СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

Р11 - 10519

Д.Д.Арнаудов

МОДЕЛЬ ИПС ДЛЯ ОЦЕНКИ ПАРАМЕТРОВ ИНФОРМАЦИОННЫХ СТРУКТУР

1977

P11 - 10519

Д.Д.Арнаудов

МОДЕЛЬ ИПС ДЛЯ ОЦЕНКИ ПАРАМЕТРОВ ИНФОРМАЦИОННЫХ СТРУКТУР

Арнаудов Д.Д

Р11 - 10519

Модель ИПС для оценки параметров информационных структур

Описана модель ИПС, предназначенная для оценки параметров инверсной, ассоциативно-адресной и частично инвертированной структуры. Параметры, относящиеся к запросам пользователей и банку данных, имеют статистический характер. Характеристики вычислительной машины относятся к параметрам устройств прямого доступа, типа магнитных дисков. Приводятся выражения для оценки среднего времени доступа в рассматриваемых структурах при работе информационно-поисковой системы, а также график функции, характеризующей запросы пользователей и частоту распределения дескрипторов в фонде ИНИС.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Сообщение Объединенного института ядерных исследований. Дубна 1977

Arnaudov D.D.

P11 - 10519

A Model for an Evaluation of Simulated Parameters of Information Structures

A model for an evaluation of simulated parameters of the multilist, inverted and partially inverted structure is described. Parameters, referring to the queries and the data-bank, have a statistical character. Parameters of the external devices refer to direct access devices, i.e. magnetic discs.

Expressions are derived for the evoluation of the average access time in these structures, during the work of an information retrieval system. The distributions of the queries and the frequencies of the descriptors in INIS data-bank are investigated.

The investigation has been performed at the Laboratory of Computing Techniques and Automation, JINR.

Communication of the Joint Institute for Nuclear Research. Dubna 1977

В информационно-поисковой системе имеется множество подсистем. Идеальная модель такой системы должна предоставлять возможность измерения и корректировки всех параметров данных подсистем. Однако этот вопрос в глобальном смысле пока еще не решен. Имеющиеся данные [I-4] показывают, что большинство работ в области моделирования поисковых систем направлены на моделирование процедур поиска и на оценку стоимости работы этих процедур. Некоторые авторы [8], однако, описывают общую структурную модель поисковой системы, используя характеристики "реальных" массивов, а другие предлагают технику моделирования входного потока в информационно-поисковой системе [6].

Цель данной работы – построить модель, которая дала бы возможность исследовать основные параметры структуры информационно-поисковых систем (ИПС) на примере трех широко распространенных структур: инверсной, ассоциативно-адресной, и частично инвертированной [7-9]. Как инверсная, так и ассоциативно-адресная организации массива представляют различные варианты структуры массива, реализация которых приводит к различным особенностям программирования. Однако обе организации могут представлять одну и ту же информационную структуру – иерархическую или ассоциативную. Они сходны в том смысле, что используют одни и те же средства логики разбиения, а именно адреса связи; в обоих случаях для разбиения массива необходимо указывать одно и то же количество адресов связи. Единственное отличие состоит в местонахождении этих адресов. Комбинируя инверсную и ассоциативно-адресную организации, можно получать различные варианты частично инвертированной структуры.

Основные параметры модели можно разбить на три группы: параметры, относящиеся к запросам пользователей, к банку данных (информационному массиву) и к характеристикам конкретной вычислительной машины, на которой реализуется система. Все параметры сведены в таблицу I, где массив ОмПОД является основным информационным массивом, а массивы ОМД и МЗД представляют управляющую часть основного информационного массива /IO/. Необходимо заметить, что большинство параметров, относящихся к запросам пользователей и к банку данных, имеют статистический характер. Кроме того, параметры вычислительной машины относятся к характеристикам устройств прямого доступа.

Все параметры модели хорошо описаны в работе /IO/. Здесь мы приведем полученные выражения для вычисления среднего времени доступа к информации, расположенной на магнитном диске, для инверсной, ассоциативно-адресной и частично инвертированной структур. Вывод этих формул подробно рассмотрен в работах /9, IO/.

$$T_{инв.} = N_p \frac{f(j)}{A} T_{z\,инв.} \qquad (1)$$

$$T_{acc.} = L \cdot T_{z\,acc.} \qquad (2)$$

$$T_{двухур.} = C_{к\,омд} T_{z_1} + a_{ompод} C_{к\,ompод} T_{z_2} . \qquad (3)$$

В формулах /1 ÷ 3/ $T_{z\,инв}$, $T_{z\,acc}$, T_{z_1}, T_{z_2} обозначают время, необходимое для чтения информации с соответствующих фрагментов дискового устройства. Если, например, величина узла меньше физической единицы записи, то время трансмиссии достаточно мало и тогда

$$T_{z\,acc.} = t_п + \frac{1}{2} t_o .$$

Таблица I

к массиву:

- N — число дескрипторов словаря системы
- Z — величина зоны в физических записях
- V — количество различных дескрипторов, используемых в поисковых образах документов
- N_z — количество документов в массиве
- $N_к$ — количество дескрипторов, индексирующих документ (среднее)
- $L = \frac{N_z N_к}{V}$ — количество документов на дескриптор (средняя длина списка)
- Q — количество символов на запись
- R_c — среднее количество записей в зоне
- $C_к = \begin{Bmatrix} C_{к\,омд} \\ C_{к\,мзд} \\ C_{к\,ompод} \end{Bmatrix}$ — среднее количество зон на дескриптор (соответственно в массивах ОМД, МЗД, ОМПОД)
- $гз = \begin{Bmatrix} гз_{омд} \\ гз_{мзд} \\ гз_{ompод} \end{Bmatrix}$ — число зон массивов ОМД, МЗД и ОМПОД

к запросу:

- $гз_з = \begin{Bmatrix} гз_{з\,омд} \\ гз_{з\,мзд} \\ гз_{з\,ompод} \end{Bmatrix}$ — среднее количество зон запроса (соответственно в массивах ОМД, МЗД, ОМПОД)
- K — коэффициент, указывающий на наличие более чем одного дескриптора в зоне I части ОМД
- N_t — количество дескрипторов в запросе
- N_p — количество положительных дескрипторов в запросе
- L_s — длина самого короткого списка дескрипторов запроса
- ρ — отношение среднего числа ответов на запрос к L
- $\rho = \frac{гз_{з\,ompод}}{C_{к\,омд}}$ — отношение среднего числа общих зон запроса в ОМПОД к $C_{к\,омд}$ (для двухуровневой схемы)
- $\delta = \frac{гз_{з\,мзд}}{C_{к\,омд}}$ — отношение среднего числа общих зон запроса в МЗД к $C_{к\,омд}$ (для трехуровневой схемы)
- $a = \begin{Bmatrix} a_{омд} \\ a_{мзд} \\ a_{ompод} \end{Bmatrix}$ — отношение среднего количества общих зон запроса к $C_к$, где $a_{омд} = \frac{гз_{з\,омд}}{C_{к\,омд}}$; $a_{мзд} = \frac{гз_{з\,мзд}}{C_{к\,мзд}}$; $a_{ompод} = \frac{гз_{з\,ompод}}{C_{к\,ompод}}$

к ЭВМ:

- A — количество адресов записей массива на физическую запись
- $t_s = \begin{Bmatrix} t_п \\ t_o \\ t_{тр} \end{Bmatrix}$ — время поиска цилиндра, время оборота, время трансмиссии

В качестве частично инвертированной структуры рассматривается двухуровневая информационная структура /7,10/.

Как уже заметили, формулы 1 ÷ 3 и таблица 1 содержат параметры, которые имеют ярко выраженный статистический характер. Так, например, N_p -указывает на среднее число положительных дескрипторов запроса, L - является средней длиной дескрипторного списка запросов, $f(j)$ - указывает на число случаев, когда j - й дескриптор используется для индексирования в поисковых образах документов. В этом случае самый общий вид рассматриваемых формул имеет вероятностный характер, и они могут быть выражены следующим образом (в параметрах частично инвертированной структуры $C_{команд}, C_{конц}, a_{запрос}$ уже учтен вероятностный характер рассматриваемых распределений /10/) :

$$T_{инв.} = T_{z\ инв.} \sum_{j=1}^{N} \frac{f(j)}{A} \cdot P(j) \quad , \qquad (4)$$

где A -число записей в одной физической записи.

$$T_{acc.} = T_{z\ acc.} \sum_{j=1}^{N} f(j)\, P(j) \quad . \qquad (5)$$

В приведенных формулах принимается, что в рассматриваемых структурах имеется $f(j)$ документов в списке, соответствующем j - му дескриптору; следовательно, j - й список имеет длину $f(j)$. Тогда время, необходимое для обработки, например, всех узлов в списке ассоциативно-адресной структуры, равно $f(j) \cdot T_{z\ acc.}$. Это есть и среднее, и максимальное время, необходимое для обработки всех узлов. Кроме того, имеется в виду, что каждая запись охватывает не больше чем одну физическую единицу записи.

Из предложенных формул видно, что выбор соответствующих дескрипторов запроса осуществляется на основе распределения функции $P(j)$,

а распределение дескрипторов информационного массива описывается функцией $f(j)$. Следовательно, для моделирования входного потока запросов пользователей и формирования банка данных необходимо иметь математическое выражение для функций $P(j)$ и $f(j)$.

Раскроем более подробно физическую сущность этих функций. Для этого необходимо напомнить, что под банком данных ИПС, или, иначе говоря, под понятием "информационного массива" системы, будем подразумевать набор записей поисковых образов документов, каждый из которых характеризуется некоторым количеством дескрипторов.

Пусть общее количество дескрипторов в дескрипторном словаре равно N. Код каждого дескриптора характеризуется некоторым числовым индексом j, где $1 \le j \le N$.

Пусть $f(j)$ - это общее число случаев, когда j - й дескриптор используется для индексирования в поисковых образах документов (иначе говоря, это частота использования данного дескриптора в различных документах массива).

Пусть S - это общее число случаев, в которых все дескрипторы встречаются в различных документах массива. Следовательно,

$$S = \sum_{j=1}^{N} f(j) \, .$$

Данная формула характеризует суммарное использование дескрипторов в различных документах массива и является, как уже заметили, характеристикой самого массива.

В качестве функции, характеризующей запросы пользователей, можно выбрать $P(j)$ - вероятность того, что выбранный в запросе дескриптор есть j - й дескриптор.

Функции $f(j)$ и $P(j)$ могут быть неизвестными или очень трудно определяемыми. Вид этих функций может зависеть как от конкретной информации в массиве, так и от метода индексирования этой информации и т.п.

Поэтому особый интерес представляет расчет отдельных параметров модели при различных распределениях функций $f(j)$ и $P(j)$, и их сравнение с параметрами модели при использовании реального информационного массива и реальных запросов пользователей. Тогда можно показать, какое распределение этих функций является самым близким к реальному, и использовать это при моделировании указанных структур на вычислительной машине.

Теоретически показано, что возможные распределения $f(j)$ и $P(j)$ лежат в области, определяемой равномерным распределением и распределением по закону Ципфа /9/. Тогда возможны четыре случая изменения функций $f(j)$ и $P(j)$.

I. Равномерное распределение предполагает, что функции равномерно распределены и постоянны для всех j, т.е.

$$f(j)=\frac{S}{N} \; ; \quad P(j)=\frac{1}{N} \; .$$

2. Функции имеют распределение по закону Ципфа, т.е.

$$f(j)=\frac{S}{j(\ln N+\gamma)} \; ; \quad P(j)=\frac{1}{j(\ln N+\gamma)} \; .$$

В этом случае используется формула Ципфа с поправкой *Lowe* /9,II/. Здесь $\gamma = 0{,}5772$ и употребляется в качестве константы Эйлера.

3. Функция $f(j)$ распределена по закону Ципфа, а $P(j)$ имеет равномерное распределение, т.е.

$$f(j)=\frac{S}{j(\ln N+\gamma)} \; ; \quad P(j)=\frac{1}{N} \; .$$

4. Функция $P(j)$ распределена по закону Ципфа, а $f(j)$ имеет равномерное распределение, т.е.

$$f(j)=\frac{S}{N} \; ; \quad P(j)=\frac{1}{j(\ln N+\gamma)} \; .$$

Основные параметры модели для массива из 3200 и I0000 документов показаны в таблицах 2,3. В таблице 4 показаны параметры модели реального массива ИНИОа /I3/ для 3200 и I0000 документов. Эксперименты проводились для двух значений величины зоны информационного массива: Z = 7 физических единиц (*PRU*) и Z = I40 физических единиц.

В качестве запросов использовались 300 различных дескрипторов, генерированных в соответствии с различными случаями (I ÷ 4), а также пакет реальных запросов пользователей (300 запросов). Запросы были представлены в конъюнктивной форме. В каждом запросе имелось в среднем по три дескриптора.

На рис.I показаны графики функции $f(j)$ для реального массива из 3200 документов (штрихпунктирная линия), для массива из I0000 документов (пунктирная линия) и сплошной линией изображен график функции Ципфа. На рис.2 показаны график распределения дескрипторов в реальном запросе пользователей (пунктирная линия) и график функции Ципфа.

Необходимо заметить, что для моделирования различных распределений (I ÷ 4) использовался хорошо известный в теории вероятности метод воспроизведения событий с заданной вероятностью. Этот метод состоит в следующем. Разобьем интервал /0,I/ на N подинтервалов $D_1, D_2 - D_N$. Правый конец δ_i интервала D_i вычисляется по формуле:

$$\delta_i=\sum_{j=1}^{i} P(j),$$

$$P(j)=\frac{1}{j(\ln N+\gamma)} \; .$$

Для каждого документа моделируемого массива генерируются N_K случайных точек / где N_K – среднее количество дескрипторов на документ/ с равномерным распределением в интервале /0,I/. Если случайная точка

Таблица 2

$$f(j)=\frac{S}{j(\ln N+\gamma)}$$

	N_z	V	L	N_z	V	L
	3200	6552	4,3	10000	9459	9,5
	$S=N_z N_к=28800$			$S=N_z N_к=90000$		
Z (PRU)	7	140		7	140	
$C_к$ омпод	3,14	1,97		6,81	4,26	
$C_к$ омд	1,52	0,47		4,6	1,9	
$r3_{омпод}$	134	7		417	21	
$r3_{омд}$	54	1,2		246	11	
число всех заголовков ОМД	20582	12912		64420	40373	
$r3_{з\,омпод}$ $P(j)=\frac{1}{j(\ln N+\gamma)}$	1	0,7		1,4	1,05	
$a_{омпод}$ при $P(j)=\frac{1}{j(\ln N+\gamma)}$	0,3	0,3		0,2	0,25	
$r3_{з\,омпод}$ $P(j)=\frac{1}{N}$	0,3	0,6		0,68	0,45	
$a_{омпод}$ при $P(j)=\frac{1}{N}$	0,09	0,3		0,1	0,1	

Таблица 3

$$f(j)=\frac{1}{N}$$

	N_z	V	L	N_z	V	L
	3200	9419	3	10000	9997	9
	$S=N_z N_к=28800$			$S=N_z N_к=90000$		
Z (PRU)	7	140		7	140	
$C_к$ омпод	3	2,54		8,9	7,4	
$C_к$ омд	2,8	2		8,8	6	
$r3_{омпод}$	134	7		417	21	
$r3_{омд}$	86	4		353	15	
число всех заголовков ОМД	28481	23972		89022	74510	
$r3_{з\,омпод}$ $P(j)=\frac{1}{j(\ln N+\gamma)}$	1,1	0,9		2,7	2,4	
$a_{омпод}$ при $P(j)=\frac{1}{j(\ln N+\gamma)}$	0,3	0,36		0,3	0,32	
$r3_{з\,омпод}$ $P(j)=\frac{1}{N}$	1,24	1,1		3,8	3,6	
$a_{омпод}$ при $P(j)=\frac{1}{N}$	0,4	0,44		0,42	0,48	

Таблица 4

Реальный массив ИНИСА

	N_z	V	L	N_z	V	L
	3200	2220	9	10000	7780	11
	$S=N_z N_K=28800$			$S=N_z N_K=89851$		
Z (PRU)	7	140		7	140	
Ск омпод	6,33	2,6		8,03	3,9	
Ск омд	4,46	0,6		5,58	1,37	
$z3_{омпод}$	132	6,5		410	21	
z3 омд	78	1,2		245	6	
число всех заголовков ОМД	20000	12100		62540	31076	
$z3_в$ омпод	0,6	1,2		1,5	0,9	
α омпод	0,1	0,2		0,18	0,23	

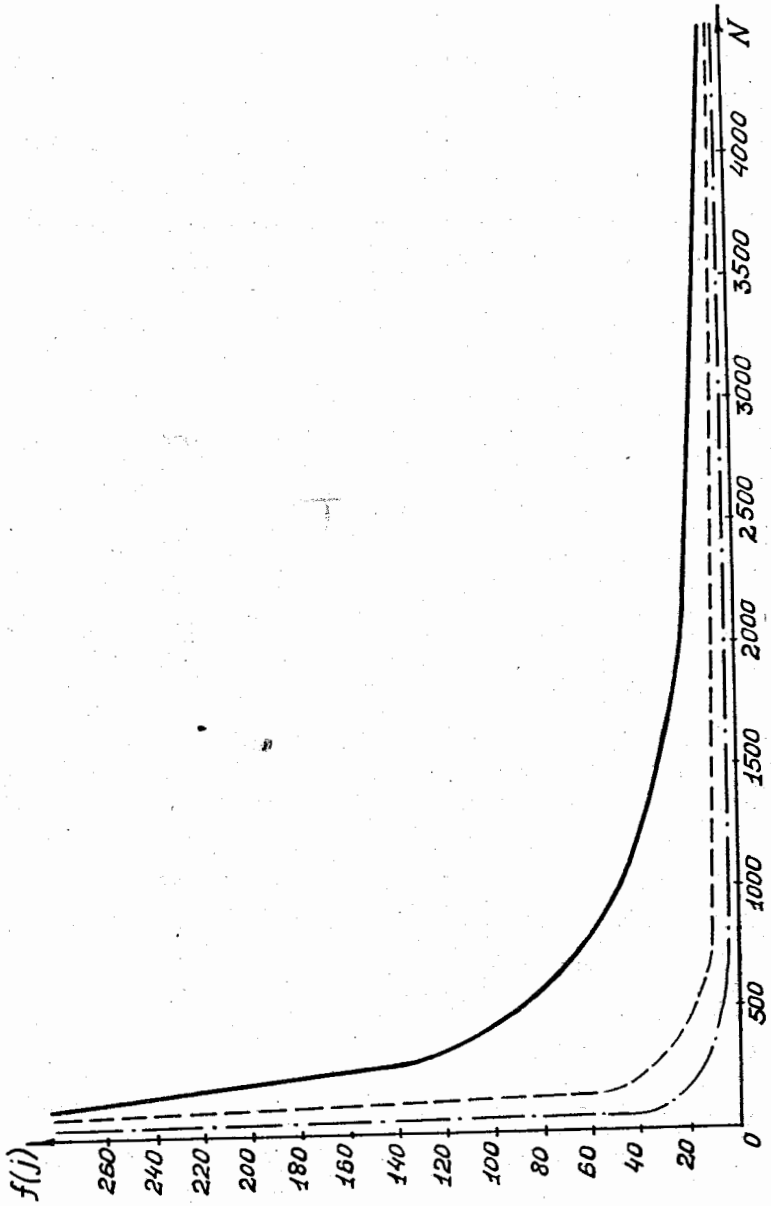

Рис.1

попадает в i- й подинтервал, считаем, что в данном документе встречается i- й дескриптор. Поскольку длина i-го подинтервала равняется $\delta_i - \delta_{i-1}$, то данная процедура приводит /при S обращений к генератору случайных чисел/ к появлению i-го дескриптора в среднем $P(j)\cdot S$ раз, т.е. в данном случае по закону Ципфа частота использования дескриптора получается равной $\frac{S}{i(\ln N+\gamma)}$.

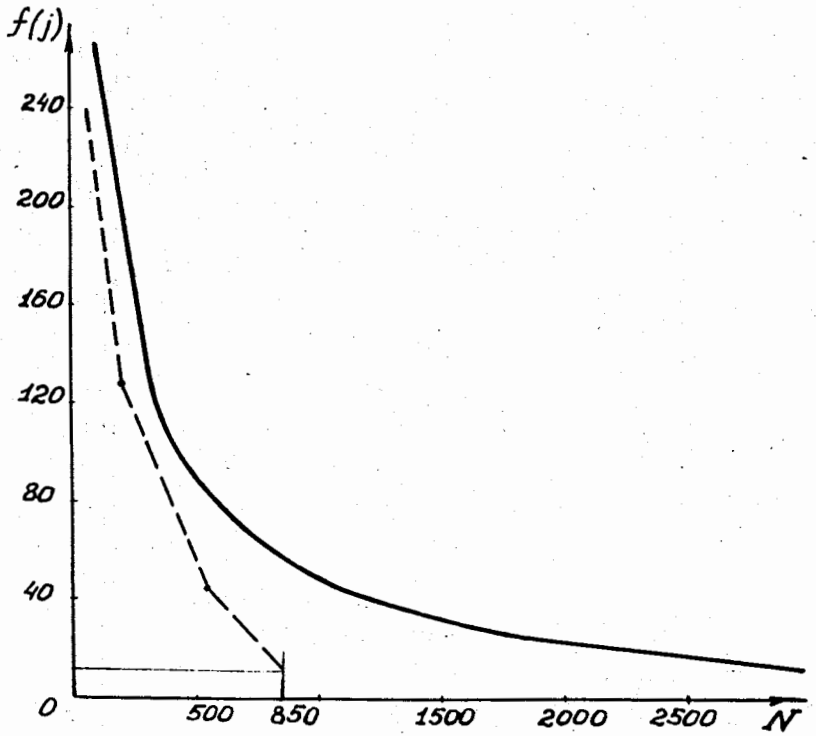

Рис.2

Расчет среднего времени доступа для инверсной, ассоциативно-адресной и частично инвертированной структур на основе банка данных ИНИСа (из 10000 документов) на ЭВМ CDC 65000.

Значения основных параметров модели сведены в таблицу 4. Кроме указанных параметров для вычисления времени доступа используются и характеристики ЭВМ CDC, которые подробно описаны в работах /9, 12/. С помощью этих характеристик вычисляются величины $T_{z\,инв}$, $T_{z\,acc}$, T_{z_1}, T_{z_2}.

В связи с этим необходимо иметь в виду, что для чтения с магнитного диска информации величиной в одну физическую единицу записи в среднем затрачивается время $t_п + \frac{1}{2}t_o$. Следовательно, для чтения информации в семь физических единиц (т.е. 7PRU) необходимо время $T_7 = t_п + \frac{1}{2}t_o + t_o$, где $t_п$=75мс, t_o=25мс. Следовательно, $T_7 = 1{,}3(t_п + \frac{1}{2}t_o)$. Для чтения информации величиной в 140 физических единиц записи необходимо $T_{140} = 6{,}7(t_п + \frac{1}{2}t_o)$. Более подробно методика расчета времени доступа описана в работе /12/. Необходимо заметить, что время доступа в двухуровневой частично инвертированной структуре рассчитано для величин зон Z_1= 7PRU и Z = 140PRU.

$T_{инв.}$ для запроса, состоящего из пяти дескрипторов, можно расчитать по формуле (I). Для этого, если на условный номер документа уходят в среднем два слова CDC, то формула (I) преобразуется в формулу (Ia), где

$$T_{инв.} = 5(t_п + \frac{1}{2}t_o + \frac{2L}{448}t_o) = 5t_п + \frac{5}{2}t_o + \frac{5L}{224}t_o,$$

при $t_п$= 75 мс, t_o= 25 мс, L=II $T_{инв.} = 5{,}1(t_п + \frac{1}{2}t_o)$.

В данном случае считается, что к каждому дескриптору в среднем относится L документов.

Из данного примера видно, что время доступа в инверсном массиве зависит как от средней длины списка, так и, в еще большей степени, от числа положительных дескрипторов запроса.

$T_{acc.}$ вычисляется по формуле (2). Для массива из I0000 документов $L = 11$ (см. табл.4). При расчете $T_{zacc.}$ принимается, что величина узла не превосходит одной физической единицы записи. Тогда

$$T_{zacc.} = t_n + \frac{1}{2}t_o \quad \text{и} \quad T_{acc.} = 11\left(t_n + \frac{1}{2}t_o\right).$$

Из данного примера видно, что среднее время обработки ассоциативно-адресного массива зависит прежде всего от средней длины списка.

$T_{двухур.}$ можно рассчитать по формуле (3). При $T_{z_1} = 1{,}3\left(t_n + \frac{1}{2}t_o\right)$ и $T_{z_2} = 6{,}7\left(t_n + \frac{1}{2}t_o\right)$, $L_{кома} = 2{,}61$. Тогда $T_{двухур.} = 9{,}5\left(t_n + \frac{1}{2}t_o\right)$.

Если обе зоны имеют одинаковую величину в семь физических единиц, то

$$T_{двухур.} = 5{,}58 \cdot 1{,}3\left(t_n + \frac{1}{2}t_o\right) + 0{,}18 \cdot 8{,}03 \cdot 1{,}3\left(t_n + \frac{1}{2}t_o\right) = 9{,}1\left(t_n + \frac{1}{2}t_o\right).$$

данный пример показывает, что, выбирая соответствующим образом параметры частично инвертированной структуры, можно улучшить среднее время поиска в этой структуре.

Выводы.

I. На основе полученных результатов можно утверждать, что самым близким к реальному распределению является распределение функций $f(j)$ и $P(j)$ по закону Ципфа.

2. Используя формулы (1-3) и параметры предложенной модели, можно рассчитать среднее время доступа в указанных структурах. Тем самым для конкретной ИПС можно выбрать самую подходящую структуру.

3. Данная модель дает возможность судить об эффективности разрабатываемой ИПС на конкретной ЭВМ в зависимости от характера $f(j)$ и $P(j)$. Следовательно, на данной ЭВМ можно построить ИПС в среднем достаточно эффективную для различного рода банков данных при условии, что все они имеют одинаковый закон распределения.

4. Описанная модель дает возможность выбирать параметры рассмотренных структур в зависимости от конкретной ЭВМ и природы банка данных.

Литература

I. IBM Corporation: Bibliography on Simulation, White Plains, C20-I649, I969.

2. Cooper M.D., J. Information Storage and Retrieval, v 9, pp. I3-32, I973.

3. Senko M.E. oth., Proceedings of IFIP, pp. 5I4-5I9, I969.

4. Griffiths J.M. oth., ASLIB Conference, pp. I4I-I52, Amsterdam, I976.

5. Cardens A.F., Comm. ACM, v I6, No.9, pp. 540-548, I973.

6. Griffiths J.M., J. Documentation, v. 3I, No.3, pp. I85-I90, I975.

7. Lefkovitz D., "File Structures for On-Line Systems", Spartan Books, I969.

8. Prywes N.S., Proceedings of IFIP, pp. 273-278, I962.

9. Арнаудов Д.Д. Препринт ОИЯИ, I0-7953, I974, Дубна.

I0. Арнаудов Д.Д. Сообщение ОИЯИ, PI0-9I78, I975, Дубна.

II. Lowe T.C. J.ACM, v I5, No.4, pp. 534-548, I968.

I2. Арнаудов Д.Д., Сообщение ОИЯИ, I0-7949, I974, Дубна.

I3. Арнаудов Д.Д. и др. Депонированная публикация ОИЯИ, БI-II-8553, I975, Дубна.

Рукопись поступила в издательский отдел
22 марта I977г.