

А.Э. МАНУКЯН

ВЫБОР ФОРМЫ ОРГАНИЗАЦИИ СВЯЗЕЙ ТИПА "ОДИН КО МНОГИМ" МЕЖДУ ОРГАНИЗАЦИОННЫМИ ЕДИНИЦАМИ БАЗЫ ЗНАНИЙ

Անարկալայան աշխարհի ինֆորմացիոն օբյեկտների միջև առկա բարդ հարաբերությունների արտապատկերման անհրաժեշտ փուլ է «մեկը շատերի հետ» կապերի ֆիզիկական կազմակերպումը: Առաջարկվում է կապերի կազմակերպման լավագույնում չէլ կախված տվյալների հոսքի բնույթից: Առաջված կանոններն առաջարկվում է օգտագործել գիտելիքների բազաների կառավարման համակարգերի ծրագրային միջոցների մշակման ժամանակ:

Организация связей типа "один ко многим" между организационными единицами базы знаний является необходимым этапом процесса отображения сложных отношений между ИО предметной области. Предлагается оптимальный способ организации связей типа "один ко многим" в зависимости от характера потока данных. Полученные результаты рекомендуется использовать при создании программных средств систем управления базами знаний.

Библиогр.: 1 назв.

The organization of links "one to many" between organizational knowledge base units is the necessary stage for the process of displaying complex relations between informational objects. Optimal organization methods for choosing links depending on the input data stream character are proposed. The results obtained are to be used for the knowledge base software system.

Ref. 1

Организация связей типа "один ко многим" между организационными единицами базы знаний (БЗ) является необходимым этапом процесса отображения сложных отношений между информационными объектами (ИО) предметной области. Применяются следующие два способа организации этой связи:

1. Связь между корневой и подчиненными ей организационными единицами выражается наличием определенной закономерности между координатами расположения этих единиц в памяти ЭВМ.

2. Связь организуется посредством списка указателей хранящегося у корневой организационной единицы.

Для автоматического выбора оптимального способа организации связей необходимо средствами системы управления базами знаний (СУБЗ) провести статистический учет о потоке поступающих данных и запросов. Это позволяет, в свою очередь, провести оценку ожидаемых потерь машинных ресурсов и автоматическое подключение средств организации связей, обеспечивающих минимум этих затрат. Введем зависимость оценок ожидаемых потерь от параметров потока данных и запросов.

Пусть ИО некоторого уровня представления БЗ отображаются набором $(m-1)$ организационных единиц нижестоящего уровня. В общем случае количество свойств и объем памяти, необходимой для хранения каждой из m организационных единиц, являются случайными величинами. Первый способ организации связи при распределении объема памяти (Q) , необходимой для хранения ИО, обычно вынуждает фиксировать количество свойств (m_0) и объем памяти (χ_0) , отводимый для представления каждого i -го свойства.

Не теряя общности дальнейших рассуждений, их можно принять равными, т.е. $Q = m_0 \chi_0$. Тогда при ведении БЗ возникают ситуации, когда размещение всех поступающих на вход системы данных об ИО становится невозможным из-за превышения объема памяти (χ_0) , требуемой для хранения данных хотя бы одного свойства с зарезервированным значением χ этого объекта. Вероятность появления такой ситуации определяется выражением

$$P_n = 1 - [1 - P \times P'(\chi_0 - \chi)]^n, \quad (1)$$

где P — вероятность появления данного свойства у ИО.

Известно, что функция распределения вероятностей объемов ИО хорошо аппроксимируется функцией гамма-распределения [1]. Поэтому для вероятности $P(\chi_0 > \chi)$ можно написать, что

$$P(\chi_0 > \chi) = \int_0^{\chi_0} \frac{r^\alpha}{\Gamma(\alpha)} \chi_0^{\alpha-1} e^{-r} dr = \frac{\Gamma(\alpha, r\chi_0)}{\Gamma(\alpha)}, \quad (2)$$

где $r = M(\chi_0) / D(\chi_0)$, $\alpha = M^2(\chi_0) / D(\chi_0)$; $D(\chi_0)$ — дисперсия случайной величины χ .

При втором способе организации связей координаты расположения организационных единиц в памяти заранее не фиксируются, они формируются и помещаются в свободных фрагментах памяти по мере поступления данных о свойствах ИО. Поэтому объем памяти, необходимой для их хранения, является случайной величиной с математическим ожиданием

$$M(Q) = M(m)(S + M(X)),$$

где S — затраты памяти на организацию одного указателя. Тогда условие эффективности применения второго способа организации связей можно представить в виде $M(m)(S + M(X)) < m_0 \chi_0$.

Преобразуем последнее неравенство к виду $A < Y$, где $Y = r\chi_0$; $A = M(m)(rS + \alpha) / m_0$. Заметим, что $\Gamma(\alpha, Y)$ является монотонно убывающей функцией от Y , так как в (2) подынтегральное выражение всегда положительно в области $0 < Y < \infty$. С учетом вышеизложенного условие $A < Y$ можно представить в виде $\Gamma(\alpha, A) > \Gamma(\alpha, Y)$, а на основе (1) и (2) условие эффективности применения второго способа организации связей принимает вид

$$\Gamma(\alpha, A) > [1 - (1 - P_n)^n] \Gamma(\alpha) / P_n. \quad (3)$$

Использование условия (3) в средствах адаптации требует статистического учета об ИО и автоматического определения

величин $P, P(x), M(x), D(x)$, если автоматическое определение $P(x)$ затруднено, то в (3) можно принять $\alpha = 2$. Значение $\alpha = 2$ для ИЭ статистически оправдано. При таком значении величины α условие (3) преобразуется в виду

$$e^{-\lambda} (1 - P_0) > [1 - (1 - P_0)^m] / P, \quad (4)$$

где

$$\lambda = 2 \frac{M(m)}{n_0} \left(\frac{S}{M(x)} + 1 \right).$$

Использование (4) в реальных условиях адаптации требует значительно меньших затрат вычислительных ресурсов. При организации БЗ к каждой из связей типа "один ко многим" могут приписываться идентификатор и различные количественные характеристики. Если для представления связей используются указатели, то значения характеристик хранятся совместно с указателями в специальных таблицах при сортированных организационных единицах. Значения характеристик могут совпадать во многих строках этой таблицы. В зависимости от количества таких строк и распределения частот их активизации, наиболее эффективным может стать использование дополнительных таблиц для групп указателей с однозначными характеристиками. В основной таблице для этих групп указателей предусматривается лишь одна строка, содержащая указание характеристик и указатель на дополнительную таблицу. Для выбора оптимального состава и структуры таблиц указателей в программных средствах адаптации возникает необходимость оценки среднего времени перехода по связям как при однотабличной так и при двухтабличной организации указателей. Выведем выражения этих оценок.

Пусть хранящаяся на диске таблица указателей из k строк с длиной строки в l байтов содержит группу из n строк, в каждой из которых l_0 байтов совпадает по содержанию и представляет собой характеристики связей. Тогда условие эффективности применения дополнительной таблицы указателей, с точки зрения минимума количества операций ввода-вывода, можно записать в виде

$$1 + \frac{k}{l_0} > \left(1 - \frac{n}{k}\right) \left(1 + \frac{(k-n)l}{l_0}\right) + \frac{n}{k} \left(2 + \frac{k-n+1}{l_0} + \frac{n(l-l_0)}{l_0}\right), \quad (5)$$

где l_0 — объем блока данных, обмениваемых между ОЗУ и диском, при операциях ввода-вывода. При выводе последнего условия была принята гипотеза о равновероятности активизации связей и случайным расположении таблиц указателей на диске.

Необходимость реорганизации уже существующих в БЗ таблиц указателей на принятие решения перехода от однотабличной системы организации указателей к двухтабличной рекомендуется выявить путем проверки условия (5).

ЛИТЕРАТУРА

1. Белоногов Г.Г., Новоселов А.П. Автоматизация процессов накопления, поиска и обработки информации. — М.: Наука, 1979. — 256 с.

ГИУД

08.02.1997