СООБЩЕНИЯ
ОБЪЕДИНЕННОГО
ИНСТИТУТА
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

Р10 - 9178

Д.Д.Арнаудов

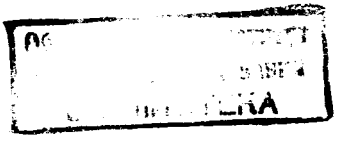
ОБ ОДНОМ СПОСОБЕ ОРГАНИЗАЦИИ МНОГОУРОВНЕВОЙ АДАПТИРУЮЩЕЙСЯ ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ

1975

Р10 - 9178

Д.Д.Арнаудов

ОБ ОДНОМ СПОСОБЕ ОРГАНИЗАЦИИ МНОГОУРОВНЕВОЙ АДАПТИРУЮЩЕЙСЯ ИНФОРМАЦИОННО-ПОИСКОВОЙ СИСТЕМЫ



Проблема многоуровневой организации информационно-поисковой системы заключается прежде всего в организации основного информационного массива и системы поиска. Система поиска в иерархической ИПС подробно рассмотрена нами в[1]. Особенность этой системы заключается в необходимости определения последовательности правил сравнения. Под этим понимается последовательно усложняющийся анализ элемента с постепенным охватом все большего числа его деталей, при этом результаты, полученные на предыдущем этапе анализа, исследуются на последующем. Необходимость в использовании последовательности этапов анализа вызвана тем, что по мере того, как растет основной информационный массив, применять самый сложный анализ к каждому элементу оказывается слишком дорогостоящим делом. В связи с этим необходимо отметить, что, со своей стороны, основной информационный массив должен иметь подходящую структуру, обеспечивающую возможность использования менее сложных механизмов индексации и промежуточных мер для определения степени соответствия, чем окончательная мера соответствия. Именно этот аспект порождает значительные трудности при разработке системы информационного поиска, связанные с согласованием объема памяти поискового массива и времени для получения ответа. Эти трудности могут быть разрешены на основе создания самоорганизующейся многоуровневой информационно-поисковой системы, где число уровней иерархии основного информационного массива и величина сегмента

массива определенного уровня могла бы динамически изменяться в процессе нарастания информационного массива согласно определенному критерию и тем самым регулировать соответствующее отношение между временем, затрачиваемым на поиск элемента массива, и объемом расходуемой памяти.

Анализируя различные организационные структуры основного информационного массива, Лефковитц /2/ отмечает, что можно различить два полюса в организации массивов - инверсную и мультисписковую (ассоциативно-адресную) организацию. Все остальные варианты организации можно рассматривать как частично инвертированные мультисписковые схемы. Проведенный нами анализ инверсной и мультисписковой организации информации /3/ показал, что с нарастанием объема основного информационного массива ухудшаются поисковые характеристики как мультисписковой, так и инверсной организации. В данной работе мы покажем, что улучшение поисковых характеристик системы можно получить на базе комбинирования признаков инверсной и мультисписковой организации, создавая многоуровневую самоорганизующуюся поисковую систему на базе определенной стратегии поиска /I/. В основе методики разработки подобной системы заложен принцип модификации мультисписковой организации в сторону "инверсности".

На рис.I,2,3 показано возможное развитие структуры информационного массива. На рис.I представлена классическая ассоциативно-адресная структура. Управляющая часть представляет набор дескрипторов словаря системы. Коды дескрипторов совпадают с номерами (индексами) элементов массива. В каждом дескрипторе имеется информация о первом члене узлового списка (АС- адрес связи) и о числе членов этого списка (ЧД - число членов). В массиве ОМПОД расположены соответствующие узловые списки.

На рис.2 показана уже частично-инвертированная структура. Все уровни списка в массиве ОМПОД сегментированы по зонам. Управляющая часть представляет сложную структуру,где хранятся заголовки дескрипторных списков. Сами эти заголовки соединены в цепочки при помощи адресов связи ($NZ1$, $AC1$). Расположение первого элемента данного списка определяется тоже с помощью адреса связи ($AC2$, $NZ2$, $AD2$). Более подробно эта структура рассмотрена в /4/. Ясно, что управляющая часть состоит из двух частей, т.е. имеет двухуровневую структуру.

Управляющая часть с трехуровневой структурой (массив ОМД из двух частей и массив МЗД) показана на рис.3. Здесь в массиве МЗД собраны заголовки дескрипторов узловых списков массива ОМПОД, а в массиве ОМД находятся начала цепных списков заголовков массива МЗД /4/.

Исходя из принятой стратегии поиска /I/, можно предложить следующие формулы для подсчета среднего времени обработки информационного массива с одноуровневой, двухуровневой и трехуровневой структурой.

$$T_{одноур} = K N_p T_Z + L_S t_S = K N_p T_Z + L_S (t_n + 1/2\, t_c), \quad (I)$$

$$T_{двухур} = K N_p T_Z + a_{смд} C_{ксмд} T_{Z_1} + a_{смпод} C_{смпод} T_{Z_2}, \quad (2)$$

$$T_{трехур} = K N_p T_Z + a_{смд} C_{ксмд} T_{Z_1} + a_{мзд} C_{кмзд} T_{Z_2} + a_{смпод} C_{ксмпод} T_{Z_3}. \quad (3)$$

Рассмотрим отдельные элементы предложенных формул. Первые одночлены всех формул одинаковы. Здесь учитывается время, которое расходуется на декодирование дескрипторов запроса. Необходимо отметить, что любой запрос представляет конъюнкцию дескрипторов /5/. В дальнейшем рассматриваются только положительные дескрипторы - N_p (дескрипторы без знака отрицания). Дескрипторы с отрицанием

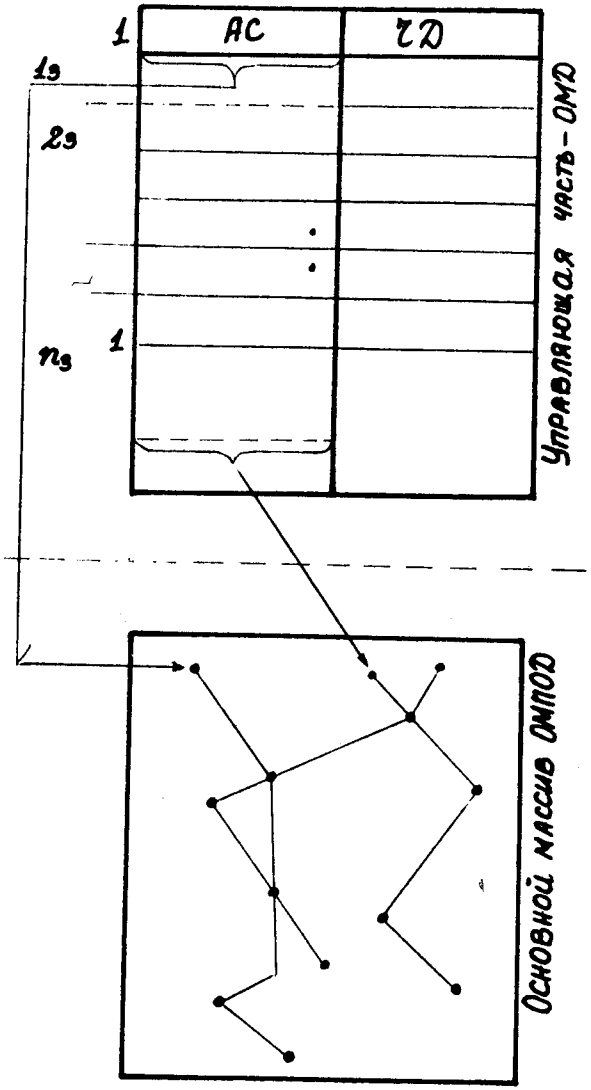

1
AC
ZD
1з
2з
nз
1
Управляющая часть - ОМД
Основной массив ОМПОД

Рис. I

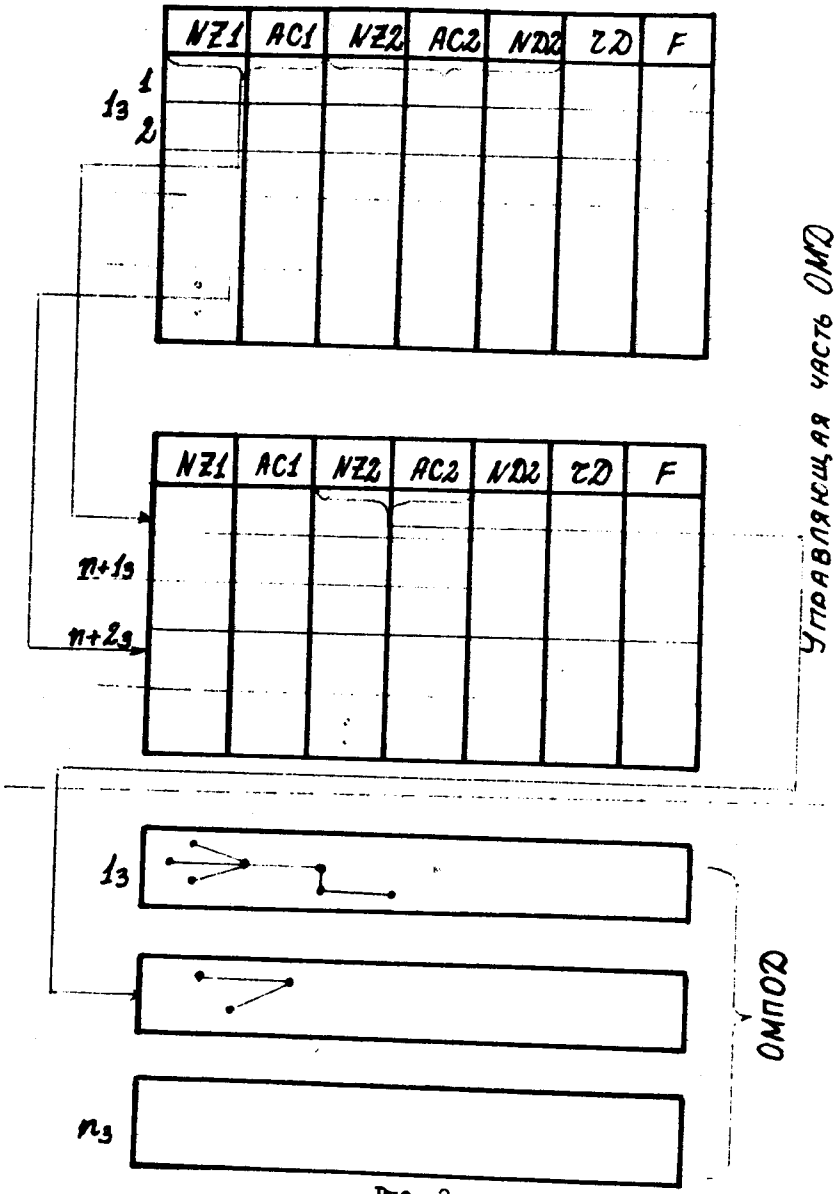

NZ1
AC1
NZ2
AC2
NDZ
ZD
F
1
1з
2
n+1з
n+2з
Управляющая часть ОМД
1з
nз
ОМПОД

Рис. 2

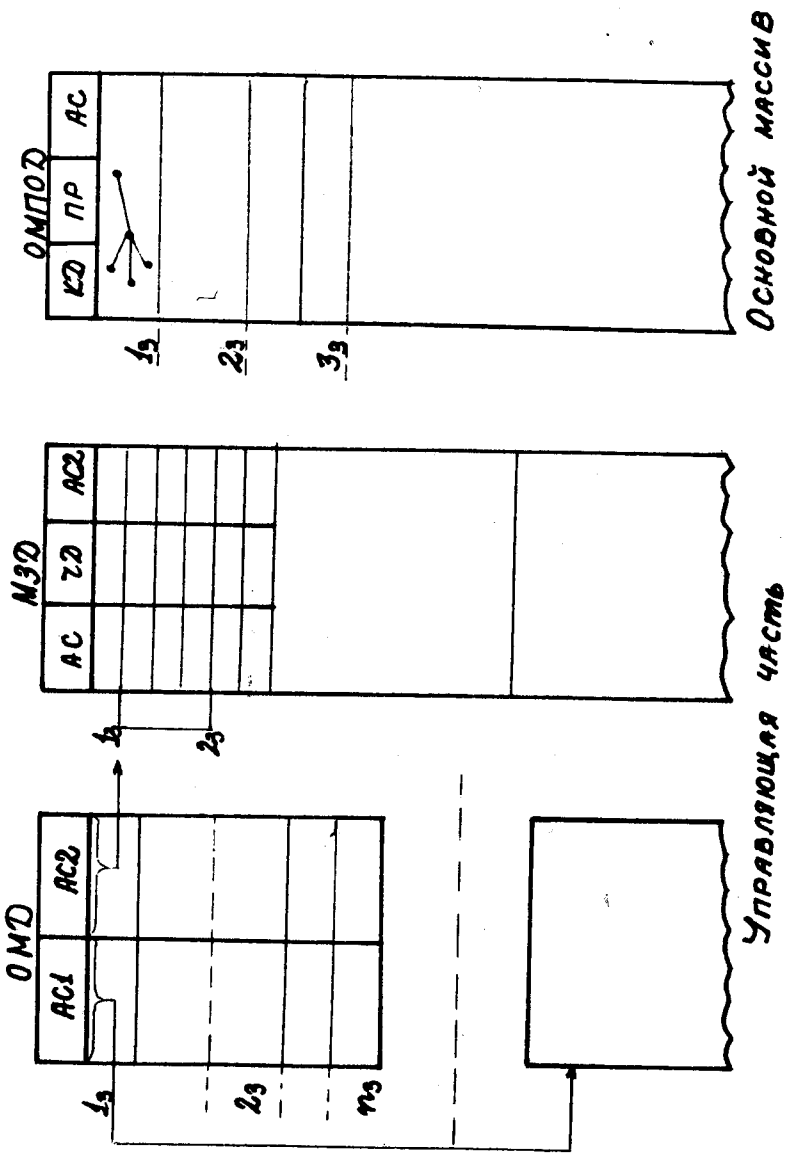


Рис. 3

используются только на последнем этапе поиска, в узлах ОМПОД, и практически не влияют на время ответа, вычисляемого по формулам I-3. Если запрос состоит из суммы произведений (т.е. представлен в дизъюнктивной форме /5/), время обработки запроса меньше, чем суммы времени обработки отдельных конъюнкций благодаря принятой стратегии параллельной обработки запросов /1/.

Коэффициент К учитывает одновременную выборку $К\Lambda р$ дескрипторов запроса. Необходимо заметить, что первая часть массива ОМД (во всех рассматриваемых структурах) логически сегментирована по зонам и возможно, что для некоторых дескрипторов запроса необходимо читать только одну зону. Это намного сокращает время выборки и учитывается коэффициентом К. Параметр T_Z обозначает время чтения зоны (аналогично T_{Z_1}, T_{Z_2}, T_{Z_3}). Параметр C_K обозначает среднее количество зон на дескриптор и имеет место в формулах 2 и 3. $C_{КСМД}$ показывает среднее количество зон на дескриптор в П-й части массива ОМД. $C_{КМЗД}$ и $C_{КСМПОД}$ показывает среднее количество зон на дескриптор соответственно в массивах МЗД и ОМПОД. Параметры $a_{СМД}$, $a_{МЗД}$, $a_{СМПОД}$ показывают отношение среднего количества общих зон запроса к $C_К$. Следовательно:

$$a_{СМД} = \frac{ЗЗ_{СМД}}{C_{КСМД}}; \quad a_{МЗД} = \frac{ЗЗ_{МЗД}}{C_{КМЗД}};$$

$$a_{СМПОД} = \frac{ЗЗ_{СМПОД}}{C_{КОМПОД}}.$$

Здесь $ЗЗ_{СМД}$, $ЗЗ_{МЗД}$, $ЗЗ_{СМПОД}$ обозначают среднее количество зон запроса соответственно в массивах ОМД, МЗД, ОМПОД.

Введем еще несколько параметров, которые встречаются в нашем изложении.

Параметры $t_п$, t_c, $t_{тр}$, t_S являются характеристиками операционной системы при работе с дисковым устройством, где

$t_п$ – время выборки цилиндра;

t_c – время оборотов дискового пакета;

$t_{тр}$ – время трансмиссии n символов;

t_S – время выборки информации одной физической записи с диска.

Они подробно рассмотрены нами в /I,4/.

Особенно важный параметр – это среднее количество документов на дескриптор, L, которым характеризуется средняя длина списка. Эта характеристика является среднестатистической величиной и связана с частотой встречаемости дескрипторов – $f(j)$/3/. В данном случае ее можно вычислить по формуле

$$L = \frac{N_z N_к}{V}, \qquad (4)$$

где N_z – количество документов основного информационного массива;

$N_к$ – средняя глубина индексирования документа;

V – количество различных дескрипторов.

В формуле (I) через L_S обозначена длина самого короткого дескрипторного списка. Анализируя эту формулу, необходимо отметить, что в схеме на рис.I поиск происходит при движении по самой короткой цепи L_S. Так как списки не сегментированы, то каждый раз происходит выбор отдельного узла с диска. Принимается, что максимальный размер узла не превышает одной физической единицы записи и поэтому

$$t_S = t_п + 1/2\, t_c$$ (более подробно см./I,3,4/).

Исходя из принятой стратегии поиска/I/ в двухуровневой и трехуровневой структуре, необходимо просмотреть все элементы цепи дескрипторов в ОМД (т.е. в высшем уровне иерархии), поэтому в среднем в этом массиве всегда просматривается среднее количество зон на дескриптор. Тогда $a_{смд} = 1$ и формулы (2) и (3) принимают следующий вид:

$$T_{двухур} = K N_р T_z + C_{кснд} T_{z_1} + a_{смпод} C_{кмпод} T_{z_2}, \qquad (2a)$$

$$T_{трехур} = K N_р T_z + C_{кснд} T_{z_1} + a_{мзд} C_{кмзд} T_{z_2} + a_{смпод} C_{кмпод} T_{z_3}. \qquad (3a)$$

Основной вопрос, который необходимо решить на основе предложенных формул (I, 2a, 3a), – это когда и в каких случаях необходимо переходить от одной структуры к другой (см.рис.I,2,3) и какие величины зон сегментированных списков являются самыми подходящими при проведении поиска в соответствующих структурах.

Так как первое слагаемое одинаково во всех формулах (I÷3), дальнейший сравнительный анализ проведем, используя следующие выражения:

(I') $T_{сонсур} = L(t_п + 1/2\, t_c)$,

(2') $T_{двухур} = C_{кснд} T_{z_1} + a_{смпод} C_{кмпод} T_{z_2}$,

(3') $T_{трехур} = C_{кснд} T_{z_1} + a_{мзд} C_{кмзд} T_{z_2} + a_{смпод} C_{кмпод} T_{z_3}$.

В формуле (I') вместо L_S берем L – среднюю длину списка в ОМПОД, т.к. $\frac{\sum_{S=1}^{N} L_S}{N} \leq L$,

где N – число всех дескрипторов. Средняя длина списка L вычисляется по формуле (4) и, используя формулу $L_{оптим}$ /3/, можно вычислить, при каком количестве документов ожидается, что все N дескрипторов словаря системы будут использованы.

$$V = \frac{N_z N_к}{\ell n N + \gamma}. \qquad (5)$$

Для информационного массива принимаем $N_к$ =9.

Тогда

$$10000 = \frac{N_z \cdot 9}{\ln 10000 + 0{,}57}\ ;\quad N_z = 10720.$$

Это означает, что для информационного массива из документов ИНИСа[6] при объеме 10720 документов ожидается, что все N дескрипторов словаря системы будут встречаться в документах информационного массива.

В дальнейшем в качестве характеристик массива будут использоваться характеристики информационного массива, получаемого на базе магнитной ленты ИНИСа[6].

На рис.4 приведена зависимость $L = f(N_z)$.

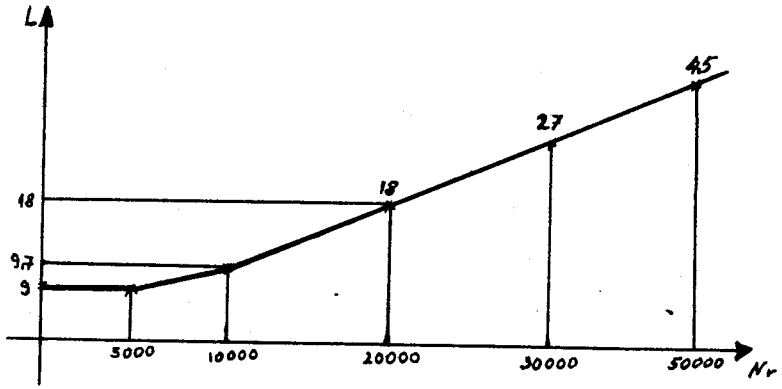


Рис.4

Величина L является основной при решении вопроса о переходе от одной структуры к другой.

Первый вопрос, который необходимо решить, - это при каком L_{min} $T_{двухур.} < T_{одноур.}$, т.е. $T_{двухур.} < L(t_п + 1/2\, t_о)$.

С целью определения L_{min} проведен машинный эксперимент. В качестве экспериментального массива моделировался массив по закону Ципфа с поправкой L_{OWE} [3]. Получено, что $T_{двухур.} < T_{одноур.}$ для массива с параметрами L =9, N_z =3220. Оптимальная вели-

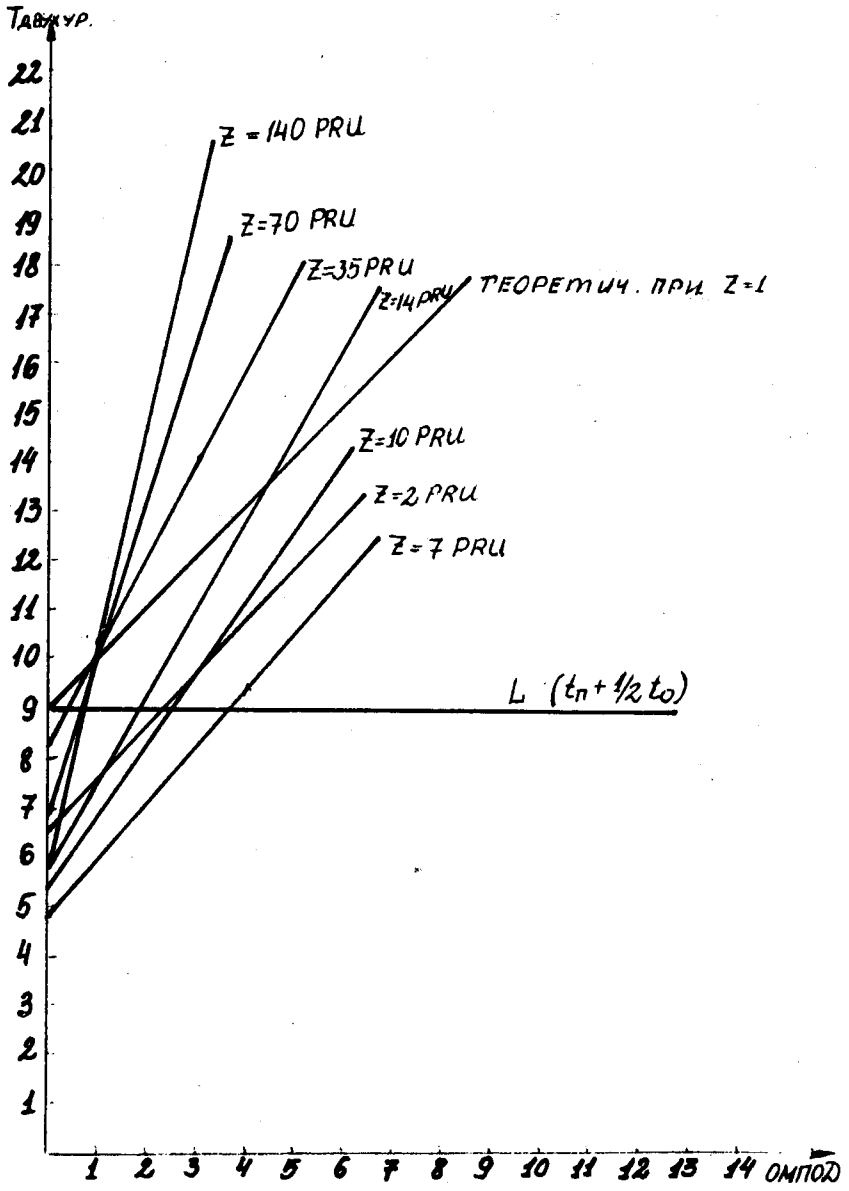


Рис. 5

чина зоны (Z), в этом случае $Z_1 = Z_2 = 7PRU$ (о вычислении величины зоны в физических единицах см.[17]). На рис.5 показан график $T_{двухур} = f(r3_{ОМПОД})$ при $Z = const$. В этом эксперименте Z меняется от $2PRU$ до $140PRU$. Самая хорошая прямая времени - при $Z = 7PRU$. Как видно из рисунка, при обработке $r3_{ОМПОД} \leq 3$ Т двухуровн. < Т одноуровн. Обработка трех зон основного массива является вполне достаточной, и в этом случае α ОМПОД = 0,43. Исследования Лефковитца[21] показали, что при запросах, состоящих не меньше чем из трех дескрипторов, α ОМПОД варьируется в пределах 0,1÷0,2. В наших экспериментах (среднее число дескрипторов запроса равно четырем) α ОМПОД ≤ 0,1. Значения параметров двухуровневой структуры при объеме массива 3200, 10000, 20000 и 50000 документов экспериментально получены и сведены в таблицу 1А. На базе данного эксперимента исследованы и получены коэффициенты K_1 и K_2, которые являются основными в расчете основных параметров массива. Об этих коэффициентах речь пойдет дальше, а сейчас приведем формулы для расчета основных параметров массива. Эти формулы будут использованы дальше для эмпирического расчета параметров в выражении (1, 2, 3) при различных объемах информационного массива.

Сначала приведем формулы расчета параметров двухуровневой структуры.

$$C_{КОМПОД} = \frac{K_1 Z_{ОМПОД}\, 64\, r3_{ОМПОД}}{2V} \quad . \qquad (6)$$

Вывод формулы покажем на базе анализа физической сущности $C_{КОМПОД}$. Так как $C_{КОМПОД}$ является средней характеристикой числа зон (основного массива ОМПОД) на дескриптор, то

$$C_{КОМПОД} = \frac{\text{число всех заголовков}}{V} \quad . \qquad (7)$$

σ_1			I,42	I,63	I,7I	I,78	2	2,32
σ_2			I,I	I,I	I,I2	I,I9	I,24	I,32
K_1			0,70	0,67	0,64	0,56	0,50	0,43
K_2			0,90	0,90	0,89	0,84	0,80	0,75
Z (PRU)			7PRU	10 PRU	14 PRU	35 PRU	70 PRU	140 PRU
L	V	Nc						
9	3220	3200	$C_{КОМПОД}$	6,33	5,8	5,3	4,0	3,2	2,6
			$C_{К\ ОМД}$	4,46	3,85	3,3	2,1	I,4	0,6
			$r3_{ОМПОД}$	I32	93	66	27	I3	6,5
			$r3_{ОМД}$	76	46	32	9	3,2	I,2
9,49	9478	I0000	$C_{КОМПОД}$	6,72	6,43	6,I3	5,34	4,73	4,09
			$C_{К\ ОМД}$	5,2	4,9	4,57	3,66	2,97	2,27
			$r3_{ОМПОД}$	40I	283	20I	8I	4I	2I
			$r3_{ОМД}$	242	I62	I09	37	I6	I
I8,08	9954	20000	$C_{КОМПОД}$	I2,8	I2,24	II,68	I0,I8	9,0I	7,79
			$C_{К\ ОМД}$	I0,96	I0,36	9,72	8,00	6,62	5,I6
			$r3_{ОМПОД}$	80I	565	40I	I62	82	4I
			$r3_{ОМД}$	523	35I	237	82	36	I6
27,0I	9995	30000	$C_{КОМПОД}$	I8,59	I8,0I	I7,47	I5,23	I3,5	II,0
			$C_{К\ ОМД}$	I6,08	I6,05	I5,09	I2,53	I0,0	8,I
			$r3_{ОМПОД}$	I20I	847	60I	242	I20	60
			$r3_{ОМД}$	782	543	367	I28	55	23

Таблица 1а

Ясно, что число всех зон массива ОМПОД равняется числу заголовков узловых списков, сегментированных в конкретных зонах. Априори ясно, что число всех заголовков меньше числа всех элементов массива ОМПОД, т.к. предполагается, что в узловых списках имеется более одного члена (в худшем случае, если есть в списках по одному члену, то число всех заголовков будет равно числу элементов массива ОМПОД). Под "элементом" массива ОМПОД (а также ОМД и МЗД) понимаем физическую величину, состоящую из двух машинных слов (более подробно см. /14/). Тогда

$$\text{Число всех заголовков} = \frac{K_1 Z_{ОМПОД} 64 Z3_{ОМПОД}}{2} \quad . \qquad (8)$$

В формуле (8) K_I является коэффициентом встречаемости и связан с числом различных дескрипторов в данной зоне (каждый дескриптор является основной частью одного элемента). Ясно, что $K_I \leq I$. Величины K_I вычислены экспериментально и сведены в таблицу IA. Видно, что величина этого коэффициента не зависит от длины массива, а зависит только от величины зоны. Следовательно, физическая природа этого коэффициента показывает, что он узко связан с характеристиками входного потока и частотой встречаемости дескрипторов - $f(j)$

Заметим, что в дальнейшем используется и обратный коэффициент $G_1 = 1/K_1$, который называется коэффициентом связанности и обозначает число членов в списке данного дескриптора в данной зоне ОМПОД. Этот коэффициент тоже вычислен экспериментально и показан в таблице IA. Видно, что для распределения, генерируемого по закону Ципфа (с поправкой L_{OWF}), видно, формируемые списки являются недостаточно плотными (в среднем по 2,32 члена в списке при $Z_{ОМПОД} = I40\ PRU$).

В используемых формулах $Z_{ОМПОД}$ обозначает величину зоны в физических единицах (PRU) ($Z_{ОМПОД}$ меняется от 7 до I40).

Коэффициент 64 обозначает, что в одной физической единице (PRU) находятся 64 машинных слова.

Параметры $Z3_{ОМПОД}$ и V были рассмотрены нами раньше. Имея в виду все вышеописанное, подставляя (8) в (7), получим формулу для вычисления $C_{КОМПОД}$ (6).

С другой стороны, можно показать взаимосвязь этого параметра с другими параметрами информационного массива. Для этого будем использовать формулу для подсчета $Z3_{ОМПОД}$.

$$Z3_{ОМПОД} = \frac{2 N_z N_k}{64 Z_{ОМПОД}} \quad . \qquad (9)$$

Подставляя (9) в (6), получим:

$$C_{КОМПОД} = \frac{K_1 N_z N_k}{V} \quad . \qquad (IO)$$

Подставляя (4) в (IO), получим:

$$C_{КОМПОД} = K_1 L \quad . \qquad (II)$$

Формулы (IO) и (II) выражают $C_{КОМПОД}$ при помощи числа документов массива - N_z, среднего числа дескрипторов на документ - N_k и средней длины узлового списка - L.

На рис.6 показан график изменения $C_{КОМПОД}$ при изменении величины зоны, т.е. $C_{КОМПОД} = f(Z_{ОМПОД})$. Здесь использованы экспериментальные данные массива на 3200 документов.

Перейдем к анализу формулы для вычисления $C_{КОМД}$.

$$C_{КОМД} = \frac{K_2 Z3_{ОМПОД} Z_{ОМД} \cdot 64}{2V} \quad . \qquad (I2)$$

Физическая сущность этого параметра в том, что он обозначает среднее число зон на дескриптор в массиве ОМД. Необходимо, однако, отметить, что речь идет о числе зон во второй части массива (см. рис.2), поскольку в первой части совершается автоматическая выборка по кодам дескрипторов и она влияет только на первое сла-

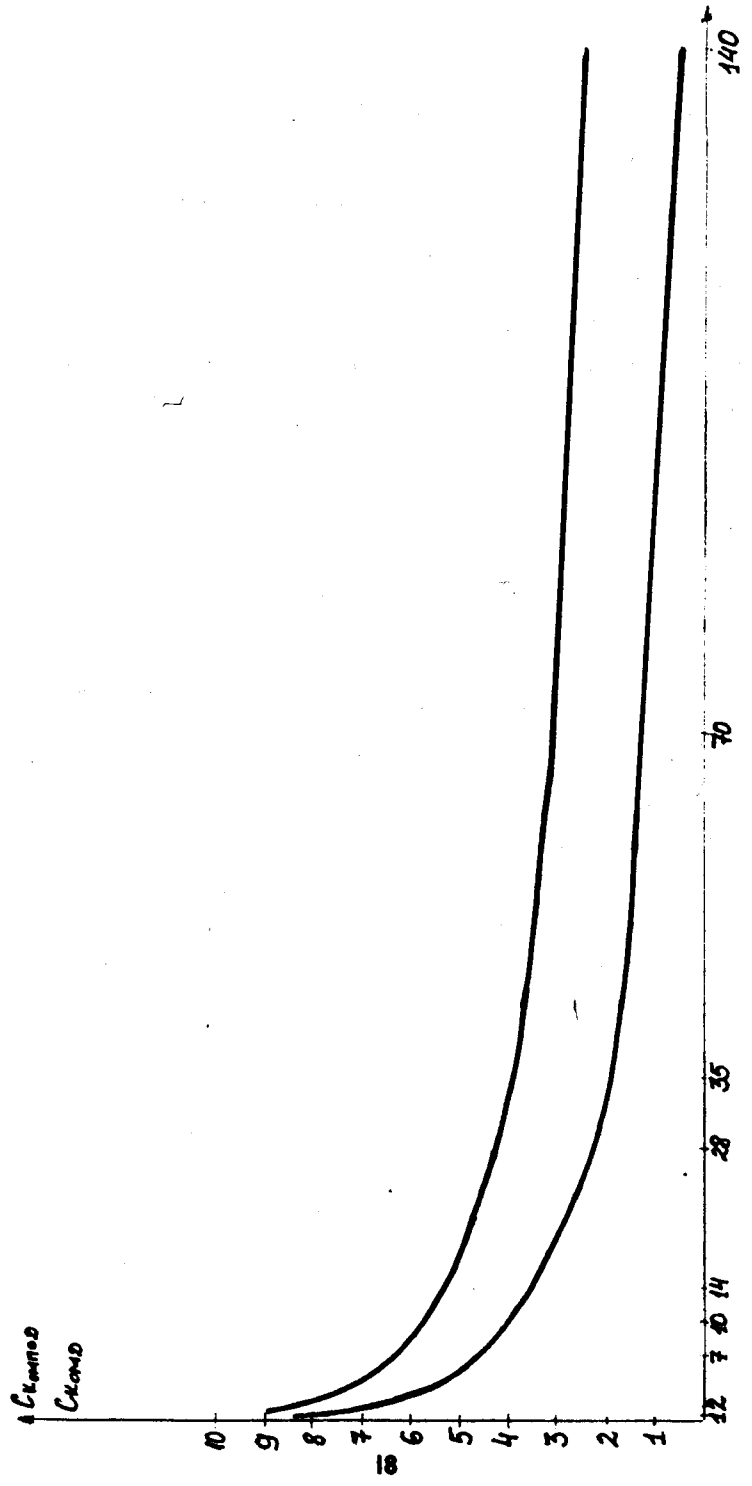


Рис. 6

гаемое формулы (2), которое является общим для всех трех формул и в сравнительном анализе не рассматривается. Так как к первой части массива ОМД относится столько элементов массива (в данной структуре элементами являются заголовки цепных списков массива ОМПОД), сколько имеется различных дескрипторов (V), то

$$C_{K_{OMD}} = \frac{K_2 (\mathcal{ZЭ}_{OMD} - V)}{V} . \qquad (13)$$

Как уже заметили,

$$\mathcal{ZЭ}_{ОМД} = \text{числу всех заголовков} . \qquad (14)$$

Здесь вводится впервые коэффициент на K_2, выполняющий такую же функцию, как и коэффициент K_I по отношению к параметру $C_{кОМПОД}$. В данном случае K_2 связан с той частью элементов массива ОМД, принадлежащих данной зоне, которые являются представителями различных дескрипторных заголовков. Ясно, что $K_2 < I$. Обратный коэффициент $G_2 = 1/K_2$ является коэффициентом связанности и показывает наличие больше чем одного заголовка одного и того же дескриптора в одной зоне ОМД. Коэффициенты K_2 и G_2 получены экспериментальным путем и сведены в таблицу IA. Из таблицы видно, что даже если зона равняется I40 PRU, то там встречается не больше чем I,32 членов в цепном списке дескриптора. А для зоны $Z = 7PRU$ - только по I,I члену.

Исходя из приведенных рассуждений, нужно отметить, что

$$\mathcal{ZЭ}_{OMD_{II}} = \mathcal{ZЭ}_{OMD} - V , \qquad (15)$$

где $\mathcal{ZЭ}_{OMD_{II}}$ - это число элементов второй части массива ОМД.

Для вычисления по формуле (I2) необходимо выразить $\mathcal{ZЭ}_{OMD_{II}}$ через $\mathcal{ZЗ}_{OMD_{II}}$ (число зон второй части массива ОМД).

$$\mathcal{ZЭ}_{OMD_{II}} = \frac{\mathcal{ZЗ}_{OMD} \cdot Z_{OMD} \cdot 64}{2} . \qquad (16)$$

Подставляя (I6) в (I5), а (I5) в (I3), получим формулу (I2).

Можно установить некоторую взаимосвязь между $C_{кОМД}$ и $C_{кОМПОД}$. Если выразить $z_{зм_{II}}$ через $z_{эм_{II}}$, то получим:

$$z_{зм_{II}} = \frac{2 z_{эм_{II}}}{Z_{смп} 64} = \frac{2(\text{число всех заголовков} - V)}{Z_{смп} 64}, \quad (17)$$

число всех заголовков = $\frac{K_1 Z_{смп} 64 \; z_{зм\,опт}}{2}$,

тогда $z_{зм_{II}} = \frac{2\left(\frac{K_1 Z_{смпод}^{2} 64 \; z_{зм\,опт}}{2} - V\right)}{Z_{смп} \cdot 64} =$

$$= \frac{K_1 Z_{смпод} 32 \; z_{зм\,опт} - V}{Z_{смп} 32}.$$

Подставляя результат в (I2), получим:

$$C_{кОМД} = K_2(C_{кОМПОД} - I). \quad (18)$$

График зависимости $K_{кОМД}$ от величины зоны при $L = const$ показан на рис.6. Необходимо отметить, что с изменением величины объема массива $C_{кОМД}$ изменяется по линейному закону.

Используя приведенные уже формулы (I,2), а также формулы (I÷I8), покажем изменение оптимальных зон (Z оптим.$_{ОМД}$, Z оптим.$_{ОМПОД}$) при изменении объема массива (соответственно изменяется и L).

В формулах (I÷I8) в качестве коэффициентов K_I и K_2 использованы полученные экспериментальным путем данные для этих коэффициентов. Они сведены в таблицу IA. Как уже заметили, эти коэффициенты являются постоянными для определенной величины зоны и не изменяются (очень слабо изменяются) с увеличением объема массива. Вычислены $C_{кОМД}$ и $C_{кОМПОД}$ для массива из I00000 и 500000 документов. Данные приведены в таблице 2а.

Используя приведенные данные и формулы (I')и (2'), получаем следующие результаты.

Таблица 2А

Тдвухур. (одна зона)	L	V	N	Z(PRU)	7	I0	I4	35	70	I40
$71(t_n + t_{zo})$	90	I0000	I00000	$C_{кОМД}$	43	53	50	4I	34,5	27
				$C_{кОМПОД}$	63	60	58	50	45	38,5
				$z_{зо.пд}$	2900	-	-	-	-	-
				$z_{зо.под}$	4000	-	-	-	-	-
$356(t_n + t_{zo})$	450	10000	500000	$C_{кОМД}$	220	-	252	-	-	-
				$C_{кОМПОД}$	315	-	289	-	-	-
				$z_{зОМД}$	I3000	-	-	-	-	-
				$K_{зОМПОД}$	20000	-	-	-	-	-

Если $Z_{ОМД} = Z_{ОМПОД} = 7PRU$, то как при N_Z =I00000, так и при N_Z =500000 Т двухур. < Т одноур. при $\alpha_{ОМПОД} \leq 0{,}43$. Это означает, что $Z = 7PRU$ является оптимальной при исследовании не больше чем 43% массива ОМПОД. Однако при этом получается довольно большая затрата памяти на массив ОМД - около 62% от массива ОМПОД. Начиная примерно с 30000 документов массив ОМД занимает большое место на диске.

Необходимо отметить, что в реальных случаях исследуется не больше, чем 20% ОМПОД (по данным Лефковитца[2]), а в наших исследованиях (для данного экспериментального массива) - не больше I0%, т.е. $\alpha_{ОМПОД} \leq 0{,}I$. В таком случае при данной оптимальной зоне $Z_{ОМП} = 7PRU$ имеется лишний запас, т.к. $\alpha_{ОМПОД} \leq 0{,}43$. Память, расходуемую на ОМД, можно уменьшить, не увеличивая Т двухур., но уменьшая границы изменения коэффициента $\alpha_{ОМПОД}$. Этого можно достичь путем увеличения размера зоны массива ОМПОД. Хорошие результаты получаются при увеличении зоны, начиная с массива величиной в 30000 документов (если массив меньше этой длины, то при увеличении зоны увеличивается Т двухур). Значение основных характеристик для массивов из 30000, I00000, I50000, 500000 документов сведены в таблицу 3А. Имеется в виду, что $Z_{ОМПОД} = I40PRU$, а $Z_{ОМД} = 7PRU$. В данном случае характеристики лучше, чем в случае $Z_{ОМД} = Z_{ОМПОД} = 7PRU$ (см.табл. 2А и 3А). Кроме того, в данном случае за счет изменения границы $\alpha_{ОМПОД}$ достигнуто соотношение между массивами $\frac{ОМД}{ОМПОД} = 0{,}37$, при котором ОМД уже составляет 37% от основного массива. Это лучшие возможные результаты, т.к. $Z_{ОМПОД}$ принимает максимально допустимое значение - I40PRU. Конечно, в зависимости от конкретных требований к поисковой системе в других случаях можно увеличить границы $\alpha_{ОМПОД}$ (например, до 0,25), но тогда необходимо, чтобы $Z_{ОМПОД} = 70PRU$.

Таблица 3А

	L	V	N_Z	L	V	N_Z	L	V	N_Z	L	V	N_Z	L	V	N_Z
	27	I0000	30000	90	I0000	I00000	I35	I0000	I50000	450	I0000	500000	900	I0000	I000000
$C_{КОМД}$	8,9			2,6			42,5			I70			347		
$C_{КОМПОД}$	II			39			58			I94			387		
$Z_{ОМД}$	450			I290			2200			7500			7300		
$Z_{ОМПОД}$	60			200			300			I000			2000		
$\alpha_{ОМПОД}$	$\leq 0{,}I$			$\leq 0{,}I$			$\leq 0{,}I$			$\leq 0{,}I$			$\leq 0{,}I$		
$T_{двухур.}$	$10{,}4(t_п + 1/2\ t_о)$			$62{,}5(t_п + 1/2\ t_о)$			$9\ G(t_п + 1/2\ t_о)$			$351(t_п + 1/2\ t_о)$			—		

Увеличение зоны ОМД больше чем на 7PRU ведет к ухудшению временных характеристик. Понятно, что это происходит из-за медленного уменьшения $C_{кОМД}$ (см.рис.6) с увеличением величины зоны, а в этом случае Т двухур. резко повышается.

Исходя из проведенных экспериментов и вышеописанного анализа, можно сделать следующие выводы.

Для основного информационного массива величиной до 3000 документов необходимо, чтобы структура системы была одноуровневой (см.рис.I).

Переход к двухуровневой структуре необходимо осуществить при массиве выше 3000 документов. Тогда при увеличении объема массива до 30000 документов оптимальной является зона величиной 7PRU, т.е. $Z_{ОМД} = Z_{ОМПОД} = 7PRU$

При массиве выше 30000 документов необходимо переключить систему на следующие оптимальные зоны: $Z_{ОМД} = 7PRU$, $Z_{ОМПОД} = 140PRU$.

График изменения $Z_{опт.ОМД}$ и $Z_{опт.ОМПОД}$ с изменением величины массива показан на рис.7.

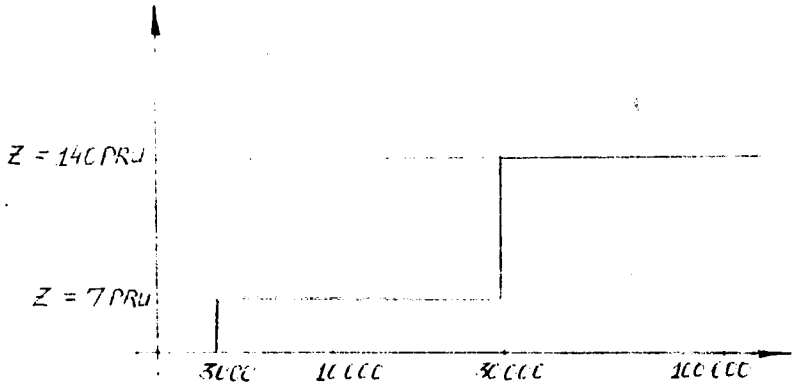


Рис.7

Здесь сплошной линией показано изменение $Z_{опт.ОМПОД}$, штриховой линией - $Z_{опт.ОМД}$.

Все до сих пор сказанное подчеркивает тот факт, что величина оптимальной зоны после достижения определенного объема фактически не зависит от величины информационного массива. В качестве оптимальных зон $Z_{ОМПОД}$ и $Z_{ОМД}$ рассматриваются такие зоны, при которых Т двухур$<$Т одноур. и при этом имеется самый маленький возможный расход памяти на управляющую часть массива (т.е. в данном случае на массив ОМД). Время обработки для одноуровневой структуры зависит прежде всего от средней длины дескрипторного списка - L. Чем длинее L, тем больше времени уходит на поиск (см.формулу I). Однако с увеличением L увеличиваются (после N_Z =I0000) прямо пропорционально коэффициенты $C_{кОМД}$ и $C_{кОМПОД}$ при уже выбранной оптимальной зоне. Напомним, что $C_{кОМПОД}$= а $C_{кОМД}=K_2$ ($C_{кОМПОД}$- I).

При достижении некоторого L^*, в данном случае это 27 (при 30000 документах), как уже говорилось,за счет изменения границы $Q_{ОМПОД}$ имеется предельная возможность (при постоянных в данном случае K_I и K_2) изменения зоны ОМПОД, т.е. достигается ее максимальное значение. И в дальнейшем, конечно, двухуровневая структура всегда будет работать лучше одноуровневой, но она не будет являться оптимальной структурой по отношению ко времени ответа для некоторого достаточно большого информационного массива. В связи с этим отметим, что оптимизация двууровневой структуры связана с уменьшением коэффициентов K_I и K_2, что реализуется путем "уплотнения" дескрипторных списков. Но даже и после применения соответствующих оптимизационных методов, после того как достигнута максимальная величина оптимальной зоны, если информационный массив продолжает расти, то приходит момент, когда возможно большое уменьшение коэффициента Р ($P = \frac{Z_{ОМПОД}}{C_{кОМД}}$).

Это означает ($C_{кОМД}$ достаточно большое), что "впустую" просматривается большое число зон в массиве ОМД, а в результате только маленькое количество из них является релевантным запросу. Это те зоны, которые в дальнейшем должны быть просмотрены в массиве ОМПОД - $З_{ОМПОД}$. В этом случае улучшение времени поиска может быть достигнуто за счет некоторого увеличения объема управляющей части вследствие перехода к новой иерархии, в данном случае к трехуровневой структуре. Рассмотрим этот вопрос более подробно. Напомним, что время трехуровневой структуры вычисляется по формуле 3. $T_{трехур} = C_{кОМД} T_{Z_1} + З_{МЗД} T_{Z_2} + З_{ОМПОД} T_{Z_3}$.

Рассмотрим отдельные параметры этой формулы. Заметим, что $З_{ОМПОД}$ и $C_{кОМПОД}$ вычисляются по уже сформулированным выражениям 6+II.

Число зон массива МЗД - $З_{МЗД}$ можно вычислить по формуле

$$З_{МЗД} = \frac{K_1 \cdot З_{ОМПОД} \cdot Z_{ОМПОД}}{Z_{МЗД}} . \qquad (19)$$

Физическая сущность этого параметра состоит в том, что он отражает число зон массива заголовков. Известно, что число всех заголовков, выходящих из зон массива ОМПОД, вычисляется следующим образом:

$$\text{число всех заголовков} = \frac{K_1 Z_{ОМПОД} \cdot 64 \cdot З_{ОМПОД}}{2} . \qquad (20$$

С другой стороны,

$$З_{МЗД} = \frac{\text{число всех заголовков} \cdot 2}{Z_{МЗД} \cdot 64} . \qquad (21)$$

Подставляя (20) в (21), получим формулу (19).

В /4/ рассмотрена подробно структура масива МЗД. Там же отмечено, что в пределах данной зоны все одинаковые заголовки соединяются в цепь. В данном случае в ОМД накапливаются только заголовки цепных списков зон массива МЗД. Тогда все элементы массива ОМД можно вычислить следующим образом:

$З_{ОМД}$ = K_2 число всех заголовков =

$$= \frac{K_2 K_1 Z_{ОМПОД} \cdot 64 \cdot З_{ОМПОД}}{2} . \qquad (22)$$

Формула (22) может быть использована для вычисления $C_{к\ МЗД}$, которое обозначает среднее число зон на дескриптор в массиве МЗД.

$$C_{к\ МЗД} = \frac{З_{ОМД}}{V} . \qquad (23)$$

Подставляя (22) в (23), получим

$$C_{к\ МЗД} = \frac{K_2 K_1 Z_{ОМПОД} \cdot 64 \cdot З_{ОМПОД}}{2V} . \qquad (24)$$

По аналогии с формулой (I7) можно привести формулу для вычисления числа зон второй части массива ОМД трехуровневой структуры.

$$З_{ОМД_{II}} = \frac{K_2 K_1 Z_{ОМПОД} \cdot 32 \cdot З_{ОМПОД} - V}{32 \cdot Z_{ОМД}} . \qquad (25)$$

Здесь же по аналогии с коэффициентами K_I и K_2 вводится коэффициент K_3, который связан с числом различных заголовков в данной зоне массива ОМД и который служит для вычисления параметра $C_{кОМД}$.

$$C_{кОМД} = \frac{K_3 З_{ОМД_{II}} \cdot Z_{ОМД} \cdot 64}{2V} . \qquad (26)$$

Подставляя (25) в (26), получим:

$$C_{кОМД} = \frac{K_3 K_2 K_1 Z_{ОМПОД} \cdot 32 \cdot З_{ОМПОД} - V}{V} , \qquad (27)$$

$$C_{кОМД} = K_3 (C_{к\ МЗД} - I) . \qquad (28)$$

Применяя далее формулы к различным объемам массива при $Z_{ОМД} = Z_{МЗД} = 7PRU$ и $Z_{ОМПОД} = I40$ и при значениях коэффициентов K_I=0,43 и K_2=0,9, K_3=0,9 (полученных экспериментальным путем), получаем, что Т трехур. > Т двухур. Различные комбинации разных величин зон не изменяют это неравенство. Дальнейший анализ пока-

зывает, что для лучшей работы трехуровневой структуры необходимо уменьшить коэффициент K_3. Однако этот коэффициент зависит прежде всего от частоты встречаемости дескрипторов $f(j)$, т.е. от характеристик входного потока, т.к. он непосредственно связан с K_1 и K_2. Поэтому если исследуемый информационный массив такой, что при реализации иерархической многоуровневой структуры K_1, K_2 и K_3 получаются близкими к единице, то необходимо после нахождения оптимальных зон двухуровневой структуры оптимизировать эту структуру (путем уплотнения списков методом автоматической сортировки[2]), а потом принимать решение о целесообразности перехода к трехуровневой структуре.

ЛИТЕРАТУРА

1. Д.Д.Арнаудов. Сообщение ОИЯИ, Р10-8622, 1975.
2. Д.Лефковитц. Структуры информационных массивов оперативных систем , М., 1973.
3. Д.Д.Арнаудов. Препринт ОИЯИ, 10-7953, Дубна, 1974.
4. Д.Д.Арнаудов. Сообщение ОИЯИ, Р10-8621, Дубна, 1975.
5. Д.Д.Арнаудов, З.И.Коженкова. Сообщение ОИЯИ, 10-8799, Дубна, 1975.
6. Д.Д.Арнаудов, Н.И.Янев. Сообщение ОИЯИ, 11-8555, Дубна, 1975.
7. Д.Д.Арнаудов. Сообщение ОИЯИ, 10-7949, Дубна, 1974.

Рукопись поступила в издательский отдел
15 октября 1975 года.