Некоторый подход к классификации исходных данных, циркулирующих в информационных системах

В. В. Митропов, кандидат военных наук, доцент
А. В. Кравцов, г. Москва
Н. В. Кравцов, г. Москва

На настоящем этапе развития цифровых систем связи Вооруженных сил РФ в руководящих документах неоднократно подчеркивался факт необходимости перехода к использованию распределенных баз данных, позволяющих повысить качество и скорость получения и обработки необходимой информации для должностных лиц органов управления (ОУ) различного ранга.

Однако уже на самых раних этапах проектирования распределенных систем необходимо предусмотреть возможность сокращения объёмов так называемой первичной информации об объектах предметной области, а также реализовать процесс объединения первичных структурообразующих элементов в более крупные образования или классы: алгоритмы, записи, файлы, массивы данных, позволяющих исключить дублирование отдельных фрагментов информации и существенно снизить избыточность её потоков. В противном случае, из-за некорректной структуризации данных будут значительно возрастать объёмы запоминающих устройств (ЗУ) на узлах сети, снижаться пропускная способность информационной системы (ИС), возрастать время доступа к ресурсам распределённой базы данных (РБД) и т.п. Причем, ошибки и просчёты, допущенные на начальных этапах построения РБД, будут множиться и достигать гигантских размеров на заключительных этапах. Поэтому, от того, насколько удачно будет сформирована начальная структура данных, в значительной степени будет зависеть эффективность функционирования всей системы РБД в целом. Имеющийся опыт создания РБД показывает, что только по этой причине ежегодно за рубежом до 30% проектов ИС оказываются неудачными. Экономические потери достигают 10 миллиардов долларов.

Под структуризацией информации будем понимать введение определенных соглашений о способах и формах упорядочения данных. Каждая представляемая информацией сущность имеет ряд характерных для неё свойств (признаков, параметров, характеристик и т.п.). В качестве значений таких признаков могут быть последовательности символов (букв, цифр, различных знаков или обозначений) или конкретные параметры объектов (например, характеристики средств связи), будет называть их структурообразующими элементами, иерархически упорядоченными во времени и пространстве, и функционально связанными между собой. По существу они являются минимальной по составу информационной совокупности распределенных систем, которые сохраняют информативность и поэтому достаточны для образования самостоятельных массивов данных, которые в дальнейшем могут существовать даже изолированно от информационной системы, имея свою форму и свои алгоритмы. В этом отношении структуры данных можно рассматривать как некоторую совокупность структурообразующих элементов.

Базу данных можно представить в виде совокупности экземпляров записей различного типа, содержащую ссылки между записями, представленные в виде набора (рис. 1.)

Таким образом, ведущая роль в общем процессе формирования массивов информации базы данных принадлежит первичным структурообразующим элемента данных (СЭД), а также способам организации процесса структуризации промежуточных типов структур БД, которые не могут выполняться произвольным образом, а только по определенным правилам. Таким образом, целью структуризации

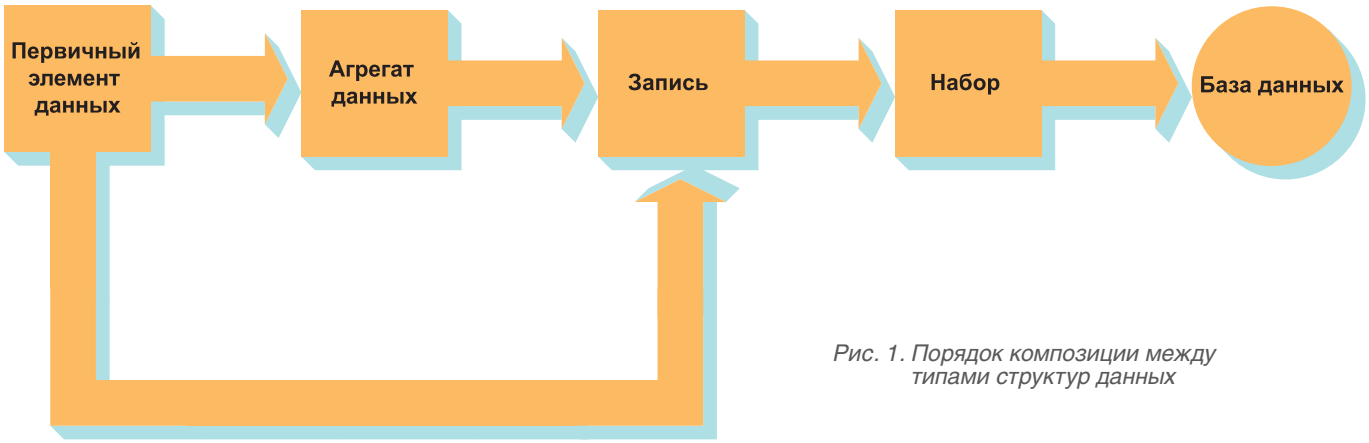

Рис. 1. Порядок композиции между типами структур данных

является такое разбиение (кластеризация) объектов на некоторые подмножества (классы), соответствующее определенным типам данных, которое в конечном итоге позволит построить наиболее рациональную форму организации базы данных. Ввиду исключительной сложности задачи в настоящее время отсутствуют теоретические разработки в законченном виде и тем более модели или алгоритмы процессов кластеризации распределенных БД. Более того, существует мнение, что задачи структуризации данных скорее всего следует относить к области искусства, чем технологическим проблемам.

Одним из таких способов решения задачи структуризации данных РБД является так называемый ***венгерский метод*** (комбинаторной оптимизации ограниченных множеств), который можно рассматривать как частный случай транспортных задач, но с более простыми алгоритмами решения, чем алгоритмы решения транспортных задач за счет использования специальных эквивалентных итерационных преобразований исходных матриц эффективности. Достоинством венгерского метода является также возможность оценки близости результата каждой из итераций к оптимальному варианту решения. Это позволяет контролировать процесс вычислений и прекращать его при достижении заданной точности результата. Такое свойство особенно важно для задач большой размерности, каковыми и являются РБД.

Основная идея рассматриваемого метода заключается в том, что на первом этапе строится начальный план в виде некоторой исходной матрицы распределения множества объектов в соответствии с условием задачи. Затем осуществляется переход к новому плану, близкому к оптимальному. Последовательное применение этого приёма за конечное число итераций в конце концов приводит к искомому решению. При этом осуществляется процедура преобразования исходной матрицы мощности межреберных связей двудольного графа $\mathbf{A} = \|\mathbf{a_{ij}}\|$ к эквивалентным $\mathbf{A'} = \|\mathbf{a'_{ij}}\|$ путем последовательной целенаправленной комбинации между собой соответствующих столбцов и строк исходной и эквивалентной матрицы связей (Таблица 1.).

В общем виде алгоритм выделения из исходного множества максимально связанных между собой подмножеств основан на вычислении и последовательном анализе коэффициента связности диагональных подматриц по результатам соответствующих расчетов, суммарных значений мощностей межреберных связей для соответствующих строк и столбцов недиагональных подматриц. В зависимости от результатов анализа промежуточных значений указанных величин осуществляется последовательная замена соответствующих строк или столбцов диагональных подматриц матрицы **A** до тех пор, пока не будет достигнуто максимальное значение целевой функции $\boldsymbol{\varphi}$. Основным признаком принадлежности объектов данных к определенной группе являются суммарные значения r_{qj} и r^n_{qj}, которые определяют общее количество связей объектов ***i*** и ***j*** с другими объектами внутри и вне **q**-го класса.

Таблица 1.

Матрица $A=\|a_{ij}\|$ мощности межреберных связей

j \ i	1	2	3	…	j	…	m
1	r^z_{qi} / r_{qj}		r^z_{1i}		r^z_{2i}	…		r^z_{ki}
2		$\mathbf{A_{11}}$		$\mathbf{A_{12}}$		…	$\mathbf{A_{1k}}$	
3	r_{1j}		$\mathbf{R_{11}}$		$\mathbf{R_{12}}$	…		$\mathbf{R_{1k}}$
.		$\mathbf{A_{21}}$		$\mathbf{A_{22}}$		…	$\mathbf{A_{2k}}$	
.	r_{2j}		$\mathbf{R_{21}}$		$\mathbf{R_{22}}$	…		$\mathbf{R_{2k}}$
i		…	…	…	…	…	…	…
.		$\mathbf{A_{k1}}$		$\mathbf{A_{k2}}$		…	$\mathbf{A_{kk}}$	
.	r_{kj}		$\mathbf{R_{k1}}$		$\mathbf{R_{k2}}$	…		$\mathbf{R_{kk}}$
n								

Работа алгоритма начинается с приведения исходной матрицы мощностей межреберных связей к виду, представленному на рис 2.4, просмотру для подматриц $\mathbf{A_{qq}}$ суммарных значений r_{qj} и выбору среди них максимальных значений $r^{\min}_{qj}$ (операторы 2÷4). Затем для недиагональных подматриц $\mathbf{A_{qp}}$ ($p\neq q$, $p=1,2,\ldots,k$) из множества r_{qj} выбирается $r^{\max}_{qj}$ (оператор 5). Диагональная подматрица, содержащая столбец с $r^{\max}_{qj}$, обозначается A_{pp}. Производится замена местами столбцов с $r^{\min}_{qj}$ и $r^{\max}_{qj}$ (оператор 6). В результате суммы элементов соответствующих подматриц формируются следующие приращения:

$$\Delta R_{qq} = r^{\max}_{qj} - r^{\min}_{qj}; \quad (1)$$

$$\Delta R_{pp} = r^{\min}_{pj} - r^{\max}_{pj}; \quad (2)$$

$$\Delta \sum_{p=1}^{k} R_{qp} = 0; \quad (3)$$

$$\Delta \sum_{q=1}^{k} P_{pq} = 0; \quad (4)$$

$$\Delta \sum R^n_{pq} = \sum_{q=1}^{k} r^{(n)\max}_{qj} - \sum_{q=1}^{k} r^{(n)\min}_{qj}; \quad (5)$$

$$\Delta \sum R^n_{qp} = \sum_{q=1}^{k} r^{(n)\min}_{qj} - \sum_{q=1}^{k} r^{(n)\max}_{qj} = \Delta \sum_{p=1}^{k} R^n_{pq} \quad (6)$$

После замены столбцов начинается просмотр для подматрицы $\mathbf{A_{qq}}$ значения r^n_{qj} и выбора среди них минимального $r^{n(\min)}_{qj}$ (оператор 8). Затем для подматриц $\mathbf{A_{qp}}$ ($p\neq q$, $p=1,2,\ldots,k$) из r^n_{qj} выбирается $r^{n(\max)}_{qj}$ (операторы 9 и 10). В результате будем иметь:

$$\varphi'^{q} = \frac{R_{qq} + \Delta R_{qq}}{\sum_{p=1}^{k} R_{qp} + \sum_{p=1}^{k} R_{pq}} \cdot \frac{R_{qq} + \Delta R_{qq}}{\sum_{p=1}^{k} R^n_{pq} + \sum_{p=1}^{k} R_{pq}}; \quad (13)$$

$$\psi = \varphi^q \cdot \varphi^p = \varphi^m_q \cdot \varphi^n_q \cdot \varphi^m_p \cdot \varphi^n_p; \quad (14)$$

$$\psi' = \varphi'^q \cdot \varphi'^p = \varphi'^m_q \cdot \varphi'^n_q \cdot \varphi'^m_p \cdot \varphi'^n_p. \quad (15)$$

При этом необходимо учитывать следующие соотношения:

$$\Delta R_{qq} = r_{qi}^{(n)\max} - r_{qi}^{(n)\min}; \quad (7)$$

$$\Delta R_{pp} = r_{pi}^{(n)\min} - r_{pi}^{(n)\max}; \quad (8)$$

$$\Delta \sum_{p=1}^{k} R_{pq}^{n} = 0; \quad (9)$$

$$\Delta \sum_{p=1}^{k} R_{qp}^{n} = 0; \quad (10)$$

$$\Delta \sum_{p=1}^{k} R_{qp} = \sum_{q=1}^{k} r_{qi}^{(n)\max} - \sum_{q=1}^{k} r_{qi}^{(n)\min}; \quad (11)$$

$$\Delta \sum_{q=1}^{k} R_{pq} = \sum_{q=1}^{k} r_{qi}^{(n)\min} - \sum_{q=1}^{k} r_{qi}^{(n)\max}. \quad (12)$$

При этом следует отметить, что разработанный алгоритм позволяет увеличивать внутригрупповую связность элементов данных только в том случае, если получено оптимальное разбиение матрицы **A** на **k** подматриц заданной размерности. Поскольку процедура перебора всех возможных вариантов замены столбцов и строк конечна, то сходимость алгоритма не вызывает сомнений.

Литература

1. Королев М.А. и др. Информационные системы и структуры данных. Под ред. проф.М.А. Королева.М.,
2. Статистика, 1977.
3. Бауэр Ф.Л., Гооз Г. Информатика. Пер. с нем.М.. Мир, 1976.
4. Бератрис А.Т. Структура данных. Пер.с англ..М., Статистика, 1974.
5. Куцык Б.С. Структура данных и управление. М., Наука,1975.
6. Папернов А.А., Подымов В.Я. Методы упорядочения информации в цифровых системах, М.,Наука,1973.
7. Скелетов С.Н., Волков Б..Г. Хранение и поиск данных в ЭВМ. М., Советское радио,1977.
8. Кожурин Ф.Д., Ярмош Н.А. Структурная обработка больших информационных массивов. Минск, Наука и техника,1975.
9. Колемаев В.А. Математическая экономика/Учебник для вузов.- М.: ЮНИТИ, 1998.
10. Ашманов С.А. Введение в математическую экономику/-М.: Наука, 1984.