q$	$1 + \sum_{j=1}^{q-1} K_j$	$x_{1i}^{1+\sum_{j=1}^{q-1} K_j}(t)$	$x_{2i}^{1+\sum_{j=1}^{q-1} K_j}(t)$	...	$x_{mi}^{1+\sum_{j=1}^{q-1} K_j}(t)$
	⋮	⋮	⋮	...	⋮
	$K - 1$	$x_{1i}^{K-1}(t)$	$x_{2i}^{K-1}(t)$	...	$x_{mi}^{K-1}(t)$
	$K$	$x_{1i}^K(t)$	$x_{2i}^K(t)$	...	$x_{mi}^K(t)$

**Рис. 1.** Реляционное отношение  $R_i$ , полученное в результате  $K$  статистических испытаний

Более того, не исключены случаи, когда некоторые универсальные сущности могут не быть объектами статистического анализа. Для этого существует несколько причин. Содержательные представления всех представителей множества  $H(t)$  не имеют полной информации обо всех универсальных сущностях, где  $H(t)$  — множество активных объектов физического или виртуального мира, которые целенаправленно воздействуют друг на друга и на другие объекты этого мира или генерируют объекты интеллектуального мира  $J$  [4]. Этот факт на протяжении продолжительного времени может быть одной из ключевых причин неполноты создаваемых систем БД.

Другим источником неполноты, а заодно, и неопределенности является значительная вычислительная сложность выбора конечного множества наиболее информативных свойств универсальных сущностей и определение с приемлемой точностью законов их распределения.

При этом будем считать, что используемые инструментальные средства позволяют получить приемлемое качество моделирования для ПрО, имеющих максимальную меру полезности.

Кроме того, часть важных универсальных сущностей, для которых достигнуто приемлемое качество моделирования, составляет величину  $\Delta \geq \alpha$ , где  $\alpha$  — заданный порог глубины моделирования ПрО. Под глубиной математического моделирования будем понимать минимальное количество универсальных

сущностей, имеющих степень важности  $\rho \geq \beta$ , где  $\rho$  — важность универсальной сущности, а  $\beta$  — нижний порог, определяющий важность универсальных сущностей, ниже которого математическое моделирование становится нецелесообразным.

Вернемся к отношению  $R_i$  на рис. 1. Будем считать, что после  $q$  итераций в реляционном отношении существует  $K = \sum_{i=1}^q K_i$  кортежей. В результате применения математических инструментальных средств на каждом шаге итерации могли изменяться (добавляться или удаляться) содержащиеся в отношениях атрибуты, в связи с изменением оценки их информационной ценности. И, конечно, в отличие от традиционных технологий формирования БД, в данном подходе для каждого кортежа одновременно с включением в БД каждого его значения включается момент времени  $t$ , в который это значение было получено.

Этот процесс продолжается до тех пор, пока для избранной универсальной сущности не будет полученное такое множество атрибутов, которое с необходимой полнотой описывает данную сущность, и при этом каждый атрибут, включенный в схему отношений, имеет информационную ценность, превосходящую информационную ценность любого атрибута, исключенного со схемы отношений на предыдущих шагах итерации. Таким образом, каждый шаг итерации использует методы и алгоритмы определения информационной ценности атрибутов и вычисление степени полноты приближенного описания рассмотренной универсальной сущности.

Предположим, что на  $(p+1)$ -й итерации получено устойчивое приближенное описание выделенной универсальной сущности. При этом на  $p$ -й итерации к описанию добавлялся хотя бы один новый атрибут. Свойство полноты и максимальной информативности выделенных атрибутов обеспечивает стойкость полученного описания универсальной сущности. Это не исключает дальнейших итераций построения приближенной математической модели ПрО. Например, достигши стойкости описания, система моделирования может не иметь приемлемую математическую модель описания закона распределения каждого атрибута. Поэтому итерационный процесс должен быть продолжен дальше.

Продолжение итерационного процесса не угрожает необходимостью радикальной перестройки схемы устойчивых отношений. Он может продолжаться только для тех объектов, для которых не получено их устойчивое описание.

Предположим, что для рассмотренной  $i$ -й универсальной сущности в момент времени  $t$  уже полученное устойчивое описание, удовлетворяющее условиям полноты и выделения наиболее важных свойств. Т.е. свойств, обеспечивающих малую вероятность возникновения условий, которые требуют поиска более информативного свойства, чем любое подмножество свойств, прежде включенных в схему отношения. Согласно отмеченным предположениям множество атрибутов  $\{a_{1,i}(t), a_{2,i}(t), \dots, a_{k,i}(t)\}$  имеет свойство устойчивости. Поэтому все предыдущие итерации для данной универсальной сущности не представляют интереса и исключены из рассмотрения. Однако информация ПрО этих итерациях может сохраняться для других универсальных сущностей, для которых полное и устойчивое описание в виде их важных информативных свойств было получено на более ранних итерациях.

Таким образом, с целью упрощения описания методов и алгоритмов проверки адекватности приближенной математической модели ПрО будем считать, что для каждой важной универсальной сущности, у которой вероятность появления в статистических испытаниях не меньше заданного порога  $\beta$ , получено устойчивое описание. На рис. 1 приведен пример представления реляционного отношения, данные которого, вместе с разработанными математическими инструментальными средствами и другими данными об универсальных сущностях, позволяют проверить адекватность приближенной математической модели ПрО.

Проверка адекватности математической модели содержит в себе следующие критерии:

- критерий проверки, к какому классу математических зависимостей принадлежит какой-либо атрибут как функция времени или как функция некоторого множества других атрибутов;
- критерий точности восстановления функциональной зависимости при использовании различных математических методов;
- критерий определения эффективности прогнозирования значений отдельных атрибутов на краткосрочную и долгосрочную перспективу;
- критерий оценки эффективности прогнозирования динамики изменения универсальной сущности по полной системе выделенных наиболее информативных атрибутов.

Проверка адекватности математической модели ПрО строится на прогнозировании в каждом реляционном отношении значений отдельных атрибутов, подмножеств атрибутов и всей совокупности атрибутов на момент времени  $T + \Delta t$ , где  $\Delta t$  — дискретный интервал времени, через который будет получено значение очередного нового кортежа. Если время получения значений атрибутов дежурного кортежа носит случайный характер, то прогноз осуществляется на основе информации, содержащейся в

$$\sum_{i=1}^q K_i - 1 \text{ кортежах для кортежа } \sum_{i=1}^q K_i .$$

Аналогично прогноз может осуществляться на момент времени  $T + \Delta t_i$  для очередного кортежа, получаемого в этот момент времени при условии, что точно известны  $\Delta t_i$ , т.е. интервалы между моментами времени получения информации. Отсюда следует, что возможны два случая при прогнозировании на  $m$  шагов развития ПрО и ее объектов. В первом случае прогноз осуществляется на основе информации, содержащейся в  $\sum_{i=1}^q K_i$  кортежах, и распространяется на  $m$  следующих наблюдаемых кортежей, о которых информация в БД отсутствует, но точно известны моменты времени их появления. При этом к моменту времени  $T$  уже получена информация  $\sum_{i=1}^q K_i$ . Прогноз осуществляется на моменты времени  $(T + \Delta t_1)$ ,  $(T + \Delta t_2)$ , ...,  $(T + \Delta t_m)$ ... . В эти же моменты времени наблюдаются реальные значения атрибутов. Величина отличия между ними и определяет степень адекватности модели на уровне данной универсальной сущности.

В том случае, когда интервалы времени  $\Delta t_1, \Delta t_2, \dots, \Delta t_m$  и, соответственно, моменты времени получения новых значений используемых атрибутов рассмотренной сущности неизвестны и не могут быть оценены даже приблизительно, используется уже накопленная в БД информация. Для прогнозирования используются первые  $\sum_{i=1}^q K_i - m$  кортежей. Прогноз осуществляется для всех моментов времени  $T + \sum_{i=1}^m \Delta t_i$  для  $m$  следующих кортежей  $\left(\sum_{i=1}^q K_i - m\right) + p$ , где  $p = \overline{1, m}$ , значения которых уже известны. При этом применяется та же мера адекватности.

Два приведенных случая имеют место при следующих предположениях. Все множество полученных  $\sum_{i=1}^q K_i$  кортежей содержит кортежи, которые, в свою очередь, не содержат значений неопределенного типа, или кортежи с неточно измеренными значениями атрибутов. Т.е., можно утверждать, что каждое реляционное отношение с помощью математических инструментальных средств построения приближенной математической модели ПрО очищено от некорректных кортежей. Рассмотренные случаи могут иметь место тогда, когда в реляционных отношениях существуют атрибуты, которые принимают значения только в определенные моменты времени. При таком условии восстановление функциональной зависимости отдельных свойств от времени и других свойств в математической модели ПрО приводит к функциям, имеющим дискретную форму.

Возникновение таких свойств связано чаще всего с атрибутами, измеренными в порядковой, балльной или номинальной шкалах. При таких условиях для некоторых атрибутов восстановление функциональных зависимостей свойств универсальной сущности от фактора времени или от свойств, измеренных в этих шкалах, в непрерывной форме является практически невозможным. Это обстоятельство и обуславливает необходимость использования двух рассмотренных вариантов, когда кортежи изменяются не только дискретно, но и определено системой интервалов времени, задающих моменты измерения значений свойств.

Приведенная схема прогнозирования появления экземпляров универсальных сущностей в процессе их динамического развития направлена на построение методов оценивания адекватности приближенных математических моделей ПрО в информационных системах. В результате прогнозирования на один интервал времени  $\Delta t_1$ , который связан с возможностью наблюдения дежурного экземпляра универсальной сущности с вероятностью, близкой к 1, получаем прогнозное и реальное значения атрибутов.

Если прогнозные значения выделенных наиболее информативных свойств сущности отличаются от реальных не более, чем на заданную величину, которая соответствует построенному критерию, и это верно для всех важных включенных в математическую модель универсальных сущностей, можно считать, что модель адекватно отображает все представления ПрО. В том случае, когда уровень адекватности сохраняется при прогнозировании на  $m$  экземпляров в каждой универсальной сущности, можно утверждать, что степень адекватности имеет глубину  $m$ . Пополнив систему БД реальной информацией о новых испытаниях над универсальными сущностями и скорректировав математическую модель ПрО с помощью созданных математических инструментальных средств, можно снова оценить адекватность модели по той же схеме.

### III. МЕРЫ ИНФОРМАЦИОННОЙ ЦЕННОСТИ СВОЙСТВ СУЩНОСТЕЙ

Предположим, что рассматривается  $r$ -я универсальная сущность  $E_r(t)$ , которая описывается выделенным с помощью инструментальных средств множеством атрибутов  $E_r(t) = \{A_{1r}(t), A_{2r}(t), \dots, A_{mr}(t)\}$ , имеющих максимальную информационную ценность. Информационная ценность этих свойств, вычисляется на основе энтропийных методов [5]. Если  $j$ -е свойство  $r$ -й универсальной сущности  $A_{jr}(t)$  является непрерывной случайной величиной, то энтропия, которая является теоретико-информационной мерой степени неопределенности случайной величины, определяется выражением:

$$H(A_{jr}(t)) = - \int_0^{\infty} f(A_{jr}(t)) \log_2 f(A_{jr}(t)) dA_{jr}(t), \quad (1)$$

где  $f(A_{jr}(t))$  — функция плотности распределения вероятности значений свойства (атрибута)  $A_{jr}(t)$ .

В том случае, когда свойство (атрибут)  $A_{jr}(t)$  носит дискретный характер и на интервалах  $\Delta t_1, \Delta t_2, \dots, \Delta t_m$  принимает значения с вероятностями  $P_1(A_{jr}(\Delta t_1)), P_2(A_{jr}(\Delta t_2)), \dots, P_m(A_{jr}(\Delta t_m))$ , энтропия этого свойства определяется выражением:

$$H(A_{jr}(t)) = - \sum_{i=1}^m P_i(A_{jr}(\Delta t_i)) \log_2 P_i(A_{jr}(\Delta t_i)). \quad (2)$$

Множество  $\{H(A_{1r}(t)), H(A_{2r}(t)), \dots, H(A_{mr}(t))\}$  рассматривается как совокупность неопределенности мер  $m$  выделенных атрибутов. Если атрибуты выделенного множества статистически независимы, то чем меньшее значение  $H(A_{jr}(t))$ , тем выше информационная ценность свойства  $A_{jr}(t)$ . Множество из  $m$  атрибутов будем относить к классу *наиболее важных* для данной универсальной сущности при выполнении следующих условий:

- атрибуты множества независимы или имеют слабую стохастическую зависимость;
- любой атрибут, не принадлежащий к данному множеству, может быть представлен в виде функциональной зависимости от некоторого подмножества этого множества атрибутов;
- суммарная энтропия для заданного множества определяет минимальное значение среди множеств из  $m$  других атрибутов.

Атрибуты множества  $\{A_{1r}(t), A_{2r}(t), \dots, A_{mr}(t)\}$  не являются в общем случае независимыми. Рассмотрим случай, когда все атрибуты принадлежат интервальной шкале измерения.

Пусть прогнозные значения атрибутов, описывающих  $r$ -ю универсальную сущность для  $(K+S)$ -го экземпляра в момент времени  $t_g$ , представлены множеством значений  $\{\bar{A}_{1r}^{K+S}(t_g), \bar{A}_{2r}^{K+S}(t_g), \dots, \bar{A}_{mr}^{K+S}(t_g)\}$ , а полученные реальные значения свойств этого экземпляра — множеством  $\{A_{1r}^{K+S}(t_g), A_{2r}^{K+S}(t_g), \dots, A_{mr}^{K+S}(t_g)\}$ . Для того, чтобы можно было сравнить точность прогнозирования для различных универсальных сущностей будем считать, что множество значений информационной ценности выделенных свойств нормировано. При этом мера информационной ценности  $m$  атрибутов определяется выражением:

$$I(A_{1r}(t), A_{2r}(t), \dots, A_{mr}(t)) = \sum_{j=1}^m \sum_{i=1}^m I(A_{jr}(t), A_{ir}(t)), \quad (2.18)$$

где  $I(A_{jr}(t), A_{ir}(t)) = H(A_{jr}(t)) + H(A_{ir}(t)) - H(A_{jr}(t), A_{ir}(t))$ .

Энтропия пары атрибутов  $H(A_{jr}(t), A_{ir}(t))$  определяется общим законом распределения этой пары при условии, что любой атрибут, не принадлежащий к этому множеству, может быть представлен в виде линейной комбинации атрибутов его некоторого подмножества с необходимой точностью. Данное условие может быть обобщено на случай, когда рассматривается не аддитивная, а аддитивно-мультипликативная комбинация заданного типа.

### IV. ВЫВОДЫ

Мера ценности для любого свойства всегда неотрицательная. При этом условии множество

значений информационной ценности выделенных свойств может быть всегда проанализировано.

Приведенные меры ценности позволяют выделить конечную совокупность наиболее важных универсальных сущностей ПрО, входящих в ее объектное ядро, при построении приближенных математических моделей ПрО. Для каждой универсальной сущности выделяется конечное множество ее наиболее информативных свойств, образующих ее описание с необходимой степенью приближения.

## ЛИТЕРАТУРА

1. *Цаленко М.Ш.* Моделирование семантики в баз данных. – М.: Наука, 1989. – 287 с.
2. *Малахов Е.В.* Оценка степени адекватности баз данных как информационных моделей предметных областей [текст] // Тр. Одес. политехн. ун-та. – 2004. – Вып. 1(21). – С. 82 – 86.
3. *Малахов Е.В.* Манипулирование метамоделями предметных областей [текст] // Восточно-европейский журнал передовых технологий. – Харьков, 2007. – Вып. 5/3(29). – С. 6 – 10.
4. *Малахов С.В.* Виділення складноструктурованих предметних областей [текст] // Матеріали Міжнародної науково-технічної конференції «Сучасні методи, інформаційне, програмне та технічне забезпечення систем управління організаційно-технологічними комплексами», Київ: НУХТ, 26-27 листопада 2009. – С. 79 – 80.
5. *Гмурман В.Е.* Теория вероятностей и математическая статистика. – М.: Высшая школа, 1999. – 432 с.