МИНИСТЕРСТВО ОБРАЗОВАНИЯ УКРАИНЫ

КИЕВСКИЙ ГОСУДАРСТВЕННЫЙ ЭКОНОМИЧЕСКИЙ УНИВЕРСИТЕТ

на правах рукописи

АМАНА НАХУДА ИБРАХИМ

ОПТИМИЗАЦИЯ СТРУКТУР БАЗ ДАННЫХ

Специальность 08.00.13. - Экономико-математические методы

А В Т О Р Е Ф Е Р А Т

диссертация на соискание ученой степени кандидата экономических наук

Киев 1993

Работа выполнена на ка
ковском инженерно-экономичес

Научный руководитель – доктор технических наук, профессор Волколупова Р.Т.

Официальные оппоненты: Доктор экономических наук, профессор Суслов О.П.

Кандидат экономических наук, доцент Витлинский В.В.

Ведущая организация: Главный научно-исследовательский институт по проблемам информатики Министерства экономики Украины

Защита диссертации состоится "24" июня 1993 года в 14 часов на заседании Специализированного совета К.068.28.05. в Киевском государственном экономическом университете /252057, г.Киев-57, Проспект Победы, 54/I, ауд. 214 /.

С диссертацией можно ознакомиться в библиотеке университета.

Автореферат разослан "24" мая 1993 года.

Ученый секретарь Специализированного совета В.П.Кулагина

I. ОБЩАЯ ХАРАКТЕРИСТИКА

I.I. Актуальность темы. Одной из важнейших задач любой системы управления является переработка информации. Важную роль в переработке информации призвана сыграть технология создания баз данных. Важнейшим элементом этой технологии является процесс выбора структуры базы данных. Выбор надежного представления такой структуры по тому или иному критерию составляет задачу логического проектирования, решение которой позволит оптимизировать базу данных, что обеспечит эффективность функционирования информационной системы по времени и стоимости.

Постоянный поиск методов, позволяющих снизить затраты на стоимость хранения данных и их переработку, послужил объективной необходимостью углубления исследований по данным вопросам и определил актуальность темы диссертационной работы.

I.2. Цель и задачи исследования. Целью диссертационной работы является исследование, разработка методов, алгоритмов и программ решения задач оптимизации структур баз данных и построениеоптимальной логической структуры.

Достижение поставленной цели обеспечивается решением следующих задач:

I. Создание модифицированного метода динамического сгущения.

2. Разработка метода и алгоритма исключения нулевых элементов.

3. Построение альтернативного множества.

4. Нахождение локального оптимума.

5. Создание модифицированного метода выбора системы отношений между элементами информации.

6. Выделение классов однотипных элементов.

7. Создание метода и алгоритма исключения бесперспективных вариантов.

8. Разработка метода и алгоритма синтеза оптимальной структуры базы данных.

I.3. Методы исследования. При выполнении данной работы применялись методы исследования операций, векторной оптимизации, кластер-анализа, теории множеств, теории матриц, теории графов.

I.4. Научная новизна результатов исследований:

- разработаны методы: модификация динамического сгущения, исключения нулевых элементов, модификация метода выбора системы отношений между элементами информации, исключения бесперспективных вариантов, векторной оптимизации;

- созданы алгоритмы и программы вышеперечисленных методов.

I.5. Практическая значимость. Разработанные методы и алгоритмы могут быть использованы для данных, позволяя оптимизировать структуры всех типов данных: иерархических, сетевых, реляционных и структуру представленную в виде гиперграфов.

I.6. Реализация результатов работы. Основные результаты работы использованы текстильным заводом СОТЕМА на Мадагаскаре для перспективного развития информационных систем (свидетельство посольства Республики Мадагаскар N318 AMBAMOSC/AFCULT).

I.7. Апробация работы. Достоверность полученных результатов и выводов подтверждена расчетом специальных математических критериев, апробацией методов оптимизации логической структуры баз данных на тестовых примерах, докладами на учредительной конференции международной ассоциации по нетрадиционным методам оптимизации (г.Дивногорск, I6 марта I992 г.), в Междуна-

родной школе "Проектирование автоматизированных систем контроля и управления сложным объектом" (г.Туапсе 17 сентября 1992 г.), а также регулярно в течение всего периода выполнения исследований на заседаниях семинара "Проблемы экономической кибернетики" Академии наук Украины (г.Харьков, ХИЭИ, кафедра информационных систем), кроме того доклад по теме диссертации направлялся во Францию на конференцию: 10ème conférence internationale sur l'analyse et l'optimisation des Systèmes approchés Fréquentielles et temporelles des Systèmes de Dimension infinie, 9-12 Juin 1992, SOPHIA-ANTIPOLIS (FRANCE) и был принят экспертной комиссией.

I.8. Публикации. По теме диссертации опубликовано 3 печатных работы.

I.9. Структура и объем диссертации. Диссертационная работа состоит из введения, трех глав, выводов по главам, заключения, списка использованной литературы и трех приложений, 7 рисунков, 8 таблиц. Список литературы включает 124 наименования, из них 9 на иностранных языках.

2. СОДЕРЖАНИЕ РАБОТЫ

Во введении обосновываются актуальность темы, выбор предмета исследований, научная новизна и практическая значимость полученных результатов, формулируются основные направления исследований.

Первая глава посвящена критическому анализу некоторых существующих методов оптимизации логической структуры баз данных, а также сформулированы задачи исследования.

В первых шагах проектирования структур баз данных требуется определение состава информационных массивов. Традицион-

ный метод сортировки массивов требует большого перебора и более того, не устраняет дублирования информации, что приводит к неоптимальной логической структуре базы данных, а значит к ухудшению качества управления.

Информация, подлежащая обработке при решении задач, представляет собой информационный массив о некотором множестве элементов исследуемого объекта. Данные о каждом отдельном элементе такого множества составляют определенную совокупность, которую будем называть информационным модулем (ИМ). Информационный модуль расчленяется на отдельные части, называемые элементарными информационными модулями (ЭИМ) и представляющие собой последовательности символов определенного алфавита. Отдельные ЭИМ несут законченную смысловую нагрузку, соответствующую качественным и количественным характеристикам элементов объектов, описываемым информационным модулем. Тогда для устранения ранее перечисленных недостаток требуется образовать альтернативные модули наименьшей мощности из всех возможных скомпанованных по исходным данным ИМ и ввести оптимизацию этого альтернативного множества согласно соответствующим критериям.

В целях получения набора альтернативных информационных модулей предлагается разработать модифицированный метод динамического сгущения и метод исключения нулевых элементов, целью которых является устранение дублирования информации в покрытии и снижение затрат на вычисление при работе с разреженной матрицей. Выбор рациональных по составу и количеству альтернативных ИМ осуществляется с помощью метода нахождения локального оптимума.

Все эти предлагаемые разработанные методы направлены на повышение качества управления.

Рассмотренный метод приведения отношений к третьей норма-

льной форме затрудняет проектирование подсхем базы данных, поскольку этот метод не учитывает наличие всех решаемых задач. В этой связи, для устранения этого недостатка необходимо применить модифицированный метод выбора системы отношений.

Оценка логической структуры базы данных осуществляется методом многокритериальной оптимизации, компромисным путем, с разработанным методом исключения бесперспективных вариантов.

Во второй главе разработаны модификация метода динамического сгущения, метод исключения нулевых элементов, применен метод локального оптимума, разработан модифицированный метод выбора системы отношений, сформулирована математическая модель названных методов, применен подход к построению альтернативного множества и приведены результаты действия алгоритмов этих методов.

Введем некоторые понятия и обозначения.

Множества элементов, определяющее ИМ, необходимое для решения задач Z_j $(j=\overline{1,s})$ обозначим как $\Omega=\{P_1,P_2,\dots,P_r,\dots,P_k\}$.

Множество непустых подмножеств $P_\ell\subset\Omega$, $(\ell=1,2,\dots,\leq k)$, объединения которых заполняют Ω, назовем его покрытием Ω' со свойствами:

1. $\forall \ell \in \{1,2,\dots,k\}$, $\ell \leq k$, $P_\ell \neq \emptyset$,
2. $\bigcup_{\ell=1}^{k} P_\ell = \Omega'$.

Тогда совокупность $P_i\subset\Omega'$ $(i\leq\ell)$, попарное пересечение которых со всеми совокупностями P_j $(j\leq\ell,\ i\neq j)$ образует пустые множества, определяет разбиение Ω' на классы P_i, P_j, предельное количество которых равно n, вычисление которого происходит с помощью соответствующего алгоритма.

Свойства классов:

1. $P_i \cap P_j = \emptyset$, где $P_{i,j}=(P_s\cup\dots\cup P_m)\cap(P_k\cup\dots\cup P_r)$,

2. $\forall s, m, \kappa, \tau \in \{1, 2, \ldots, l\}$.

Проектирование информационных модулей можно сформулировать как задачу:

- определения рациональных по составу модулей в том смысле, что каждый информационный модуль содержит минимальные избыточные и максимальные необходимые элементарные модули для решения задач;

- определения минимального количества набора информационных модулей.

Данная задача разделена на две подзадачи.

Первая подзадача: образование альтернативных ИМ наименьшей мощности, а вторая - оптимизация этого альтернативного множества.

Для решения первой задачи предлагается применить модифицированный метод динамического сгущения, а вторую задачу предполагается решить с помощью метода нахождения локального оптимума.

Образуем исходную матрицу. Исходные данные, состоящие из $\Omega' = \{P_\tau\}$, необходимые для решения задач $Z = \{Z_j\}$ могут быть представлены матрицей информационных потребностей $M = (x_\tau^j)$ элементы которой $x_\tau^j = 1$, если $P_\tau \in \Omega_j$, и $x_\tau^j = 0$ в противном случае, где Ω_j - множество ЭИМ, которое необходимо для решения Z_j.

Как правило, слабозаполненная или разреженная матрица представляет собой матрицу с большим количеством нулей. Для хранения такой матрицы при расчетах потребуются чрезмерные затраты машинного времени и объема памяти.

Разработан метод и алгоритм исключения нулевых элементов, которые позволяют значительно снизить стоимость не только хра-

нения информации, но и стоимость затрат на вычисления.

Данный метод предполагает хранить в списке только значения индексов i и j матрицы, с ненулевыми элементами. Таким образом, он позволяет преобразовать исходную матрицу M на список значений индексов ненулевых элементов матрицы M .

Используем процедуру разбиения на классы. Разобъем исходную матрицу на блоки согласно следующих требований, т.е. определяя блоки наименьшей мощности и производя разбиение на минимальное количество блоков. Пусть, в качестве примера, исходная матрица выглядит следующим образом:

$M_1 =$

	Z_3	Z_4	Z_5	Z_6	Z_7	Z_1	Z_2	Z_8	Z_9	Z_{10}
P_3	0	0	0	0	0	1	0	1	1	0
P_6	0	0	0	0	0	0	1	0	1	0
P_7	0	0	0	0	0	0	1	1	0	1
P_4	1	0	1	0	0	1	0	1	0	1
P_5	0	1	0	1	1	0	1	0	1	0
P_8	1	0	1	0	0	1	0	1	0	1
P_9	0	1	0	0	0	1	0	0	0	1
P_1	1	0	1	0	0	1	0	1	0	1
P_2	0	1	0	1	1	0	1	0	1	0
P_{10}	0	1	0	0	1	0	1	0	0	0

Разбиение $P'=(P_1', P_2', P_3')$ множества Ω' и $Q'=(Q_1', Q_2')$ множества Z соответственно находятся по строкам и столбцам матрицы.

Тогда задача состоит в нахождении однородных пар классов (P_k', Q_ℓ'), т.е., либо заполненных единицами, либо нулями, и каждая пара (P_k', Q_ℓ') классов ассоциируется с банарным идеальным значением 0 или 1.

I. Таким образом, создается бинарная матрица по K строкам и L столбцам, так называемое ядро, и обозначается как $L=(a_k^l)$, где $a_k^l \in \{0,1\}$ $\forall k=\overline{1,K}$ и $\forall l=\overline{1,L}$; a_k^l есть представительство пар классов (P_k', Q_l').

Уточним понятие идеального значения. Когда в паре классов (P_k', Q_l') в матрице M_1 стоит большее количество единиц, чем количество нулей, то $a_k^l = 1$. В противном случае $a_k^l = 0$.

Исходя из вышеизложенного, суть модифицированного метода динамического сгущения состоит в определении элементов, которые могли бы служить представителями всех пар классов (P_k', Q_l'), элементов всех классов, таких чтобы сумма расстояния от всех пар классов (P_k', Q_l') до их представительств была минимальной.

Итак, математическая модель модифицированного метода динамического сгущения описывается следующим образом

$$\min W(P', Q', L) = \sum_{k=1}^{K} \sum_{l=1}^{L} \sum_{z_i \in P_k'} \sum_{z_j \in Q_l'} | x_\tau^j - a_k^l | , \qquad (I)$$

где $x_\tau^j = \begin{cases} 1, \text{ если } Z_j \text{ используем ЭИМ } P_\tau, \\ 0 \text{ в противном случае} \end{cases}$

$a_k^l = \begin{cases} 1, \text{ если пара } (K,l) \text{ классов заполнена единицами,} \\ 0 \text{ в противном случае.} \end{cases}$

Из (I) можно определить ряд $(H^n, L^n): H^0 = P'$ и $L = L^*(P', Q')$

Пусть две функции g и f соответственно для представительства и назначения.

Функция представительства предназначена для определения ядра или представительств

$$g: H^n \to L^n, \quad \text{где } L^n = g(H^n).$$

Функция назначения предназначена для отнесения элементарных информационных модулей либо задач к классу

$$f: L^{n-1} \to H^n, \quad \text{где } H^n = f(L^{n-1})$$

Поскольку из предыдущего представительства L^{n-1} определяется следующий класс ЭИМ либо задач $H^n = f(L^{n-1})$,

$$L^n = g(H^n) = L^*(H^n, Q')$$

, то ряд стационирует (H^n, L^n) и уменьшает значение W до стационарности. Если будем считать, что ряд стационарный с N порядка, получим:

$$W(H^0, Q', L^0) > \ldots > W(H^n, Q', L^n) = W(H^{N+1}, Q', L^{N+1}) = \ldots$$

В случае достижения стационаронсти $N=0$, т.е. с самого начала, можно симметричным образом конструировать разбиение Q'' и ядро L', которое уменьшает критерий W. В этом случае образуется ряд (P'^n, Q'^n, L^n), который выполняет соотношение

$$W(P'^0, Q'^0, L^0) \geq W(P'^1, Q'^1, L^1) \geq \ldots \geq W(P'^n, Q'^n, L^n) \geq \ldots \qquad (2)$$

Алгоритм модифицированного метода конкретизируется двумя промежуточными этапами.

Можно выразить функцию $W(P', Q', L)$ меры сходства:

$$W(P', Q', L) = \sum_{\kappa} \sum_{P_\kappa \in P'_\kappa} \sum_{\ell} \underbrace{\sum_{z_i \in Q'_\ell} |x_\tau^i - a_\kappa^\ell|}_{A}$$

Пусть

$A = \sum_{z_i \in Q'_\ell} |x_\tau^i - a_\kappa^\ell|$ и $q_\ell = card\, Q'_\ell$, т.е. мощность Q'_ℓ

Воспользуемся переменной y_τ^ℓ, где $y_\tau^\ell = \sum_{z_i \in Q'_\ell} x_\tau^i$. Это означает, что вычисляется только сумма $\sum_{z_i \in Q'_\ell} x_\tau^i$ всех единиц в классе. Тогда A выражается в таком виде: $A = |y_\tau^\ell - q_\ell a_\kappa^\ell|$

поскольку $A = \begin{cases} y_\tau^\ell, & \text{если } a_\kappa^\ell = 0; \\ q_\ell - y_\tau^\ell, & \text{если } a_\kappa^\ell = 1 \end{cases}$

В этом случае можно записать:

$$W(P', Q', L) = \sum_{\kappa} \sum_{P_\kappa \in P'_\kappa} \sum_{\ell} |y_\tau^\ell - q_\ell a_\kappa^\ell| \to min \qquad (3)$$

Выразим через y_k^e матрицу (RQ). Все ядра представляются в виде $(q_1 a_k^1, q_2 a_k^2, \dots, q_L a_k^1)$, где $a_k^e \in \{0,1\}$. Необходимо находить для всех классов элементы $(q_1 a_k^1, q_2 a_k^2, \dots, q_L a_k^L)$ которые минимизируют

$$\sum_{P_r \in P_k'} \sum_{e} |y_r^e - q_e a_k^e| = \sum_{e} \sum_{P_r \in P_k'} |y_r^e - q_e a_k^e| . \qquad (4)$$

Это приводит к нахождению элементов $q_e a_k^e$, которые минимизируют $\sum_{P_r \in P_k'} |y_r^e - q_e a_k^e|$.

Если $a_k^e = 0$, то мы получим $\sum_{P_r \in P_k'} y_r^e$.

Если $a_k^e = 1$, тогда

$$\sum'(q_e - y_r^e) = s_k q_e - \sum y_r^e , \qquad s_k = card(P_k') ,$$ т.е. мощность P_k'.

Таким образом, можно определить значения ядра. Если $a_k^e = 0$, то сумма элементов в (P_k', Q_e') ближе к нулю и $a_k^e = 1$, если эта сумма приближается к $s_k q_e$.

В общем виде можно выразить A таким образом

$$A = \sum_{P_i \in P_k'} \sum_{z_j \in Q_e'} |x_r^j - a_k^e| .$$

Пусть

$$y_k^e = \sum_{P_i \in P_k'} \sum_{z_j \in Q_e'} x_r^j ,$$ тогда

$$A = \begin{cases} y_k^e, \text{если } a_k^e = 0 \\ s_k q_e - y_k^e, \text{если } a_k^e = 1 \end{cases} \quad \text{или} \quad A = |y_k^e - s_k q_e a_k^e| \qquad (5)$$

Тогда задача состоит в определении $s_k q_e a_k^e$ для всех классов, которые минимизируют

$$|y_k^e - s_k q_e a_k^e| .$$

В результате для нашего примера получим

$M_2 =$

	Z_1	Z_3	Z_5	Z_8	Z_{10}	Z_2	Z_4	Z_6	Z_7	Z_9
P_3	I			I					I	
P_7				I	I	I				
P_9	I				I		I			
P_1	I	I	I	I	I					
P_3	I	I	I	I	I					
P_4	I	I	I	I	I					
P_2						I	I	I	I	I
P_5						I	I	I	I	I
P_6						I	0	0	0	I
P_{10}						I	I	0	I	0

$$L = {}^{*}\begin{pmatrix} 0 & 0 \\ I & 0 \\ 0 & I \end{pmatrix}$$

Отметим, что альтернативное множество ИМ представляет собой множество ИМ, полученных с помощью некоторых операций, имеющее меньшую мощность по сравнению с первоначальным множеством возможных вариантов ИМ. Для построения альтернативного ИМ, имеется пара смежных классов задач $Q'_{\ell-1}$ и Q'_{ℓ} из матрицы M_2 . Пусть S_ℓ - множество ЭИМ покрывающих информационные потребности класса задач Q'_{ℓ} , элементный состав которого определяется столбцами матрицы M_2 . Каждая пара смежных классов задач порождает некоторые информационные модули. Несколько из них могут быть получены с помощью операции объединения $(\cup)$, пересечения $(\cap)$, разности $(\setminus)$ и дополнения $\bar{S}_\ell$.

Используя эти операции, можно выражать одни множества через другие. Тогда пара смежных классов задач порождает $S_\ell, S_{\ell-1}$ $S_\ell \cap S_{\ell-1}$, $S_{\ell-1} \setminus S_\ell$, $S_\ell \setminus S_{\ell-1}$, $S_\ell \cup S_{\ell-1}$, $\bar{S}_\ell$ и $\bar{S}_{\ell-1}$

Можно выражать $\bar{S}_{\ell-1}$ и $\bar{S}_{\ell}$ через $S_{\ell-1} \setminus S_{\ell}$ и $S_{\ell} \setminus S_{\ell-1}$. Исключаем $S_{\ell-1} \cup S_{\ell}$ в связи с тем, что построение информационных массивов в виде единого набора ЭИМ приводит к практически полному отсутствию гибкости структуры баз данных.

Такая пара смежных классов задач $(Q'_{\ell-1}, Q'_{\ell})$ порождает некоторое множество ИМ, которые могут быть включены в альтернативное множество:

$$\Omega'_{\ell} = \{S_{\ell-1}; S_{\ell}; S_{\ell-1} \cap S_{\ell}; S_{\ell-1} \setminus S_{\ell}, S_{\ell} \setminus S_{\ell-1}\}.$$

Таким образом, рассмотренный подход построения множества альтернативных информационных модулей позволяет значительно уменьшить возможные варианты ИМ с 2^{τ} до $4(L-1)+1$.

Для данного тестового примера $\tau = 10$ и $L = 2$. Тогда уменьшение идет от I024 вариантов ИМ до 5 вариантов. В дальнейшем выбор наилучшего варианта по составу, количеству и качеству ИМ осуществляется из 5 вариантов с помощью метода нахождения локального оптимума, что значительно сокращает затраты на время.

Введем понятие так называемого идеального информационного модуля, который содержит те и только те элементарные информационные модели, которые необходимы для Z_j.

Далее введем величину C_j для оценки ИМ $F_{j'}$, $j' = \overline{1,\tau'}$. Отметим, что $F_{j'}$ определяет j'-й ИМ.

Введем следующие величины:

$\tau'_{jj'} = P'_{jj'} / P_j$, $\tau''_{jj'} = P''_{jj'} / P_j$. P_j – мощность $\mathfrak{D}_j$ (необходимые данные для Z_j), $P_{jj'}$ – мощность множества ЭИМ отсутствующих в $F_{j'}$ и необходимых для Z_j, $P''_{jj'}$ – мощность множества лишних для Z_j ЭИМ, входящих в состав $F_{j'}$. Тогда величина $\tau_{jj'} = \tau'_{jj'} + \tau''_{jj'}$ будет определять оценку отклонения эле-

ментного состава ИМ $F_{j'}$ от идеального ИМ для Z_j . Суммарная оценка F_j' всех задач выражается в виде $C'_{j'} = \tau_{j'} = \sum_{j=1}^{I} \tau_{jj'}$ представляет сумму всех лишних и отсутствующих нужных ЭИМ в j -ом ИМ при решении задач $\{Z_j\}$.

Тогда задача оптимизации информационных массивов, исходя из указанных выше требований к выделяемым из альтернативного множества Ω'_1 , можно выразить требованием минимизации критерия качества $C'_{j'}$.

Итак, математическая модель структуризации информационных массивов можно представить в виде следующей задачи линейного программирования с булевыми переменными

$$\Phi = \sum_{j'=1}^{J'} c_{j'} x_{j'} \qquad (6)$$

при ограничении

$$\sum_{j'=1}^{J'} b_{ij'} x_{j'} = 1, \quad x = \{0,1\}$$

Из множества Ω'_2 выбираем непересекающиеся ИМ, т.е. из

$$\Omega'_2 = \{S_{l=1}; S_2; S_1 \cap S_2; S_1 \setminus S_2; S_2 \setminus S_1\}, \quad l = \overline{1,2}$$

Тогда получим

$$\Omega'' = \{\{S_1, S_2 \setminus S_1\}, \{S_2, S_1 \setminus S_2\}, \{S_1 \setminus S_2, S_1 \cap S_2, S_2 \setminus S_1\}\}$$

Из всех m возможных вариантов ИМ необходимо ввести выбор по k непересекающимся $(k<m)$ ИМ для покрытия информационной потребности. Каждый ИМ имеет информационное качество $C_{j'}$. Пусть множество $J'' \subset \{1,2,3,\dots,m\}$ $|J''| = k$ - его мощность. Тогда задача оптимизации состоит в нахождении J'' так, чтобы

$$\Phi = f(J'') = \sum_{j'=1}^{m} c_{j'} x_{j'} \to min \qquad (7)$$

при ограничениях

$$\sum_{j'=1}^{m} b_{ij'} x_{j'} = 1, \quad x = \{0,1\}$$

Далее, имеется множество J'', $\forall_{j'} \in J''$. Найдем, если су-

ществует, лучшую оценку, то есть, если существует $j'' \in \{1,2,3,\dots,m\} \setminus J''$, такое, что $f(J'' \cup \{j''\} \setminus \{j'\}) < f(J'')$. Если такой случай проверен, то меняем J'' на $J_1 = J'' \cup \{j''\} \setminus \{j'\}$ и процедуру повторяем до тех пор, пока найдем локальный оптимум.

В результате получим $\{S_1, S_2 \setminus S_1\}$. Тогда в состав базы данных входят ИМ $\{S_1, S_2 \setminus S_1\}$. Таким образом, метод нахождения локального оптимума позволяет найти набор ИМ, который формирует базу данных с самой минимальной суммой всех лишних и отсутствующих нужных ЭИМ при решении всех задач $\{Z_j\}$.

Для создания логической структуры баз данных кроме необходимых и достаточных используемых элементов всеми решаемыми задачами, надо знать отношения между ними. Выбор системы отношений осуществляется с помощью модифицированного метода выбора системы отношений между элементами информации.

Пусть каждый пользователь декларирует свою систему отношений U_m^j , где m - номер пользователя и j - номер задачи / $j = \overline{1,r}$, $m = \overline{1,M}$ / в задаваемом им множестве элементов информации. Причем, среди этих отношений могут быть эквивалентные, независимые и пересекающиеся. Поэтому возникает необходимость анализа исходных информационных структур, представляемых пользователем, с целью их совмещения в единой структуре.

Исходные структуры пользователей можно представлять в виде ориентированных графов. Если каждому элементу ИМ S_1 и $S_2 \setminus S_1$ идентифицировать некоторую вершину графа x_i , геометрический образ которой на графе будет задан точкой, то совокупность их определенная на графе с помощью отрезок кривых,

идентифицирует U_j . Они образуют множество отношений $U_j \in U$ Совокупность элементов ИМ S_1 , $S_2 \setminus S_1$ составляют элементы множества X , которое вместе с множеством U представляют собой граф $G(X, U)$, моделирующий структуру рассматриваемой системы.

Из каждой системы отношений U_m^i можно образовать множество декларируемых отношений A_m^i .

$$A_m = \bigcup_{j=1}^{\tau} A_m^j$$

Ориентированные графы $G_m^i(X_m^i, U_m^j)$ соответствуют матрице смежностей $\| \tau_{i'j'}^{A_m^i} \|$

$$\tau_{i'j'}^{A_m^j} = \begin{cases} 1, \text{ если существует ориентированная дуга в множестве отношений, соединяющая } i' \text{ с } j' ; \\ 0 \text{ в противном случае.} \end{cases}$$

Эти графы могут быть связными и содержать одни и те же отношения. Поэтому преобразуем исходные графы пользователей следующим образом.

Приводим все матрицы $\| \tau_{i'j'}^{A_m^i} \|$ к одному размеру и формируем матрицу $\| R_{i'j'}^j \|$ того же размера.

$$R_{i'j'}^j = \sum_{m=1}^{M} \tau_{i'j'}^{A_m^j} \quad , \text{ где } m = \overline{1,M} \; ; \; j = \overline{1,\tau}.$$

Для каждой матрицы $\| R_{i'j'}^j \|$ формируем матрицу $\| \tau_{i'j'}^j \|$ таким образом, чтобы

$$\tau_{i'j'}^j = \begin{cases} 1, \text{ если } R_{i'j'}^j \geq 1 \; , \\ 0 \text{ в противном случае.} \end{cases}$$

Устраняем транзитивные отношения таким образом. Определим все транзитивные отношения: там, где стоят единицы, ставим нули. Так например, имеются три элемента x_1, x_2, x_3. Они формируют такие отношения как $(x_1, x_2), (x_2, x_3), (x_1, x_3)$ Из этих отношений (x_1, x_3) является лишним. Тогда в матрице смежности ставим нуль для (x_1, x_3) и получим матрицу $\| r^{ij}_{i'j'} \|$. Матрица сложности окончательной синтезированной логической структуры БД определяется

$$R_{i'j'} = \sum_{j=1}^{J} r^{ij}_{i'j'}.$$

Формируем для матрицы $R_{i'j'}$ также матрицу $\| R'_{i'j'} \|$ таким образом, чтобы

$$R'_{i'j'} = \begin{cases} 1, \text{ если } R_{i'j'} \geq 1 \\ 0 \text{ в противном случае.} \end{cases},$$

Для всех задач Z_j, $j = \overline{1,J}$, условие независимости описывается следующим образом.

$$card\left((A_s \cup \ldots \cup A_m) \cap (A_\kappa \cup \ldots \cup A_\tau)\right) \leq \Delta_n,$$

где A_1, A_m, ..., A_κ, ..., A_τ - множества декларируемых отношений для всех решаемых задач.

Δ_n - некоторая мера независимости множеств отношений, которая может быть задана, исходя из необходимой точности последующих вычислений.

В результате применения таково модифицированного метода выбора системы отношений получим все отношения, приведенные в табл. I.

Третья глава посвящена выделению классов однотипных элементов, реализации элементарных запросов, разработке метода и алгоритма синтеза оптимальной структуры баз данных с помощью метода исключения бесперспективных вариантов, а также постро-

Таблица I.

Необходимые и достаточные отношения для создания логической структуры БД фрагментов предметных областей о заработной плате рабочих.

№ отношения	Тип отношения	Начало отношения	Конец отношения	Отношения
I	I:I	TN	Fio	(TN,Fio)
2	I:I	TN	K	(TN,K)
3	I:I	TN	C	(TN,C)
4	I:I	IZ	NIZ	(IZ,NIZ)
5	I:I	IZ	KZ	(IZ,KZ)
6	I:I	IZ	NV	(IZ,NV)
7	I:M	SM	U	(SM,U)
8	I:M	PD	BD	(PD,BD)
9	I:I	Fio	BD	(Fio,BD)
I0	I:I	SM	KZ	(SM,KZ)
II	I:I	SM	NV	(SM,NV)
I2	I:I	IZ	KOP	(IZ,KOP)
I3	I:M	PR	K	(PR,K)
I4	I:I	K	WO	(K,WO)
I5	I:I	SI	RC	(SI,RC)
I6	I:I	IZ	KZAT	(IZ,KZAT)
I7	I:I	NIZ	SI	(NIZ,SI)
I8	I:M	C	U	(C,U)
I9	I:I	C	NC	(C,NC)
20	I:I	TN	SM	(TN,SM)
2I	I:I	IZ	KP	(IZ,KP)
22	I:I	KRZ	K	(KRZ,K)
23	I:I	K	PD	(K,PD)

ению оптимальной логической структуры. В главе показано, что целесообразно выделить классы однотипных элементов, таких, что структуры отношений элементов принадлежат к одному классу. Для этого используется метод динамического сгущения.

Имеется множество Ω''', в состав которого входят элементы базы данных. Образуем два непересекающихся подмножества множества Ω''', $\mathcal{D}$ и V, мощностью которых являются n атрибутов и p количественных переменных соответственно. Если предположить, что все элементы находятся в матрице X, тогда элементы множества $\mathcal{D}$ расположены по n строкам и элементы множества V по p столбцам.

$$X=(x_i^j), \quad i\in\mathcal{D} \quad \text{и} \quad j\in V,$$

где x_i^j - значение переменной j для атрибута i.

Предполагается, что каждому атрибуту приписан некоторый вес p_i, сумма которых равна $\sum_{i\in\mathcal{D}} p_i = 1$, где

$$p_i = \frac{1}{n} \quad \forall i \in \mathcal{D}.$$

Предположим также, что каждый атрибут i можно представить как точку. Тогда, для атрибута i можно ассоциировать вектор $x_i = (x_i^1, \ldots, x_i^p) \in R^p$. Множество векторов x_i с весами p_i образуют облако $N(\mathcal{D}) \in R^p$.

Предположим, что матрица X, центрированная по столбцам, то есть, сумма элементов каждого столбца равна нулю. В противном случае необходимо ввести преобразование для выполнения этого условия. Это требование получается в результате того, что центр тяжести облака $N(\mathcal{D})$ находится в начале пространства R^p.

$$\sum_{i\in\mathcal{D}} p_i x_i^j = 0, \quad \forall j \in V.$$

Также для каждой переменной j аналогичным образом ассоциируем ее с вектором $x^j = (x_1^j, \dots, x_n^j)$. Множество x^j с весами q_j образуют облако $N(V) \in R^n$.

Имеется пространство атрибутов с квадратичной метрикой, определенной диагональной матрицей, где коэффициенты важности q_j находятся по диагональной линии. Тогда

$$d^2(x_i, x_{i'}) = \sum_{j \in V} q_j (x_i^j - x_{i'}^j)^2.$$

В этом выражении могут существовать значения переменной j с разными единицами измерения. Чтобы каждая переменная j сыграла одинаковую роль, необходимо ввести нормализацию коэффициентов важности q_j таким образом:

$$q_j = \frac{1}{\sigma_j^2}, \qquad \text{где } \sigma_j^2 = \sum_{i \in D} p_i (x_i^j)^2.$$

σ_j - представляет собой дисперсию переменной j.

Имеется пространство переменных с весами $\mathcal{D}_p$ и метрика. Тогда

$$d^2(x^j, x^{j'}) = \sum_{i \in D} p_i (x_i^j - x_i^{j'})^2.$$

Введем понятие инерции. Рассмотрим облако из n точек $(x_1, x_2, \dots, x_n)$ пространства R^p, причем каждой точке x_k приписан некоторый вес μ_k $(k = 1, 2, \dots, n)$. Инерцией множества $(x_1, x_2, \dots, x_n)$ вокруг точки $a \in R^p$ называется величина

$$I(a) = \sum_{k=1}^{n} \mu_k d^2(x_k, a).$$

Вклад каждой точки в эту инерцию есть $\mu_k d^2(x_k, a)$.

$$I(G)=\sum_{k=1}^{n}\mu_k d^2(x_k,G)=\sum_{j=1}^{k'}\sum_{x_k\in C_j}\mu_k d^2(x_k,G_j)+$$

$$+\sum_{j=1}^{k'}m_j d^2(G_j,G),\quad \text{где}\quad G_j=\sum_{x_k\in C_j}\mu_k x_k\Big/\sum_{x_k\in C_j}\mu_k -$$

центры тяжести k' классов. Эти центры образуют облаки из k' новых обьектов, которым соответствуют веса

$$m_j=\sum_{x_k\in C_j}\mu_k .$$

$\sum_{j=1}^{k'}m_j G_j\Big/\sum_{j=1}^{k'}m_j$ – центр тяжести облака $\{x_k\}$, $k=\overline{1,n}$ с весами $\{\mu_k\}$, $k=\overline{1,n}$ является также центром тяжести облака $\{G_j\}$, имеющегося веса $\{m_j\}$, $j=\overline{1,k'}$.

Инерция облака $I(G)$ вокруг центра тяжести также называется полной инерцией облака и обозначается T. Величина $\sum_{j=1}^{k'}m_j d^2(G_j,G)$ есть инерция облака центров тяжести $\{G_j\}$. Ее называют межклассовой инерцией и обозначают B. Каждое выражение $\sum_{x_k\in C_j}\mu_k d^2(x_k,G_j)$ представляет инерцию точек, составляющих класс C_j, вокруг центра тяжести этого класса. Сумма этих k' инерции называется внутриклассовой инерцией и обозначается W.

Уравнение разложения инерции записывается как

$$T=W+B .$$

Отметим, что метод приводит к минимизации внутриклассовой инерции и наоборот, к максимизации межклассовой инерции.

Определяется инерция облака $N(\mathcal{D})$ по отношению к началу координаты, которая представляет также центр тяжести:

$$I(\mathcal{D},p_i,q_i)=\sum_{i\in\mathcal{D}}p_i d^2(0,x_i)=\sum_{i\in\mathcal{D}}p_i\sum_{j\in V}q_j(x_i^j)^2=$$

$$=\sum_{i\in\mathcal{D}}\sum_{j\in V}p_i q_j(x_i^j)^2 .$$

Далее, из исходной матрицы X сформируем новую матрицу, в которой по строкам расположено разбиение множества $\mathcal{D}$ на непустые подмножества $P'=(P'_1,\dots,P'_{K'})$ на K' классов и по столбцам разбиения множества V на непустые подмножества $Q'=(Q'_1,\dots,Q'_L)$ на L классов. Элементы матрицы $X(P',Q')$ определяются таким образом:

$$x^{\ell}_{\kappa}(P',Q')=\sum_{i\in P'_{\kappa}}\sum_{j\in Q'_{\ell}}p_i q_j x^j_i \Big/ \sum_{i\in P'_{\kappa}}\sum_{j\in Q'_{\ell}} p_i q_j .$$

С новой матрицей $X(P',Q')$ связываются некоторые весы $\gamma_{\kappa'}=\sum_{i\in P'_{\kappa}} p_i$, $\forall \kappa'=1,\dots,K'$ и $\beta_{\ell}=\sum_{j\in Q'_{\ell}} q_j$ $\forall \ell=1,\dots,L$.

Исходя из ранее изложенного, задача нахождения однотипных элементов по классам в системе состоит в нахождении разбиения P' и Q' множеств $\mathcal{D}$ и V соответственно такого, чтобы новая матрица $X(P',Q')$, имеющая весы $(\gamma_{\kappa'})$ и (β_{ℓ}), потеряла меньше информации, то есть

$$I\big(X(P',Q'),(\gamma_{\kappa'}),(\beta_{\ell})\big)=\sum_{\kappa'}^{K'}\sum_{\ell}^{L}\gamma_{\kappa'}\beta_{\ell}(x^{\ell}_{\kappa'})^2\to max.$$

Таблица 2

Расчет количественных атрибутов

i \ j	KZ	KP	NV	RC	KOLD	KZAT	SI
I	2	3	4	5	6	7	8
TN	I5	8	5	0	0	0	0
F1O	I5	0	0	0	0	0	0
K	0	0	0	0	0	0	0
PR	0	0	0	0	0	0	0
KRZ	0	0	0	0	0	0	0
C	I500	I000	40	0	0	0	0

Окончание табл. 2

I	2	3	4	5	6	7	8
U	I50	92	40	0	0	0	0
SM	I50	92	5	0	0	0	0
WO	0	0	0	0	0	0	0
NC	I500	I00	40	0	0	0	0
IZ	I50000	100000	90	3200	30	1500	15
NIZ	I50000	I00000	30	3200	30	I500	I5
KOP	0	0	0	0	0	0	0
PD	0	0	0	5	0	4	0
BD	0	0	0	45	0	4I	0
Сумма	303330	20I292	250	6450	60	3045	30
$\bar{X}$	20222	I34I9,4I	I6,67	430	4	203	2

Таблица 3.

Результативная таблица.

i \ j	KZAT	KOLD	SI	RC	NV	KP	KZ
I	2	3	4	5	6	7	8
TN	0,40	0,39	0,39	0,40	0,46	0,39	0,40
FIO	0,40	0,39	0,39	0,40	0,66	0,40	0,40
SM	0,40	0,39	0,39	0,40	0,46	0,39	0,39
C	0,40	0,39	0,39	0,40	-0,92	0,37	0,37
U	0,40	0,39	0,39	0,40	-0,92	0,39	0,39
NC	0,40	0,39	0,39	0,40	-0,92	0 39	0,37
KRZ	0,40	0,39	0,39	0,40	0,66	0,40	0,40

Окончание табл. 3.

I	2	3	4	5	6	7	8
IZ	-2,55	-2,55	-2,55	-2,55	-2,89	-2,55	-2,55
NIZ	-2,55	-2,55	-2,55	-2,55	-0,52	-2,55	-2,55
KOP	0,40	0.39	0,39	0,40	0,66	0,40	0,40
PR	0,40	0,39	0,39	0,40	0,66	0,40	0,40
K	0,40	0,39	0,39	0,40	0,66	0,40	0,40
WO	0,40	0,39	0,39	0,40	0,66	0,40	0,40
BD	0,32	0,39	0,39	0,35	0,66	0,40	0,40

$I = 0{,}0000305$; $q_1 = 0{,}0000196$; $q_2 = 0{,}0000294$;

$q_3 = 0{,}393580$; $q_4 = 0{,}0009204$; $q_5 = 0{,}0980581$;

$q_6 = 0{,}0019653$; $q_7 = 0{,}1961161$.

Тогда все классы однотипных элементов описываются таким образом:

$$P_1' = \{TN, FIO, SM\},\ P_2' = \{C, U, NC, KRZ\},\ P_3' = \{IZ, NIZ, KOP\}.$$

$$P_4' = \{PR, K, WO, PD, BD\};\ Q_1' = \{KZAT, KOLD, SI\}.$$

$$Q_2' = \{RC, NC, KP, KZ\}.$$

Реализация элементарных запросов осуществляется следующим образом.

Пусть имеются некоторые множества отношений $\{A_j\}$ $j = \overline{1,T}$, необходимые для решения задач $\{Z_j\}$.

Тогда проблема состоит в формировании связного подграфа для каждого множества A_j , чтобы для любого его элемента существовали пути /дуги/ от него к остальным элементам этого множества.

В этой главе разработан метод и алгоритм синтеза оптимальной логической структуры базы данных.

Логическая структура представляет собой множество элементов и отношений между ними. Тогда можно представить ее в виде ориентированного графа $G=(X,U)$, где $X=\{x_i\}$, $i=\overline{1,n}$ множество вершин, $U=\{u_j\}$, $j=\overline{1,m}$ - множество дуг, соединяющих между собой пары вершины. Каждая вершина x_i графа G однозначно сопоставляется информационными элементами определенного типа, каждая дуга u_j - связь между двумя какими-либо информационными элементами.

Итак, модель синтеза оптимальной логической структуры БД описывается следующими соотношениями.

Определить множество дуг $\{x_{IJ}^{ijk}=1\}$, входящих в оптимальное решение при минимальной годовой стоимости поиска, хранения и передачи информации

$$W=\sum_{k=1}^{K}\sum_{J=1}^{P}\sum_{I=1}^{P}\sum_{i=1}^{n}\sum_{j=1}^{n} a f_{ijk} x_{IJ}^{ijk} \to \min ,$$

где a - средняя стоимость обработки одного указателя связи /включает стоимость преобразования ключа в адрес, стоимость поиска, чтения соответствующего элемента информации и передачи информации; f_{ijk} - частота связи (i,j) в K-ом запросе; x_{IJ}^{ijk} - бинарная переменная, используемая запросами.

$$x_{IJ}^{ijk}=\begin{cases} 1, \text{ если дуга } (i,j)\in I\times J \text{ в } K\text{-ом запросе;} \\ 0, \text{ в противном случае.} \end{cases}$$

Принимая во внимание следующие условия

$$\lim_{n\to+\infty}\left(S+\frac{\tau}{2^n}\right)\to S ,$$

где S - минимальная стоимость хранения логической структуры базы данных;

τ - радиус окружности;

n - искомый параметр.

$l \leq \tau$ и $i < j$ поскольку граф ориентированный и без петель.

Приведенная модель синтеза оптимальной логической структуры БД легко сводится к задаче целочисленного линейного программирования.

Пронумерованы классы и их элементы, полученные в результате действия метода динамического сгущения

$I = \{1, 2, 3\}$, $2 = \{4,5,6,7\}$, $3 = \{8,9,10\}$,

$4 = \{11,12,13,14,15\}$, $5 = \{16,17,18\}$, $6 = \{19,20,21,22\}$,

где TN = I; FIO = 2; SM = 3; C = 4; V = 5; NC = 6;

KRZ = 7; IZ = 8; NIZ = 9; KOP = I0; K = II; PR = I2;

WO = I3; PD = I4; BD = I5; KZAT = I6; KOLD = I7;

SI = I8; RC = I9; NV = 20; PK = KP = 21; KZ = 22.

Таблица 4.

Частота решения задач.

Z_j	f_j	Z_j	f_j
I	I80	6	30
2	I50	7	20
3	I00	8	I0
4	80	9	I0
5	50	I0	5

В результате действия алгоритма синтеза оптимальной логической структуры базы данных получается для нашего примера логическая структура со стоимостью хранения y = 960 руб. и суммарной стоимостью поиска и передачи информации W = 10760 руб.

Элементарные запросы реализуются таким образом.

- { /8,9/, /8,20/, /8,22/, /3,22/, /1,3/, /1,4/, /1,5/, /1,2/, /1,11/, /1,12/} ;
- {/3,20/, /3,22/, /3,5/, /1,3/, /1,2/, /2,15/, /1,15/, /14,15/};
- { /8,9/, /8,17/, /8,22/, /3,22/, /1,3/, /8,10/, /1,2/, /10,17/};
- { /8,9/, /9,18/, /18,19/, /8,20/, /3,20/, /1,3/, /1,11/, /7,11/, /11,14/, /11,12/, /11,13/ } ;
- { /8,9/, /8,16/, /9,18/, /8,19/, /8,18/};
- { /8,9/, /8,20/, /8,21/, /3,21/, /1,3/, /1,2/, /4,5/, /4,6/, /1,4/, /1,5/ } ;
- { /8,9/, /4,5/, /8,16/, /8,19/, /8,18/, /3,20/, /3,5/};
- { /8,9/, /8,22/, /1,3/, /1,2/, /1,11/, /1,12/, /1,13/ } ;
- { /8,9/, /8,16/, /8,20/, /1,3/, /1,2/};
- {/1,4/, /4,6/} .

В этой главе также построена оптимальная логическая структура. Она представлена в виде гиперграфов на рис. 1.

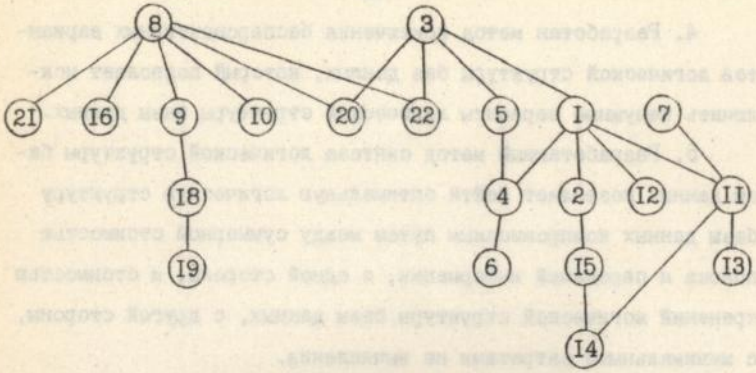

Рис. I. Логическая структура в виде гиперграфа.

Основные выводы.

I. Анализ оптимизации распределения информационных потоков с применением метода сортировки, показывает, что этот метод не позволяет устранить дублирование информации, что приводит к неоптимальной логической структуре баз данных. Для устранения этого недостатка применен модифицированный метод динамического сгущения.

2. Исследование метода динамического сгущения показывает, что он не только не позволяет устранить дублирование информации в покрытии, но и не может найти предельное количество попарных пересечений классов, нужных для нахождения разбиения.

Модифицированный метод динамического сгущения позволяет устранить эти недостатки.

3. Разработан также метод исключения нулевых элементов, который позволяет снизить стоимость хранения и вычисления.

4. Разработан метод исключения бесперспективных вариантов логической структуры баз данных, который позволяет исключить ненужные варианты логической структуры базы данных.

5. Разработанный метод синтеза логической структуры базы данных позволяет найти оптимальную логическую структуру базы данных компромиссным путем между суммарной стоимостью поиска и передачей информации, с одной стороны, и стоимостью хранений логической структуры базы данных, с другой стороны, с минимальными затратами на вычисления.

6. Разработанный модифицированный метод выбора системы отношения учитывает наличие всех решаемых задач, а также позволяет найти логическую структуру базы данных в виде гиперграфа. Такая новая модель базы данных использует ассоциативный метод доступа.

Основные опубликованные работы, отражающие содержание диссертации

I. Волколупова Р.Т., Амана Н.И. Идентификация подструктур баз данных. Сборник научных трудов "Экономические проблемы работы предприятий в новых условиях хозяйствования". - К.: УМКВО, 1988. - С.52-61.

2. Волколупова Р.Т., Амана Н.И. Модифицированный метод динамического сгущения. Материалы докладов учредительной конференции международной ассоциации по нетрадиционным методам оптимизации. - Красноярск: КИКТ, 1992.

3. Волколупова Р.Т., Бузько Я.П., Амана Н.И. Метод нахождения локального оптимума. Сборник научных трудов "Использование математических методов и информационных технологий в технических и экономических системах". - К.: ИК АН Украины, 1992.

Подп. к печ. 20.05.93. Формат 60×84 1/16.
Бумага тип. № 3 . Способ печати офсетный. Услови. печ. л. 1,63
Услови. кр.-отт. 1,86 . Уч.-изд. л. 1,0 .
Тираж 100 . Зак. № 4534 . Бесплатно.

Фирма «ВИПОЛ»
252151, г. Киев, ул. Волынская, 60.