УЧЕНЫЕ ЗАПИСКИ ВЫСШИХ УЧЕБНЫХ ЗАВЕДЕНИЙ ЛИТОВСКОЙ ССР
ЭКОНОМИКА, XIV, ВЫП. 2, 1977

ПОИСК В МАССИВЕ С ВЕРОЯТНОСТНОЙ ЗАВИСИМОСТЬЮ МЕЖДУ АДРЕСОМ ЗАПИСИ И ЗНАЧЕНИЕМ КЛЮЧА СОРТИРОВКИ

А. БАРТКУС

Скорость поиска в массиве, организованном последовательным способом, зависит от количества обращений к устройствам внешней памяти для поиска одиночных записей. Если между значением ключевых реквизитов и номером записи (условным адресом, определяемым расстоянием от начала массива до определенной записи) можно найти аналитическую зависимость, то адрес искомой записи может быть определен однозначно и запись найдена за одно обращение к устройствам внешней памяти. Однако для подавляющего большинства массивов, функционирующих в автоматизированных системах управления предприятиями, такой зависимости определить нельзя.

В связи с этим при использовании запоминающих устройств с непосредственным доступом (магнитные диски, магнитные барабаны) может быть использован метод дихотомии. Среднее число обращений для этого метода поиска может быть определено следующим образом. Пусть имеется массив, состоящий из N зон (запись считается найденной, если в ОЗУ введена зона массива, содержащая искомую запись). Вероятность нахождения записи за одно обращение равна $\frac{1}{N}$, а вероятность необходимости больше чем одного обращения $\frac{N-1}{N}$. Вероятность нахождения искомой записи за два обращения равна $\frac{1}{\frac{N-1}{2}} \cdot \frac{N-1}{N}$ и т. д. Формула для подсчета математического ожидания числа обращений K_g выглядит следующим образом:

$$K_g = \frac{1}{N} + \frac{N-1}{N}\left(2\frac{1}{\frac{N-1}{2}} + \quad + \frac{a_1-1}{a_1}\left((i+1)\frac{1}{a_{i+1}} + \quad\right)\cdots\right),$$

где

$$\frac{1}{a_{i+1}} = \begin{cases} \frac{2}{a_1-1}, & \text{если } \frac{2}{a_1-1} < 1 \\ 1, & \text{если } \frac{2}{a_1-1} \geqslant 1 \end{cases}$$

где $a_1 > 1$.

Если $N = 1000$, то рассчитанное по этой формуле математическое ожидание числа обращений равно 9,03.

Обследование конкретных массивов показывает, что между номером записи и значениями ключевых реквизитов может быть обнаружена корреляционная зависимость. Действительно, поскольку записи массива упорядочены по возрастанию значения ключей сортировки, то чем дальше запись расположена от начала массива, тем бóльшие значения имеют

ее ключи. Методом наименьших квадратов можно определить сглаженную аналитическую зависимость между значением ключа к и адресом а записи $a=f(к)$ и параметры отключения фактического адреса от расчетного как случайной величины ζ.

Пусть ζ имеет нормальный закон распределения и $M(\zeta)=0$.

При наличии таких данных можно аналитически определить границы первоначальной области поиска записи со значением ключа $к_и$ (алгоритм I) следующим образом:

$$a_н^1=\max\{f(к_и)-3\sigma,\ a_1\},$$
$$a_в^1=\min\{f(к_и)+3\sigma,\ a_N\},$$

где $a_н^1$ — первоначальная нижняя граница области поиска,
$a_в^1$ — первоначальная верхняя граница области поиска,
a_1 — адрес первой зоны массива,
a_N — адрес последней зоны массива,

$\min\{x_1, x_2\}$ означает, что выбирается наименьшее значение из величин, заключенных в фигурные скобки.

Ясно, что при $\sigma<\frac{N}{3}$ первоначальная область поиска всегда меньше a_N-a_1, что дает возможность найти искомую запись за меньшее число обращений, чем методом дихотомии. После определения $a_н^1$ и $a_в^1$ дальнейший поиск производится методом дихотомии.

Рассмотрим следующий алгоритм поиска (алгоритм II). Определяется расчетный адрес искомой записи $a^р$. Соответствующее ему фактическое значение ключа $к_ф$ сравнивается с ключом искомой записи $к_и$. Если $к_и<к_ф$, то

$$a_н^1=a^р-3\sigma,\ a_в^1=a^р;$$

если $к_и>к_ф$, то

$$a_н^1=a^р,\ a_в^1=a^р+3\sigma.$$

Хотя первоначальная область в алгоритме II в два раза уже, чем в алгоритме I, но в алгоритме II для определения этой области необходимо одно обращение к внешним ЗУ, так что общее количество обращений в обоих алгоритмах одинаковое.

Рассмотрим алгоритм, основанный на предварительной сегментации первоначальной области поиска (алгоритм III). Пусть количество обращений по алгоритму II равняется z. Разобьем теперь первоначальную область на четыре сегмента. Ясно, что количество обращений для поиска в одном сегменте равно $z-2$, но появляются дополнительные обращения, необходимые для определения, находится ли искомая запись в данном сегменте. Для первого сегмента от границы, определяемой величиной $a^р$, необходимо одно дополнительное обращение, для второго — два и т. д. В общем случае математическое ожидание количества обращений определяется по формуле

$$z_г=\sum_{i=1}^{n}p_i(z-\log_2 n+i-1),$$

где n — количество сегментов,
p_i — вероятность нахождения искомой записи в сегменте i. Учитывая нормальный закон распределения ζ, математическое ожидание количества обращений в случае четырех сегментов будет

$$z_{0,75}=(z-1)\cdot 0{,}7112+z\cdot 0{,}2549+(z+1)\cdot 0{,}0324+(z+2)\cdot 0{,}0015=$$
$$=z-0{,}6758.$$

Зависимость математического ожидания количества обращений от величины сегмента, выраженного в долях σ, показана на рис. 4, из которого видно, что оптимальный сегмент имеет величину 0,4σ. Алгоритм III при оптимальной сегментации первоначальной области требует в среднем на 0,95 обращений меньше, чем алгоритм II.

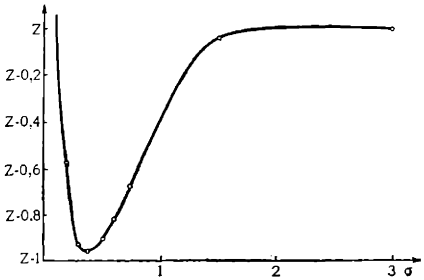

Рис. 4. Зависимость количества обращений от величины шага

Естественно, что оптимальные сегменты могут быть определены также и для других законов распределения.

Зависимость количества обращений по алгоритму III от значения определена следующим путем. Была составлена математическая модель массива, состоящего из 50 000 записей, распределенных по 1000 зон. Величина σ менялась от 25 до 1000 записей. Для каждого значения моделировалось 200 независимых поисков, причем $к_и$ и $к_ф$ определялись на основании величин, полученных из генератора случайных чисел. Результаты испытаний представлены на рис. 5.

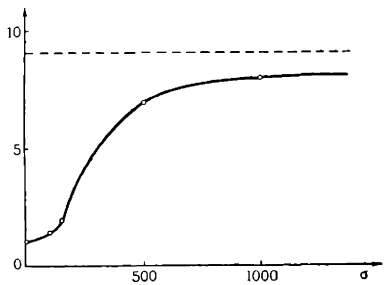

Рис. 5. Зависимость числа шага от среднеквадратного отклонения

Анализ кривой подтверждает, что алгоритм III при любых значениях σ эффективнее алгоритма I, причем эффективность увеличивается при уменьшении σ. Во всех случаях алгоритм III эффективнее алгоритма I, для которого количество обращений не зависит от значения σ (пунктирная линия на рис. 5).

Алгоритм III перекрывает весь диапазон от метода дихотомии до аналитического метода определения адреса искомой фразы. Следовательно, предлагаемый метод является наиболее универсальным среди методов, не требующих дополнительной памяти.

Вильнюсский государственный
университет им. В. Капсукаса
Кафедра экономической
информации

Редколлегии вручено
в декабре 1975 г.

PAIEŠKA MASYVE SU TIKIMYBINE PRIKLAUSOMYBE TARP ĮRAŠO ADRESO IR RŪŠIAVIMO RAKTO REIKŠMĖS

A. BARTKUS

Reziumė

Informaciniams masyvams, naudojamiems AVS, galima nustatyti statistinę priklausomybę tarp įrašo adreso ir rūšiavimo rakto reikšmės, kurią galima panaudoti paieškai paspartinti.

Tam tikslui apskaičiuojama išlyginta funkcinė priklausomybė tarp įrašo adreso ir rūšiavimo rakto reikšmės ir atsitiktinio dydžio — faktinio įrašo adreso nukrypimo nuo apskaičiuojamojo — parametrai.

Jeigu šis atsitiktinis dydis turi žinomą pasiskirstymo dėsnį, tai galima analitiniu būdu apskaičiuoti pradinį paieškos intervalą, kuris visais atvejais ne didesnis už masyvą. Taigi sumažėja kreipimųsi į išorinę atmintį skaičius, palyginus su dichotomijos metodu.

Jeigu nukrypimas turi normalinį pasiskirstymo dėsnį, tai kreipimųsi skaičių galima dar sumažinti, pasirinkus optimalų pradinio intervalo segmentą. Įrodoma, kad šis segmentas yra lygus 0,46.