УДК 004.62

Т. И. Москвичева[*]

АЛГОРИТМ ДОСТУПА К ОБЪЕКТАМ В ИНФОРМАЦИОННОМ МАССИВЕ

Развитие информационных технологий приводит к тому, что государственные организации и коммерческие предприятия вынуждены обрабатывать большие объемы данных. Современная теория информационного поиска предоставляет различные подходы для построения алгоритмов поиска, добавления и удаления данных в информационных массивах, основанные на известных моделях: лексикографических, древовидных, реляционных и других [1, 2]. Вместе с тем, даже для таких широко известных программных средств, как СУБД компании *Oracle* или сообщества *OpenSource*, отмечается значительное время поиска, несоответствующее решаемым задачам и необходимость мощной и, как следствие, дорогой элементной базы. Причиной тому является несовершенство моделей хранения данных для эффективного решения задач поиска. Поэтому актуальной является задача разработки математической модели, позволяющей сократить время поиска элементов, а также операций добавления и удаления.

Одной из эффективных, с точки зрения доступа к данным, моделей является адресный массив, рассмотренный, например, в работах Э. Э. Гасанова, В. Б. Кудрявцева и Ю. П. Луговской [1, 2]. В нем поиск элемента осуществляется за константное время путем вычисления адреса объекта на основе его значения. Вместе с тем, в реальных условиях, характеризующихся большими объемами данных в информационных массивах такая модель практически не применима, что ограничивает ее использование. Поэтому на практике широкое распространение получили модели с высоким (логарифмическим) временем доступа к элементам, но не требующие выделения больших объемов дополнительной памяти, а также гибридные модели [2].

Построим математическую модель адресного массива.

Пусть $X = \{x_1, x_2, x_3, \ldots, x_n\}$ – множество хранимых элементов. Даже если элементы множества X изначально не числовой природы, факт хранения их на электронных носителях предполагает, что они представлены в виде конечной битовой последовательности, каждую из которых можно рассматривать как число. Например, битовую последовательность, соответствующую элементам X, можно рассматривать как бинарную запись целого числа и трактовать, таким образом, храни-

[*] Работа выполнена под руководством канд. техн. наук, доцента ФГБОУ ВПО «ТГТУ» А. В. Яковлева.

мые элементы как целые числа. Другой подход – трактовать битовые последовательности элементов X, как мантиссы некоторых дробных чисел со степенями 1, тогда хранимые элементы представляют собой числа из интервала (0; 1). Заметим, что оба примера трактовки хранимых элементов рассматривают их как рациональные числа (Q). Тогда без ограничения общности предположим:

$$X \subset [x_1; x_2] \subset Q,\ x_1 < x_2 < \ldots < x_n. \quad (1)$$

Строгость данного неравенства обуславливается тем, что хранение идентичных объектов представляет собой хранение пары: объекта и числа его появлений.

Для построения адресного массива вычислим его шаг

$$\delta = \min_{i=\overline{2, n}} (x_i - x_{i-1}). \quad (2)$$

Определим размер адресного массива как

$$m = \frac{x_n - x_1}{\delta}. \quad (3)$$

Разобьем отрезок $[x_1; x_2]$ на m отрезков вида $[y_{i-1}; y_i]$, $i = \overline{1, m}$ следующим образом

$$y_0 = x_1,\ y_i = y_{i-1} + \delta,\ \forall i = \overline{2, m}. \quad (4)$$

Из (4) очевидно, что $y_i = y_0 + i\delta,\ \forall i = \overline{2, m}$ и соответственно $y_m = y_0 + m\delta$ или в силу (3) и (4) $y_m = x_1 + \frac{x_n - x_1}{\delta}\delta = x_n$.

Заметим, что каждый хранимый элемент при таком разбиении $[x_1; x_2]$ попадает в одно из множеств: $\{y_0\}$, $(y_0, y_1]$, $(y_1, y_2]$, …, $(y_{m-1}, y_m]$. Причем в каждое из этих множеств попадает только один хранимый элемент. Если предположить, что в некоторый полуинтервал $(y_{i-1}, y_i]$, $\forall i = \overline{2, m}$ попадет два элемента X, то расстояние между ними должно быть строго меньше δ – длины отрезка $[y_{i-1}, y_i]$. Однако, согласно (2) и (1), δ является минимальным расстоянием между элементами X. Для множества, состоящего из одного элемента $\{y_0\}$, предложенное утверждение тем более очевидно.

Пусть хранимые элементы некоторым образом располагаются в памяти. Создадим массив указателей из m элементов: $p_1, p_2, \ldots, p_m$. Идея адресного массива заключается в том, чтобы сопоставить каждому полуинтервалу $(y_{i-1}, y_i]$ элемент p_i. А именно, для $\forall x \in X \left| x \in (y_{y-1}; y_i]\ i = \overline{1, m}\right.$, p_i присвоим адрес x. Остальным указате-

лям присвоим нулевые значения, сигнализирующие о том, что они не указывают ни на какой элемент.

Поиск идентичных объектов состоит в поиске хранимого элемента, идентичного запросу, т.е. в качестве запроса без ограничения общности выступает некоторый элемент $x \in Q$. Задача же поиска идентичных объектов – определить истинность выражения $x \in X$. Идея поиска на основе адресного массива состоит в том, чтобы проверить, принадлежит ли x $(x_1; x_n]$ и если да, то определить, какому из полуинтервалов $(y_0, y_1]$, $(y_1, y_2]$, …, $(y_{m-1}, y_m]$ принадлежит x. Вычислив интервал, найдем и соответствующий ему p_i, в котором либо лежит адрес элемента, идентичного x, и тогда выражение $x \in X$ истинно, либо выражение $x \in X$ – ложно.

Пусть $x \in (y_{i-1}, y_i]$ $i = \overline{1, m}$, тогда

$$y_{i-1} < x \le y_i \tag{5}$$

или в силу (4)

$$y_0 + (i-1)\delta < x \le y_0 + i\delta.$$

Тогда, в силу (3) и (4) получим

$$i = \left\lceil m \frac{x - x_1}{x_n - x_1} \right\rceil, \tag{6}$$

где $\lceil\ \rceil$ – округление в большую сторону.

Формула (6) позволяет вычислить индекс массива p_i, в котором содержится адрес элемента, идентичного запросу с константной, т.е. не зависящей от числа элементов, асимптотической сложностью.

Основной проблемой такого подхода является быстрый рост объемов лишней памяти с увеличением числа хранимых данных. А именно, памяти, в которой хранятся нулевые значения указателей. Обозначим избыточную память N, которая определяется по формуле

$$N = m - n. \tag{7}$$

Для хранения n элементов необходимо n ячеек памяти, а используется m ячеек ($m > n$).

Рост требуемой памяти в модели адресного массива определяется неравномерностью распределения хранимых данных в информационном массиве. Если распределение данных в памяти осуществлялось равномерно, т.е. с интервалом, определяемым (2),

$$\delta = \frac{x_n - x_1}{n},$$

то выражение (3) приняло бы вид

$$m = n\frac{x_n - x_1}{x_n - x_1} \quad \text{или} \quad m = n,$$

и избыточная память в силу (7) фактически отсутствовала бы. Однако в случае реальных данных предложенный подход требует огромных объемов дополнительной памяти [1, 2]. Таким образом, из-за неравномерности распределения хранимых элементов такие математические модели на сегодняшний момент не применяются в системах хранения данных.

В данной работе предлагается подход к хранению данных на основе формирования их отображения на адресное пространство, адаптирующегося к неравномерности их распределения. Таким образом, сложность поиска адреса элемента будет определяться сложностью построения отображения, а выделение дополнительной памяти сведено к минимуму.

Хранение данных в абсолютном большинстве случаев подразумевает возможность добавления новых данных. Модель адресного массива обладает замечательным свойством, которое заключается в том, что если добавляемый элемент не меняет δ, он добавляется за константное время, равное времени поиска. Действительно, добавление элемента в этом случае сводится к присваиванию адреса нового элемента соответствующему p_i. Чтобы сохранить это свойство в предлагаемом методе, а именно добавление элемента без необходимости менять что-либо, в том числе менять отображение, достаточно, чтобы f могло быть использовано на произвольном конечном подмножестве $[x_1, x_n]$. Для сохранения этого свойства наложим на него условие биективности на всех конечных подмножествах $[x_1, x_n]$ [2].

Решение задачи в общем виде крайне сложно из-за бесконечного числа элементов в множестве F. Однако ее решение даже на некотором подмножестве F может доказать эффективность предложенного в данной работе подхода.

Список литературы

1. *Гасанов, Э. Э.* Теория хранения и поиска информации / Э. Э. Гасанов, В. Б. Кудрявцев. – Москва : ФИЗМАТЛИТ, 2002. – 288 с.

2. *Гасанов, Э. Э.* Константный в худшем случае алгоритм поиска идентичных объектов / Э. Э. Гасанов, Ю. П. Луговская // Дискретная математика. МГУ. – 1999. – Т. 11, № 4. – С. 139 – 144.

Кафедра «Информационные системы и защита информации»
ФГБОУ ВПО «ТГТУ»