ОБЪЕДИНЕННЫЙ
ИНСТИТУТ
ЯДЕРНЫХ
ИССЛЕДОВАНИЙ
ДУБНА

Р10 - 10368

Д.Д.Арнаудов

НЕКОТОРЫЕ ВОПРОСЫ ОРГАНИЗАЦИИ БОЛЬШИХ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

1977

Р10 - 10368

Д.Д.Арнаудов

НЕКОТОРЫЕ ВОПРОСЫ ОРГАНИЗАЦИИ БОЛЬШИХ ИНФОРМАЦИОННО-ПОИСКОВЫХ СИСТЕМ

Направлено в журнал "Научно-техническая информация"

Арнаудов Д.Д. Р10 - 10368

Некоторые вопросы организации больших информационно-поисковых систем

В работе рассматриваются три основные проблемы: проблема отображения информационного массива на устройствах памяти, проблема логической организации массива и проблема поиска. Особое внимание обращается на вопросы построения самоорганизующейся структуры массива, а также на вопросы допоискового прогнозирования и работы с автоматизированным информационным архивом на магнитных лентах.

Работа выполнена в Лаборатории вычислительной техники и автоматизации ОИЯИ.

Препринт Объединенного института ядерных исследований. Дубна 1977

Р10 - 10368

Arnaudov D.D.

Some Problems of the Organization of Large Information Retrieval Systems

Three main problems are considered in this work: problems connected with the"reflection"of the information file on the external memory devices, problems connected with the logical organization of the file structure, and problems of the search strategy.Special attention is paid to the organization of the selfadapting structure of the file. At the same time some aspects of the question of presearch prognosis of the relevant documents and the work with the archive information on the magnetic tapes are described.

Preprint of the Joint Institute for Nuclear Research. Dubna 1977

В основе разрешения проблем своевременной обработки информации различного характера (не только научно-технической, но и политической, экономической, медицинской, военной и т.п.) стоят задачи организации больших информационно-поисковых систем. Они охватывают достаточно широкий круг вопросов, связанных с хранением, поиском и обработкой информации. При решении этих задач необходимо иметь в виду, что они характеризуются как большим объемом, так и частыми обращениями к запоминаемой информации, хранимой во внешних запоминающих устройствах. В этих условиях время обработки определяется не столько быстродействием вычислителя, сколько временем обращения к внешним запоминающим устройствам. Поэтому создание рациональных методов организации хранения и поиска данных в машинной памяти приобретает существенное значение.

Мы остановимся на нескольких проблемах, которые, на наш взгляд, имеют важное значение при конкретной реализации информационно-поисковой системы на современной вычислительной машине.

Первая проблема посвящена вопросам отображения логической структуры массива на устройствах памяти. Эти вопросы особенно актуальны при создании эффективной организации информационного массива, так как учитывают конкретные особенности операционной системы ЭВМ и соответствующих устройств памяти.

Вторая проблема – это проблема организации информационного массива, которая остро встает, когда поисковый массив становится настолько большим, что невозможно или, по крайней мере, неэкономично просматривать каждый элемент массива для проверки его соответствия. Следовательно, необходимо, чтобы поисковый массив имел структуру, обеспечивающую возможность использования менее сложных механизмов индексации и промежуточных мер для определения степени соответствия, чем окончательная мера соответствия. Именно этот аспект порождает трудности при разработке системы информационного поиска, связанные с согласованием объема памяти поискового массива, времени, необходимого для получения ответа, и точности ответа. Для правильного разрешения этой проблемы необходимо, чтобы организация поискового массива удовлетворяла специфическим требованиям поиска, включающим такие аспекты, как время ответа, логика запроса, тип запоминающего устройства и т.п.

Наряду с этим возникает и третья основная проблема, которая охватывает вопросы создания эффективной стратегии поиска документов. Здесь следует отметить особенно задачу допоискового прогнозирования числа релевантных документов, методы работы с информационным архивом, вопросы поиска как по определенным наборам дескрипторов, связанных операторами булевой алгебры, так и по характеристикам, не являющимся дескрипторами.

Отображение логической организации массива на физических устройствах памяти

При разработке информационно-поисковой системы нельзя лишь в теоретическом плане определять оптимальную организацию информационного массива, так как она зависит как от функциональных особенностей разрабатываемой системы, так и от конкретных особенностей операционной системы ЭВМ и соответствующих устройств памяти. Можно определить четыре основных этапа при разработке методики отображения.

Первый этап характеризуется определением физических уровней и подуровней данного устройства. Понятия (известные в литературе как "блоки", "цилиндры", "страницы", "дорожки") используются при составлении спецификации физического устройства. В связи с этим для каждого физического уровня должны выполняться следующие требования:

1. Уровень должен быть определен в терминах своих компонентов – физических подуровней, которые он охватывает.

2. Должно быть определено место уровня в иерархической организации памяти во время работы.

3. Необходимо предусмотреть меры на случай его переполнения.

Второй этап характеризуется определением формата физического устройства в терминах физических уровней. Фактически он сводится к описанию физического устройства с помощью его уровней /I/.

На третьем этапе определяются средства и методика адресации для данного устройства /2/.

На четвертом этапе определяется величина логической записи массива /IO/.

Так, например, для дискового пакета 84I на ЭВМ СДС-6400 в качестве физического уровня дискового устройства можно выбрать один цилиндр. Решающее значение для подобного выбора имеет способ физической записи и чтения информации на диске /I/.

На рис.I показана схема "вертикального" разреза цилиндра дискового устройства.

	0	1	2	...	12	13
0	PRU 0	PRU 28	PRU 1		PRU 6	PRU 34
1	PRU 7					
2	PRU 14					
⋮						
19						

Рис.I

Информация на диске записывается (или считывается) последовательно,секторами. На каждой дорожке цилиндра находятся I4 секторов. Каждый сектор имеет величину *1 PRU (Physical Record Unit)* – это величина физической записи на магнитном диске. Каждая физическая запись состоит из 64 слов. Запись (чтение) производится сначала по четным секторам (0,2,4,....I2), при обходе всех двадцати дорожек цилиндра, а потом – по нечетным (I,3,5,...I3). Это означает, что для заполнения одного цилиндра необходимы минимум сорок оборотов дискового пакета.

При определении способа адресации (на третьем этапе) необходимо иметь в виду,что при работе операционной системы каждый массив, размещенный на диске, физически сегментируется на блоки (например, при работе дисковой операционной системы СДС один блок содержит *56PRU*). Адресацию можно вести как в действительных, так и в относительных адресах[2,3]. Например, для идентификации данного элемента массива можно ввести некоторое семизначное число ХУ. Первые три цифры (Х) определяют номер зоны,а последние четыре цифры (У) – номер элемента зоны. Установление необходимого соотношения между относительными адресами и действительными адресами на диске показано нами в[II].

Выбор величины и типа логической записи массива описан в [3,4,IO]; при этом нужно учитывать как естественную сегментацию памяти на цилиндрах, так и функциональные особенности стратегии поиска с целью минимизации времени ответа. Более подробно этот вопрос рассматривается нами дальше в этой работе при определении адаптирующейся структуры поискового массива.

Необходимо заметить, что современные вычислительные машины характеризуются наличием памяти более чем на двух уровнях. Так, например, СДС-6500 имеет трехуровневую память: 0-ой уровень – оперативная память, I-ый уровень – расширенная память, 2-ой уровень – память на МД. Поэтому особое значение при отображении структуры массива имеет определенное место отдельных частей массива в иерархической организации памяти во время работы системы поиска. В этом случае мы предполагаем, что весь информационный массив (поскольку он достаточно большой) хранится в памяти на самом низком уровне – на магнитных дисках. В процессе его обработки, однако, происходит перемещение блоков данных на более высокий уровень иерархии. Поэтому правильное распределение отдельных фрагментов структуры массива на соответствующих уровнях памяти имеет большое значение для эффективности поиска. В работе[6] показано, что введение иерархической структуры машинной памяти дает возможность уменьшить на I3% время доступа в достаточно сложной ассоциативно-адресной структуре.

Организация информационного массива

Как уже отмечалось,проблема организации информационного массива на устройствах памяти ЭВМ остро встает лишь тогда,когда информационный массив становится настолько большим,что возникают зна-

чительные трудности, связанные с согласованием объема памяти поискового массива и времени для получения ответа. Эти трудности могут быть разрешены на основе создания адаптирующейся многоуровневой информационно-поисковой системы, где число уровней иерархии информационного массива и величина сегмента массива определенного уровня могли 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 /7/ отмечает, что можно различить два полюса в организации массивов - организацию инверсную и мультисписковую(ассоциативно-адресную).Все остальные варианты можно рассматривать как частично инвертированные мультисписковые схемы. Проведенный нами анализ инверсной и мультисписковой организации информации /2/ показал, что с нарастанием объема основного информационного массива ухудшаются поисковые характеристики как мультисписковой, так и инверсной организации. В работе /IO/ мы показали, что улучшения поисковых характеристик системы можно достигнуть путем комбинирования признаков инверсной и мультисписковой организации, создавая многоуровневую самоорганизующуюся поисковую систему на базе определенной стратегии поиска /4/.

Здесь отметим некоторые основные характеристики этого оригинального метода, позволяющего адаптировать структуру массива при увеличении его объема путем изменения двух параметров структуры: величины сегмента (зоны) и числа уровней иерархии.

В основе предложенной методики лежит принцип последовательного раскрытия критерия соответствия /4/, что дает возможность исключить из рассмотрения те части массива, для которых априорно ясно,

что наверняка не встретится релевантный запросу документ. Принятый принцип раскрытия критерия соответствия предполагает организацию иерархической структуры информационного массива. На рис.2, 3,4 показана структура основного информационного массива ОМПОД (основного массива поисковых образов документов) с различным числом уровней иерархии управляющей части. На рис.2 представлена классическая ассоциативно-адресная структура. Управляющая часть представляет набор дескрипторов словаря системы. Коды дескрипторов совпадают с номерами (индексами) элементов массива. Каждому дескриптору соответствует информация об адресе первого члена соответствующего ему списка (АС - адрес связи) и о числе членов этого списка (ЧД - число членов). На рис.3 показана уже частично-инвертированная структура. Все узловые списки 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 ($AC2$, $NZ2$, $ND2$). Более подробно эта структура рассмотрена в /I2/. Управляющая часть состоит из двух частей, т.е. имеет двухуровневую структуру.

Управляющая часть с трехуровневой структурой (массив ОМД из двух частей и массив МЗД /I2/) показана на рис.4.

Для более полного анализа принципа адаптации структуры массива при увеличении его объема необходимо иметь в виду его основные характеристики, которые связаны как с его структурой, так и с характеристиками дискового пакета и операционной системы ЭВМ, а также с накопленной статистикой запросов пользователей. Все характеристики приведены в таблице I.

Таблица 1

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_K – количество дескрипторов, индексирующих документ (среднее)

$L = \frac{N_z N_K}{V}$ – количество документов на дескриптор (средняя длина списка)

Q – количество символов на запись

R_c – среднее количество записей в зоне

$C_K = \begin{Bmatrix} C_{K\,омд} \\ C_{K\,мзд} \\ C_{K\,омпод} \end{Bmatrix}$ – среднее количество зон на дескриптор (соответственно в массивах

$rз = \begin{Bmatrix} rз_{омд} \\ rз_{мзд} \\ rз_{омпод} \end{Bmatrix}$ – число зон массивов ОМД, МЗД и ОМПОД

$rз_a = \begin{Bmatrix} rз_{з\,омд} \\ rз_{з\,мзд} \\ rз_{з\,омпод} \end{Bmatrix}$ – среднее количество зон запроса (соответственно в массивах ОМД, МЗД, ОМПОД)

K – коэффициент, указывающий на наличие более чем одного дескр. в зоне I части ОМД

N_t – количество дескрипторов в запросе

N_p – количество положительных дескрипторов в запросе

L_s – длина самого короткого списка

ρ – отношение ср. числа ответов на запрос к L

$p = \frac{rз_{з\,омпод}}{C_{K\,омд}}$ – отношение среднего числа общих зон запроса в ОМПОД к $C_{K\,омд}$ (для двухуровневой схемы)

$\delta = \frac{rз_{з\,мзд}}{C_{K\,омд}}$ – отношение среднего числа общих зон запроса в ОМПОД к $C_{K\,омд}$ (для трехуровневой схемы)

$a = \begin{Bmatrix} a_{омд} \\ a_{мзд} \\ a_{омпод} \end{Bmatrix}$ – отношение среднего количества общих зон запроса к C_K, где

$$a_{омд} = \frac{rз_{з\,омд}}{C_{K\,омд}};\ a_{мзд} = \frac{rз_{з\,мзд}}{C_{K\,мзд}};\ a_{омпод} = \frac{rз_{з\,омпод}}{C_{K\,омпод}}$$

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.

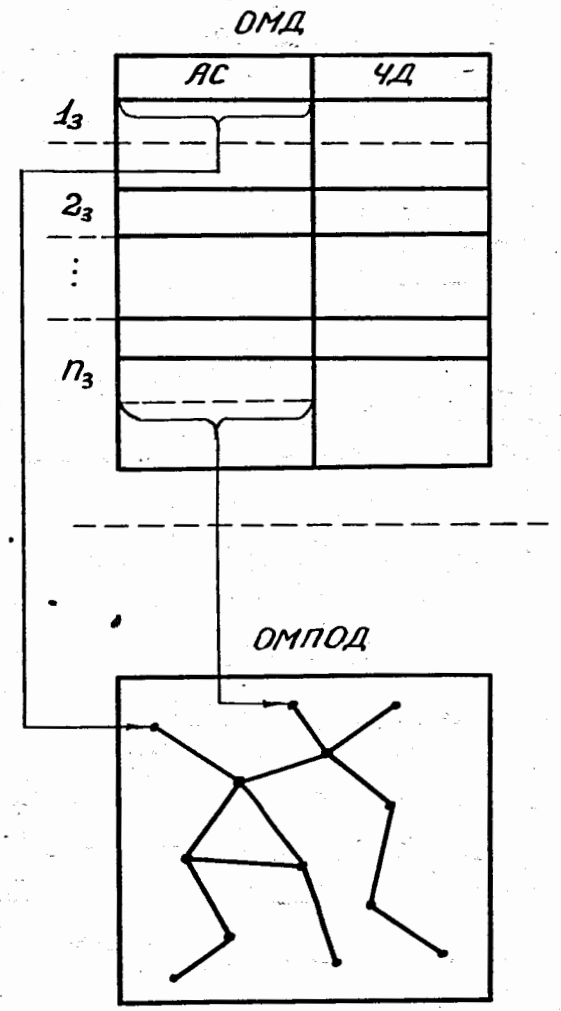

Рис. 2

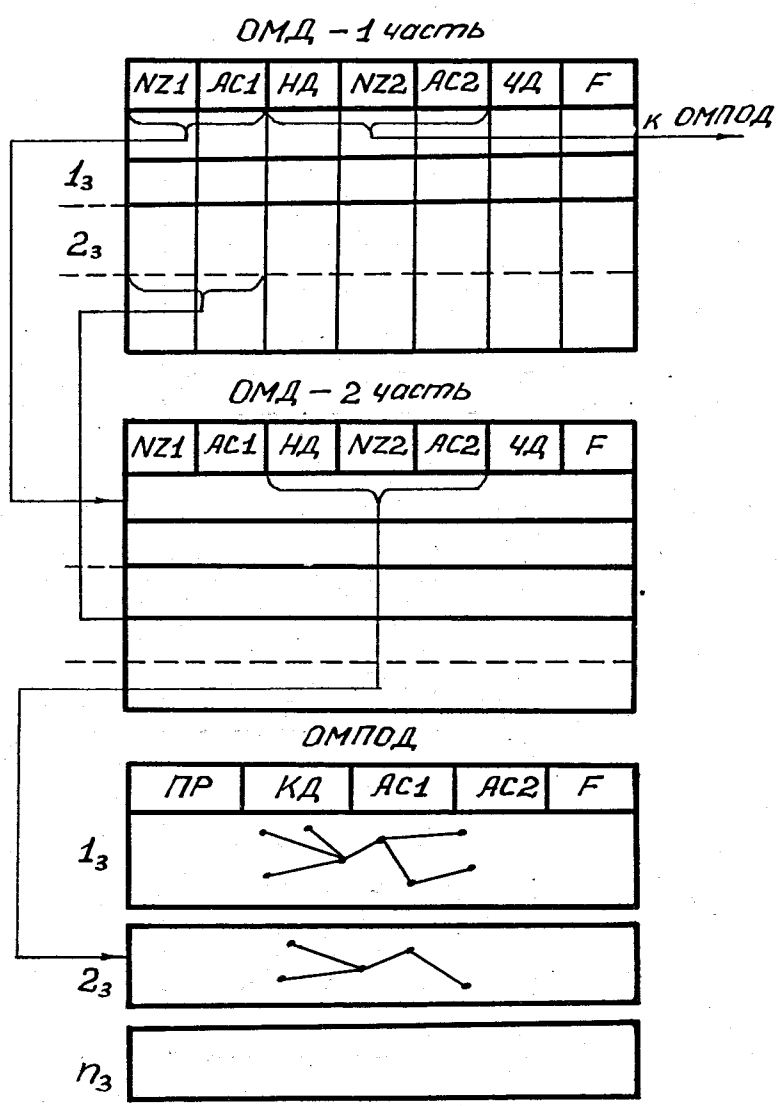
ОМД – 1 часть
NZ1
АС1
НД
NZ2
АС2
ЧД
F
к ОМПОД
1з
2з
ОМД – 2 часть
NZ1
АС1
НД
NZ2
АС2
ЧД
F
ОМПОД
ПР
КД
АС1
АС2
F
1з
2з
nз

Рис. 3

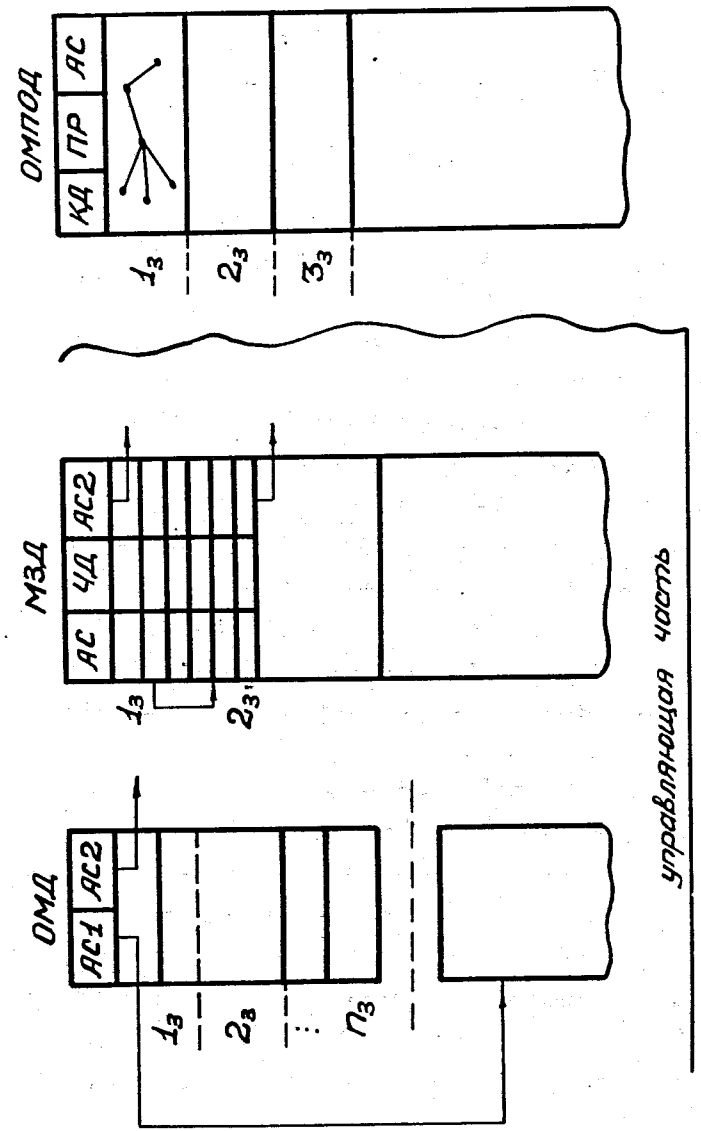
ОМПОД
АС
ПР
КД
1з
2з
3з
МЗД
АС2
ЧД
АС
1з
2з'
ОМД
АС2
АС1
1з
2з
nз
управляющая часть

Рис. 4

На основе принятой стратегии поиска (на вопросе о стратегии поиска мы остановимся дальше в нашем изложении), предлагаются следующие формулы для подсчета среднего времени обработки информационного массива с одноуровневой, двухуровневой и трехуровневой структурой.

$$T_{одноур.} = KN_p T_z + L_S T_S = KN_p T_z + L_S (t_п + \tfrac{1}{2} t_o) ; \quad (I)$$

$$T_{двухур.} = KN_p T_z + a_{омд} C_{комд} T_{z_1} + a_{ompод} C_{компод} T_{z_2} ; \quad (2)$$

$$T_{трехур.} = KN_p T_z + a_{омд} C_{комд} T_{z_1} + a_{мзд} C_{кмзд} T_{z_2} + a_{ompод} C_{компод} T_{z_3} . \quad (3)$$

Первые одночлены всех формул одинаковы. Здесь учитывается время, которое расходуется на декодирование дескрипторов запроса. Необходимо отметить, что любой запрос представляет конъюнкцию дескрипторов. В дальнейшем рассматриваются только положительные дескрипторы N_p (дескрипторы без знака отрицания). Дескрипторы с отрицанием используются только на последнем этапе поиска, в узлах ОМПОД, и практически не влияют на время ответа, вычисляемое по формулам I-3. Если запрос состоит из суммы произведений (т.е. представлен в дизъюнктивной форме), то время обработки запроса меньше суммы времени обработки отдельных конъюнкций из-за принятой стратегии параллельной обработки запросов[4].

Коэффициент K учитывает одновременную выборку KN_p 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 сокращает время выборки.

Параметр T_Z означает время чтения зоны (аналогично T_{Z_1}, T_{Z_2}, T_{Z_3}). В формуле (I) через L_S обозначена длина самого короткого дескрипторного списка. Анализируя эту формулу, необходимо отметить, что в структуре на рис.2

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 $t_S = t_п + \frac{1}{2} t_o$ (более подробно см.[3,5,6]).

Исходя из принятой стратегии поиска[4] в двухуровневой и трехуровневой структурах, необходимо просмотреть все элементы цепи дескрипторов в ОМД (т.е. в высшем уровне иерархии), поэтому в этом массиве всегда просматривается среднее количество зон на дескриптор. Тогда $a_{омд} = 1$; и формулы (2) и (3) принимают следующий вид:

$$T_{двухур.} = KN_p T_z + C_{комд} T_{z_1} + a_{ompод} C_{компод} T_{z_2} ; \quad (4)$$

$$T_{трехур.} = KN_p T_z + C_{комд} T_{z_1} + a_{мзд} C_{кмзд} T_{z_2} + a_{ompод} C_{компод} T_{z_3} . \quad (4а)$$

Основной вопрос, который необходимо решить на основе предложенных формул (2,4,4а),- это когда и в каких случаях необходимо переходить от одной структуры к другой (см.рис.2,3,4) и какие величины зон сегментированных списков являются самими подходящими при проведении поиска в соответствующих структурах.

Так как первое слагаемое одинаково во всех формулах (I-3) (оно меньше на порядок, чем другие слагаемые), то сравнительный анализ проведен на основе следующих выражений:

(I) $T_{одноур.} = L(t_п + \frac{1}{2} t_o)$;

(2) $T_{двухур.} = C_{комд} T_{z_1} + a_{ompод} C_{компод} T_{z_2}$;

(3) $T_{трехур.} = C_{комд} T_{z_1} + a_{мзд} C_{кмзд} T_{z_2} + a_{ompод} C_{компод} T_{z_3}$.

В формуле (I) вместо L_S берем L - среднюю длину списка в ОМПОД, так как $\frac{\sum_{S=1}^{N} L_S}{N} \leq L$,

где $L = \frac{N_z N_K}{V}$.

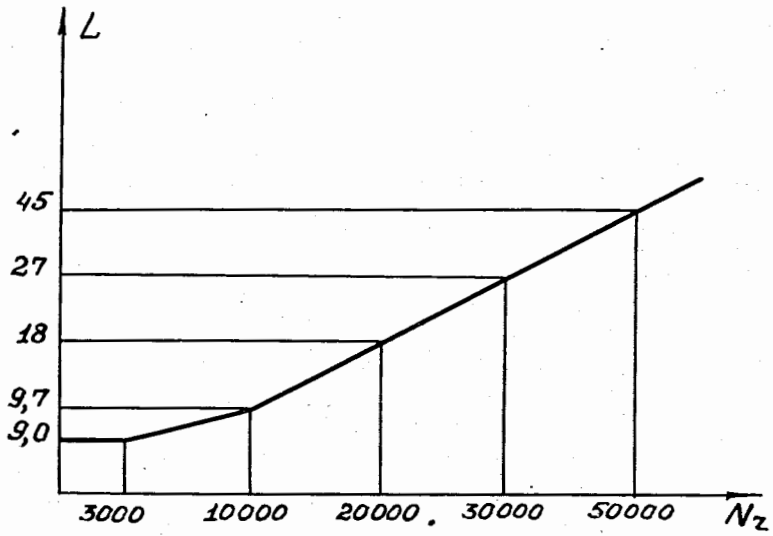

Рис.5

На рис.5 приведена зависимость $L = f(N_z)$.

В работе[10] предложены формулы для расчета всех параметров одноуровневой, двухуровневой и трехуровневой структур. Анализ временных характеристик одноуровневой и двухуровневой структур показал, что величина зоны зависит как от величины средней длины списка, так и от возможностей оперативной памяти ЭВМ и особенностей операционной системы. На рис.6 показана зависимость величины зоны (Z) от изменения средней длины списка. Интерес представляет L^*, когда эта кривая пересекает линию Z_{max}, что определяет максимально возможную зону. Это означает, что дальше зону увеличивать нельзя. Известно, что чем больше зона, тем больше элементов списка могут расположиться в данной зоне, тем меньше число элементов в цепях ОМД. Однако в условиях, когда нельзя увеличивать зону больше, наступает момент, когда характеристика ρ становится очень мала (порядка сотых или даже тысячных долей) и необходимо перейти к трехуровневой структуре.

На основе предложенных формул, а также на основе исследования характеристик информационного массива ИНИСа[9], который использовался в качестве экспериментального массива, показано, что число уровней иерархии управляющей части массива зависит от его величины. Мы определили и экспериментально проверили величины оптимальных зон основного массива ОМПОД, а также величины зон управляющей части для двухуровневой и трехуровневой структур. Показано (при использовании ЭВМ СДС-6400), что для массива до 3000 документов лучшей структурой является одноуровневая; при числе документов с 3000 до 30000 массив должен иметь двухуровневую структуру с величинами зон $Z_{ОМД} = Z_{ОМПОД} = 7 PRU$ и при числе документов 30000÷150000 - двухуровневую структуру с $Z_{ОМД}=7$, $Z_{МЗД}=140$, $Z_{ОМПОД}=140$.

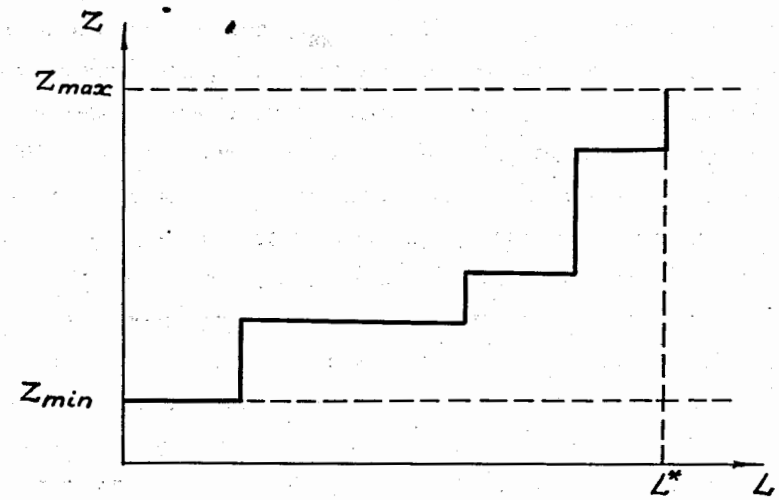

Рис.6

Описанная методика дает возможность построить самоорганизующуюся ИПС, в которой время поиска существенно меньше, чем в инверсной и ассоциативно-адресной структурах.

Необходимо отметить, что разработанный метод организации адаптирующейся иерархической структуры массива с эффективными параметрами имеет большое значение для реализации ИПС на ЭВМ с развитыми операционными системами типа СДС-6400, ЕС и т.п. В этом случае есть возможность оптимизировать структуру системы как со стороны функциональных требований ИПС, так и со стороны конкретного учета технико-математического обеспечения современных ЭВМ.

Стратегия поиска

Стратегия поиска в большой ИПС характеризуется принципом последовательного раскрытия критерия смыслового соответствия[4]. Под этим понимается последовательно усложняющийся анализ элемента с постепенным охватом все большего числа его деталей, причем результаты, полученные на предыдущем этапе анализа, исследуются на последующем. Необходимость в использовании последовательности этапов анализа вызвана тем, что по мере того, как растет информационный массив, применять самый сложный анализ к каждому элементу оказывается слишком дорогостоящим делом. На основе этого в работах[4,8] мы предложили метод раскрытия смыслового соответствия, который оказался эффективным при поиске в большом информационном массиве. Сущность этого метода заключается в предварительном отсеве нерелевантных документов. Здесь фактически идет речь о реализации уже известной из литературы плодотворной идеи выработки "негативных стратегий", позволяющих заранее исключить области с бесполезным перебором. Предложенный нами метод нашел свое практическое применение при разработке ИПС Объединенного института ядерных исследований[3,4,12].

Здесь мы остановимся еще на одном методе[5], который дает возможность оптимизировать стратегию поиска в большой ИПС. Это связано с допоисковым прогнозированием числа релевантных документов. Конечно, для решения этого вопроса используются свойства включающей ИПС.

Пусть $Д$ означает конечное множество документов (количество документов - K), которые используются для обслуживания Q запросов. Содержание документов индексируется дескрипторами, выбираемыми из дескрипторного словаря N. Элементы множества являются любыми наборами из элсментов множества N, связанными знаками конъюнкции, дизъюнкции и отрицания. Пусть K означает некоторую функцию (поисковую), дефинированную в Q, со значениями в $2^{Д}$ (где $2^{Д}$ означает множество всех подмножеств Д). Тогда каждому запросу $q \in Q$ сопоставляется определенный элемент из $2^{Д}$, который называют выдачей на запрос q (документ, релевантный запросу q).

ИПС может формально быть определена как совокупность конечного множества документов $Д$, дескрипторного языка N, поисковой функции R и языка запросов Q. ИПС частично упорядочена, если Q является частично-упорядоченным множеством, если на Q определено транзитивное и антисимметричное отношение $\geq$. Причем, если $q \in Q$ и $S \in Q$, то q меньше S, если $q \wedge S = q$. Такое соотношение введено и для элементов множества $2^{Д}$. Вследствие этого ИПС является включающей, если функция R такова, что для $q < S$ $R(q) > R(S)$, причем $\{q, S\} \in Q$. ИПС ОИЯИ является включающей информационно-поисковой системой. Это означает, что свойство включения гарантирует нахождение всех найденных по запросу S документов и по запросу q. Имеется в виду, что q тождественно S за исключением того факта, что данный дескриптор, включенный в S, исключен при формировании q. При этом

по запросу q, однако, может быть найдено дополнительное число документов, не найденных ранее по S, так что $R(q)$ может быть больше, чем $R(S)$.

Взаимосвязь между порядком, определенным в пространстве запросов Q, и соответствующим порядком в пространстве документов $2^Д$ задается следующей теоремой.

Теорема. Образом уменьшающейся цепи в Q при отображении R во включающей ИПС является увеличивающаяся цепь в $2^Д$, а образом увеличивающейся цепи в Q – уменьшающаяся цепь в $2^Д$.

Мы использовали эту теорему при решении задачи прогнозирования числа релевантных документов. Для этой цели введем еще два обозначения: $\bar{R}(q)$ и γ, где $\bar{R}(q)$ означает ожидаемое число релевантных документов (напомним, что $R(q)$ – это точное число релевантных документов, найденное на основе проведенного поиска в системе), а γ – это пороговое число ожидаемых документов. Его физический смысл следующий. Если вследствие прогнозирования ожидается получить $\bar{R}(q)$ документов, причем $\bar{R}(q) > \gamma$, то тогда необходима переформулировка запроса (т.е. либо увеличение числа дескрипторов конъюнкции, либо использование "более узких" дескрипторов).

Итак, задача прогнозирования сводится к установлению верности следующих неравенств:

$$\bar{R}(q) \le \gamma ; \qquad (5)$$

$$\bar{R}(q) > \gamma . \qquad (6)$$

При установлении верности неравенства (5) запрос считается "корректным" и производится поиск по запросу согласно принятой в ИПС стратегии поиска. Если, однако, установлено неравенство (6), то следует, как уже отметили, переформулировка запроса.

Заметим, что при первом приближенном решении задачи прогнозирования в качестве ответа $\bar{R}(\alpha_{i^*})$ используется частота встречаемости дескрипторов в поисковых образах документов – $f(i^*)$ /10/. При этом в таблицу сведены только те $f(i^*)$, для которых выполняются неравенства

$$f(i^*) > \gamma .$$

Если при исследовании $f(i^*)$ (i = I,....) всех дескрипторов конкретного запроса q окажется, что среди них имеется хоть один, у которого $f(i^*) \le \gamma$ (т.е. его частота встречаемости не указана в таблице), то запрос, как уже заметили, считается корректным и подлежит дальнейшей поисковой обработке. При этом в качестве $f(i^*)$ неравенства (7) используется минимальная частота дескрипторов запроса. Если, однако, условие (7) не выполняется, то тогда переходим к дальнейшему исследованию запроса, рассматривая все возможные пары дескрипторов запроса.

Наши работы показывают/5/, что для запросов, состоящих из трех-четырех дескрипторов, можно принять в качестве оценки прогнозирования исследование неравенства до второго ранга включительно (т.е. рассмотрение пар дескрипторов).

На базе предложенной методики прогнозирования можно создать модель допоискового прогнозирования ИПС. Однако его необходимо вводить в ИПС только тогда, когда имеется значительный объем информационного массива. Тогда, выбирая подходящее пороговое число для коротких запросов (не больше трех-четырех дескрипторов), можно ожидать удовлетворительных результатов.

В связи с программной реализацией модуля допоискового прогнозирования разработан метод формирования частотной таблицы в ИПС ОИЯИ/5/. Так, например, частотная таблица пар дескрипторов массива из 3200 документов первоначально имела 604450 элементов.

Фактически необходимы для прогнозирования только 1975, которые имеют частоту больше критической. Предложенный метод формирования дает возможность создавать подобные таблицы для большого объема документов. Так, например, для массива из 300000 документов необходимо не больше 6 часов машинного времени.

Экспериментальные исследования массива ИНИС, состоящего из 200000 документов, показали, что при использовании модуля допоискового прогнозирования время поиска уменьшается в 4-5 раз.

Особое значение при создании системы поиска в большой ИПС имеет создание методики работы с автоматизированным информационным архивом. Она может быть разработана на основе стратегии обмена документов между оперативным хранилищем (на магнитных дисках) и архивом (на магнитных лентах). Эта стратегия характеризуется следующими параметрами:

а) среднее количество поступающих документов в единицу времени;

б) емкость оперативного хранилища документов;

в) критерий перемещения данного документа из оперативного хранилища в архив;

г) критерий возврата данного документа из архива в оперативное хранилище;

д) промежуток времени, через который происходит периодическое обновление оперативного хранилища и возврат некоторых документов из архива в оперативное хранилище.

Необходимо заметить, что большое значение при изучении числа обращений к отдельным документам имеет год издания документа, т.е. его возраст. Однако практика показывает, что возраст документа не является самой важной характеристикой, и во многих случаях лучше считать количество обращений к данному документу за

последние промежутки времени. Для перемещения данного документа из оперативного хранилища в архив используется критерий, представляющий собой комбинацию возраста документа и числа обращений к нему.

Пусть $\alpha_i(t)$ обозначает возраст i-го документа в момент времени t; пусть $N_i(t_1, t_2)$ обозначает число обращений к i-му документу в промежуток времени (t_1, t_2). Тогда критерий для перемещения i-го документа в архив можно выразить следующим образом: i-й документ перемещается в архив, если:

$$\alpha_i(t) > T \qquad (8)$$

или $$x \le \alpha_i(t) \le T \quad и \quad N_i(t-y, t) < K, \qquad (9)$$

где T - константа, обозначающая предельный возраст нахождения документа в оперативном хранилище;

K - число обращений (запросов) к документу в предыдущие y единиц времени, если этот документ находится в оперативном хранилище хотя бы x единиц времени ($x \ge y \ge 0$ и $x \le T$).

Константу T требуется определить таким образом, чтобы она была достаточно большой, а K, x, y должны подбираться для соответствующей системы динамически (для каждого t решается вопрос о перемещении документа в архив в зависимости от величины места, которое нужно освободить в оперативном хранилище).

Критерий для перемещения документа в архив может быть сформулирован следующим образом: i-й документ будет перемещен в архив, если его возраст больше, чем предельный возраст T, или он находится в оперативном хранилище x единиц времени ($x \le T$) и за последние y единиц времени ($x \ge y \ge 0$) был запрошен меньше K раз.

Очевидно, что при так сформулированном критерии из оперативного хранилища согласно условию (8) будут перемещены в архив все документы с возрастом более чем T. При этом существует опасность перемещения в архив "старых", но "ценных" документов, поэтому наложим следующее дополнительное условие. Если $N_i(t-y, t) \geq \bar{K}$, то независимо от того, что $\alpha_i(t) > T$; i - й документ остается в первичном хранилище, где $\bar{K}$ -константа, определенная для данной ИПС. Это дополнительное условие вместе с условиями (8) и (9) определяют критерий для перемещения документов в архив в следующем виде: i - й документ перемещается в архив, если

$$\alpha_i(t) > T \quad и \quad N_i(t-y, t) < \bar{K}; \tag{10}$$

$$X \leq \alpha_i(t) \leq T \quad и \quad N_i(t-y, t) < K. \tag{11}$$

При введенных уже обозначениях i - й документ будет возвращен из архива в оперативное хранилище, если:

$$N_i(t-y, t) \geq K \quad и \quad \alpha_i(t) \leq T \tag{12}$$

или

$$N_i(t-y, t) \geq \bar{K}. \tag{13}$$

Это означает, что i - й документ возвращается в оперативное хранилище, если за последние y единиц времени он был запрошен больше чем K раз и его возраст не больше, чем T, или число обращений к нему за последние y единиц времени больше или равно $\bar{K}$ независимо от возраста документа.

На основе принятой стратегии и характеристики входного информационного потока разработан метод, позволяющий в динамическом режиме организовать работу обновления оперативного хранилища при максимальном использовании емкости хранилища. На базе этого метода программно реализован алгоритм на ЭВМ СДС-6400. Экспериментальные исследования показывают, что для однократного обновления оперативного хранилища из 200000 документов, при архиве из 1000000 документов, необходимы 5-6 часов машинного времени.

ЛИТЕРАТУРА

1. Д.Д.Арнаудов. Выбор оптимальной структуры основного информационного массива ИПС для размещения на магнитном диске . Сообщение ОИЯИ, 10-7949, Дубна, 1974.

2. Д.Д.Арнаудов. Анализ методов доступа информации к внешним ЗУ на МД при работе ИПС, реализованной на ЭВМ третьего поколения . Препринт ОИЯИ, 10-7953, Дубна, 1974.

3. Arnaudov D.D.,Govorun N.N. Information Retrieval System of JINR. E10-8855, Dubna , 1975. (Доклад на У-ой Международной конференции по поисковым системам в Крэнфилде, 1975).

4. Д.Д.Арнаудов. Стратегия поиска ИПС ОИЯИ . Сообщение ОИЯИ, Р10-8622, Дубна, 1975.

5. Д.Д.Арнаудов, Н.И.Янев. Допоисковое прогнозирование числа релевантных документов в ИПС ОИЯИ . Сообщение ОИЯИ, Р10-9051, Дубна, 1975.

6. J.Salasin. Hierarchical Storage in Information Retrieval , Comm. of the ACM,V.16, No5, 1973.

7. Д.Лефковитц. Структуры информационных массивов оперативных систем . Сов.радио, М., 1973.

8. Д.Д.Арнаудов. Автоматизирани информационно-търоеши системи . Монография. Изд-во "Техника", София, август 1975.

9. Д.Д.Арнаудов, Н.А.Бирюков. Анализ основных характеристик информационной системы ИНИС . Депонированная публикация ОИЯИ, Б1-11-8553, Дубна, 1975.

10. Д.Д. Арнаудов. Об одном способе организации адаптирующейся многоуровневой информационно-поисковой системы . Сообщение ОИЯИ, Р10-9178, Дубна, 1975.

11. Д.Д.Арнаудов, Н.И.Янев. Алгоритмы формирования основных массивов ИПС ОИЯИ . Сообщение ОИЯИ, Р10-8901, Дубна, 1975.

12. Д.Д.Арнаудов. Структурно-функциональная организация основных поисковых массивов ИПС ОИЯИ . Сообщение ОИЯИ, Р10-8621, Дубна, 1975.

Рукопись поступила в издательский отдел
5 января 1977 года.